Advanced Machine Learning Subsidary Notes

Lecture 17: Gaussian Processes

Adam Prügel-Bennett

April 19, 2021

1 Keywords

- Gaussian Processes, regression

2 Main Points

2.1 The Big Picture

- Gaussian processes are a machine learning method for regression
  - You can use them for classification but it is regression where they excel
- They put a prior on functions (preferring smooth functions)
- Conceptually they are easy
- They are easy to use
- But they are mathematically fiddly to understand
- Gaussian processes were first used for spatial modelling where the technique was know as *krigging*
- They are one of the most powerful methods we have of doing regression
- They are quite general purpose as the prior we use is reasonable for a large number of applications

2.2 Gaussian Processes

- Gaussian Processes are also mathematical objects
- A Gaussian Process assigns a probability to functions depending on their spatial smoothness
  - We assume that the probability of $f(\boldsymbol{x})$ taking a particular value is normally/Gaussian distributed in a way that depends on the value of value of the function at other points
- This will be governed by a *covariance* or *kernel* function $k(\boldsymbol{x},\boldsymbol{y})$ which tells us the covariance of the prior
$$\mathbb{E}_f\left[(f(\boldsymbol{x})-m(\boldsymbol{x}))\,(f(\boldsymbol{y})-m(\boldsymbol{y}))\right]=k(\boldsymbol{x},\boldsymbol{y})$$
  - This is an expectation over all functions from our probability distribution
  - $m(\boldsymbol{x})$ encodes a prior belief in the value of $f(\boldsymbol{x})$—normally we choose this to be 0 as most often we don't have any prior knowledge of the function
  - Because the covariance is quadratic it will be positive semi-definite

- The kernel function is critical to determining the type of function we will learn
- Choosing this kernel function correctly is essential to make GP work

- The values at each point $\boldsymbol{x}$ is normally or Gaussianly distributed

$$p(f|m,k) \propto \mathrm{e}^{-\frac{1}{2}\int (f(\boldsymbol{x})-m(\boldsymbol{x}))\,k^{-1}(\boldsymbol{x},\boldsymbol{y})\,(f(\boldsymbol{y})-m(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}}$$

  - It is a little unclear exactly what this means for functions: working with function spaces one has to be very careful (e.g. to normalise this we have to sum over all functions)
  - Nevertheless it is clear that we can use this for comparing the likelihood of different functions
  - If you discretise the domain of $\boldsymbol{x}$ then

$$p(f|m,k) \propto \mathrm{e}^{-\frac{1}{2}\sum_{i,j}(f(\boldsymbol{x}_i)-m(\boldsymbol{x}_i))\,k^{-1}(\boldsymbol{x}_i,\boldsymbol{x}_j)\,(f(\boldsymbol{x}_j)-m(\boldsymbol{x}_j))}$$

    * the probability function is now a well-define multi-dimensional normal distribution
    * here it is clear that the value at each point is normally distributed
    * to get back to the integral form we have to take a limit where our discrete points become every closer together (this stresses out mathematicians who point out that nasty things can happen, but for the rest of us we just go ahead)
  - We can sample Gaussian processes from this distribution
    * in low dimensions this allows us to visualise the type of functions we are likely to learn
    * you can try this out in the experiment section

2.3 Bayesian Inference

- We can use Gaussian Processes as a prior for performing regression
- We assume we have a set of data points $\mathcal{D} = \big((\boldsymbol{x}_i, y_i)\big|i = 1,\ldots,m\big)$
- The likelihood for data is taken to be a normal distribution

$$p(\mathcal{D}|f) = \prod_{i=1}^{m} \mathcal{N}\left(y_i\middle|f(\boldsymbol{x}_i),\sigma^2\right)$$

  - We assume that $f(\boldsymbol{x}_i)$ is the true value of the function at $\boldsymbol{x}_i$, but, because of measurement error, y_i will be distributed according to a normal distribution
  - σ is the accuracy of our measured data (it is another hyper-parameter we must choose)
  - Note we are assuming the data points are independent of each other (which is usually a good approximation)

- We can now compute a posterior

$$p(f|\mathcal{D},m,k,\sigma) = \frac{p(\mathcal{D}|f,\sigma)\,p(f|m,k)}{p(\mathcal{D}|m,k,\sigma)}$$

  - I have explicitly written in the dependencies on they hyper-parameters m, k and σ
  - The Gaussian Process prior is a conjugate distribution for the normal likelihood
  - Computing probabilities over functions is slightly wild

- It is more useful to compute the probability at a particular point $\boldsymbol{x}^*$

– To compute this we marginalise over all other points
  * This is non-trivial but doable
  * If you have a multivariate normal distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\mathbf{C}) = \frac{1}{\sqrt{|2\,\pi\,\mathbf{C}|}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}\,(\boldsymbol{x}-\boldsymbol{\mu})}$$

    and you integrate over x_i (marginalise it out) then

$$\int_{-\infty}^{\infty}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\mathbf{C})\,\mathrm{d}x_i = \mathcal{N}(\hat{\boldsymbol{x}}|\hat{\boldsymbol{\mu}},\hat{\mathbf{C}})$$

    where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{\mu}}$ are identical to $\boldsymbol{x}$ and $\boldsymbol{\mu}$ except with their i^{th} component removed and $\hat{\mathbf{C}}$ is identical to $\mathbf{C}$ except with the i^{th} row and column removed
  * Intuitively this seems very natural (if x_i can take any value it doesn't change the mean or covariance between other random variables)
  * You can prove this algebraically but it is not a trivial calculation
  * Because of this for a Gaussian Process we only care about those set of points where we have data values
– For all point $\boldsymbol{x}$ (except the point $\boldsymbol{x}^*$) we integrate over the possible values that $f(\boldsymbol{x})$ can take
– Where we have no prior this integral is just a constant—and will cancel with the denominator $p(\mathcal{D}|m,k,\sigma)$
– We are left with

$$p(f(\boldsymbol{x}^*)|\mathcal{D}) \propto \left(\prod_i \int \mathrm{d}f(\boldsymbol{x}_i)\right) \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-f(\boldsymbol{x}_i))^2}\;\mathrm{e}^{-\frac{1}{2}\sum_{i,j=1}^{m}(f(\boldsymbol{x}_i)-m(\boldsymbol{x}_i))\,k^{-1}(\boldsymbol{x}_i,\boldsymbol{x}_j)\,(f(\boldsymbol{x}_j)-m(\boldsymbol{x}_j))}$$
$$\mathrm{e}^{-\sum_{j=1}^{m}(f(\boldsymbol{x}^*)-m(\boldsymbol{x}^*))\,k^{-1}(\boldsymbol{x}^*,\boldsymbol{x}_j)\,(f(\boldsymbol{x}_j)-m(\boldsymbol{x}_j))}$$
$$\mathrm{e}^{-\frac{1}{2}(f(\boldsymbol{x}^*)-m(\boldsymbol{x}^*))\,k^{-1}(\boldsymbol{x}^*,\boldsymbol{x}^*)(f(\boldsymbol{x}^*)-m(\boldsymbol{x}^*))}$$

  * Writing $f^* = f(\boldsymbol{x}^*)$, $f_i = f(\boldsymbol{x}_i)$, $m_i = m(\boldsymbol{x}_i)$, $k^{-1}_{ij} = k^{-1}(\boldsymbol{x}_i,\boldsymbol{x}_j)$, $k^{*-1}_j = k^{-1}(\boldsymbol{x}^*,\boldsymbol{x}_j)$ and $k^{*-1} = k^{-1}(\boldsymbol{x}^*,\boldsymbol{x}^*)$ then

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2}k^{*-1}(f^*-m^*)^2}\left(\prod_i \int \mathrm{d}f_i\right)\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-f_i)^2-\frac{1}{2}\sum_{i,j=1}^{m}(f_i-m_i)\,k^{-1}_{ij}\,(f_j-m_j)}$$
$$\mathrm{e}^{-\sum_{j=1}^{m}(f^*-m^*)\,k^{*-1}_j\,(f_j-m_j)}$$

  * Doing Gaussian integrals is a pain—*I don't expect you to do this*
  * But if you are brave enough to try let $g_i = f_i - m_i$, $g^* = f^* - m^*$ and $\hat{y}_i = y_i - m_i$ then we can rewrite this as

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2}k^{*-1}g^{*2}}\left(\prod_i \int \mathrm{d}g_i\right)\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(\hat{y}_i-g_i)^2-\frac{1}{2}\sum_{i,j=1}^{m}g_i\,k^{-1}_{ij}\,g_j-g^*\sum_{j=1}^{m}k^{*-1}_i\,g_j}$$

    · where we made a change of variable in the integral
  * In matrix form we can write this as

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2}k^{*-1}g^{*2}}\left(\prod_i \int \mathrm{d}g_i\right)\mathrm{e}^{(\sigma^{-2}\,\hat{\boldsymbol{y}}-g^*\boldsymbol{k}^{*-1})^{\mathsf{T}}\,\boldsymbol{g}-\frac{1}{2}\boldsymbol{g}^{\mathsf{T}}\,(\mathbf{k}^{-1}-\sigma^{-2}\mathbf{I})\,\boldsymbol{g}}$$

- I've dropped a term $\mathrm{e}^{-\frac{\|\hat{\boldsymbol{y}}\|^2}{2\,\sigma^2}}$ since this is a constant that can be absorbed into the constant of proportionality

* To perform the integrals over g_i we *complete the square*. To make the algebra easier to follow we define $\mathbf{M} = \mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}$ and $\boldsymbol{b} = \sigma^{-2}\,\hat{\boldsymbol{y}} - g^*\boldsymbol{k}^{*-1}$. We now rewrite the exponent, E, of the integral as

$$\begin{aligned} E &= (\sigma^{-2}\,\hat{\boldsymbol{y}} - g^*\boldsymbol{k}^{*-1})^{\mathsf{T}}\,\boldsymbol{g} - \frac{1}{2}\boldsymbol{g}^{\mathsf{T}}(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I})\,\boldsymbol{g} = \boldsymbol{b}^{\mathsf{T}}\,\boldsymbol{g} - \frac{1}{2}\boldsymbol{g}^{\mathsf{T}}\mathbf{M}\boldsymbol{g} \\ &= -\frac{1}{2}(\boldsymbol{g} - \mathbf{M}^{-1}\boldsymbol{b})^{\mathsf{T}}\mathbf{M}\,(\boldsymbol{g} - \mathbf{M}^{-1}\boldsymbol{b}) + \frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{b} \end{aligned}$$

  - You simply have to expand out all the terms to see these are the same

* Defining $\boldsymbol{h} = \boldsymbol{g} - \mathbf{M}^{-1}\boldsymbol{b}$ and substituting this back into $p(f^*|\mathcal{D})$ we get

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2}k^{*-1}g^{*2} + \frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{b}} \left(\prod_i \int \mathrm{d}h_i\right) \mathrm{e}^{-\frac{1}{2}\boldsymbol{h}^{\mathsf{T}}\mathbf{M}\,\boldsymbol{h}}$$

* The last integral is a Gaussian integral that integrates to a constant (and can be absorbed into the constant of proportionality)

* We are left with

$$\begin{aligned} p(f^*|\mathcal{D}) &\propto \mathrm{e}^{-\frac{1}{2}k^{*-1}g^{*2} + \frac{1}{2}\boldsymbol{b}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{b}} \\ &= \mathrm{e}^{-\frac{1}{2}k^{*-1}g^{*2} + \frac{1}{2}(\sigma^{-2}\,\hat{\boldsymbol{y}} - g^*\boldsymbol{k}^{*-1})^{\mathsf{T}}\mathbf{M}^{-1}(\sigma^{-2}\,\hat{\boldsymbol{y}} - g^*\boldsymbol{k}^{*-1})} \\ &\propto \mathrm{e}^{-\frac{1}{2}(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})g^{*2} - \frac{g^*}{\sigma^2}\hat{\boldsymbol{y}}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1}} \end{aligned}$$

  - where we have used $\boldsymbol{b} = \sigma^{-2}\,\hat{\boldsymbol{y}} - g^*\boldsymbol{k}^{*-1}$
  - we have also thrown away terms not involving g^* (as they are constants that can be absorbed into the constant of proportionality)

* We can again *complete the square* this time for g^*

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2}(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})(g^* - \frac{\hat{\boldsymbol{y}}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1}}{\sigma^2(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})})}$$

  - We've again dropped terms that don't contain g^*

* But $g^* = f^* - m^*$ so we can write this as

$$\begin{aligned} p(f^*|\mathcal{D}) &\propto \mathrm{e}^{-\frac{1}{2}(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})(f^* - m^* - \frac{\hat{\boldsymbol{y}}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1}}{\sigma^2(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})})} \\ &\propto \mathcal{N}\left(f^* \middle| m^* + \frac{\hat{\boldsymbol{y}}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1}}{\sigma^2(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1})}, k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{k}^{*-1}\right) \end{aligned}$$

  - where we use the fact that we end up with a term that has the form of a normal distribution
  - because the posterior term is normalise, in fact it has to be exactly equal to this normal distribution

* Using $\mathbf{M} = \mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}$

$$p(f^*|\mathcal{D}) = \mathcal{N}\left(f^* \middle| m^* + \frac{\hat{\boldsymbol{y}}^{\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1}}{\sigma^2(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1})}, \frac{1}{k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1}}\right)$$

* In other words the expected mean value of $f^* = f(\boldsymbol{x}^*)$ is

$$\mathbb{E}[f(\boldsymbol{x}^*)] = m(\boldsymbol{x}^*) + \frac{\hat{\boldsymbol{y}}^{\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1}}{\sigma^2(k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1})}$$

(note that usually $m(\boldsymbol{x}^*) = 0$) and the expected variance in the value is

$$\mathbb{V}\mathrm{ar}[f(\boldsymbol{x}^*)] = \frac{1}{k^{*-1} - \boldsymbol{k}^{*-1\mathsf{T}}\left(\mathbf{k}^{-1} - \sigma^{-2}\mathbf{I}\right)^{-1}\boldsymbol{k}^{*-1}}$$

    - The result is pretty horrible because it involves inverting a matrix K with components $k(\boldsymbol{x}, \boldsymbol{y})$ evaluated at the set of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m, \boldsymbol{x}^*\}$
    - This is usually feasible but there is a simpler form that occurs because of identities involving inverse matrices
    - This is best obtained using a second way of deriving the results
    - I went to the pain of deriving the result this way because it is just a consequence of Bayes' rule

- In the second derivation we consider joint probability of the observations $\{y_i | i = 1, 2, \ldots, m\}$ and $f^* = f(\boldsymbol{x}^*)$
  - Even this derivation is a pain and you are **not** expected to know it
  - Now we assume $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (that is our observations have a normally distributed error) then

$$\mathbb{E}[y_i\, y_j] = \mathbb{E}[(f(\boldsymbol{x}_i) + \epsilon_i)(f(\boldsymbol{x}_j) + \epsilon_j)] = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sigma^2\, \delta_{ij}$$

    - since the observation error, ϵ_i, is assumed independent of the function value, $f(\boldsymbol{x}_i)$
    - here we assume $m(\boldsymbol{x}) = 0$ just to make life simple
  - Similarly

$$\mathbb{E}[y_i\, f^*] = \mathbb{E}[(f(\boldsymbol{x}_i) + \epsilon_i) f(\boldsymbol{x}^*)] = k(\boldsymbol{x}_i, \boldsymbol{x}^*)$$

    and

$$\mathbb{E}\left[(f^*)^2\right] = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$$

  - Thus

$$p(f^*, \boldsymbol{y}) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{y}\\ f_*\end{pmatrix} \middle| \boldsymbol{0}, \begin{pmatrix}\mathsf{K} + \sigma^2\,\mathsf{I} & \boldsymbol{k}^*\\ \boldsymbol{k}^{*\mathsf{T}} & k^*\end{pmatrix}\right)$$

    - where K is a matrix with components $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{k}^*$ is a vector with components $k(\boldsymbol{x}_i, \boldsymbol{x}^*)$ and $k^* = k(\boldsymbol{x}^*, \boldsymbol{x}^*)$
  - We now use an identity involving matrix inverses

$$\begin{pmatrix}\mathsf{K} + \sigma^2\,\mathsf{I} & \boldsymbol{k}^*\\ \boldsymbol{k}^{*\mathsf{T}} & k^*\end{pmatrix}^{-1} = \begin{pmatrix}\left(\mathsf{K} + \sigma^2\,\mathsf{I} - \frac{\boldsymbol{k}^{*\mathsf{T}}\boldsymbol{k}^*}{k^*}\right)^{-1} & -\frac{(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*}{k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*}\\ -\frac{\boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}}{k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*} & \frac{1}{k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*}\end{pmatrix}$$

    - these identities involving inverses of matrices seem to come from nowhere—they make working with normal distributions a real pain
    - you can prove the identity by multiply the right-hand side by $\begin{pmatrix}\mathsf{K} + \sigma^2\,\mathsf{I} & \boldsymbol{k}^*\\ \boldsymbol{k}^{*\mathsf{T}} & k^*\end{pmatrix}$ and showing you get the identity
    - if you are tempted to do this then set $\sigma^2 = 0$ and rename $\boldsymbol{k}^*$ and k^* so its easier to do the algebra—it is a tedious calculation (see exercises)
  - Now we use $p(f^*|\mathcal{D}) = p(f^*|\boldsymbol{y}) = p(f^*, \boldsymbol{y})/p(\boldsymbol{y})$
  - All you need to do is collect the terms in $p(f^*, \boldsymbol{y})$ involving f^*

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{f^{*2}}{2(k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*)} + \frac{f^*\,\boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{y}}{k^* + \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*}}$$

  - *Completing the square*

$$p(f^*|\mathcal{D}) \propto \mathrm{e}^{-\frac{1}{2(k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*)}\left(f^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K}+\sigma^2\,\mathsf{I})^{-1}\boldsymbol{y}\right)^2}$$

  - But as $p(f^*|\mathcal{D})$ must be normalised

$$p(f(\boldsymbol{x}^*)|\mathcal{D}) = \mathcal{N}\left(f(\boldsymbol{x}^*) \middle| \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K} + \sigma^2\,\mathsf{I})^{-1}\boldsymbol{y},\, k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathsf{K} + \sigma^2\,\mathsf{I})^{-1}\boldsymbol{k}^*\right)$$

- That is the mean value of f at a point $\boldsymbol{x}^*$ is

$$\mathbb{E}[f(\boldsymbol{x}^*)] = m(\boldsymbol{x}^*) + \boldsymbol{k}^{*\mathsf{T}}(\mathbf{K} + \sigma^2\,\mathbf{I})^{-1}\boldsymbol{y}$$

  - I've re-instated $m(\boldsymbol{x}^*)$ (even though it is usually taken to be zero)
- The variance is given by

$$\mathbb{V}\mathrm{ar}[f(\boldsymbol{x}^*)] = k^* - \boldsymbol{k}^{*\mathsf{T}}(\mathbf{K} + \sigma^2\,\mathbf{I})^{-1}\boldsymbol{k}^*$$

- This may look completely different from the previous expression but surprisingly it is identical (you can demonstrate this numerically)
- The advantage over the previous methods is that we only need to do one matrix inversion (and we get a slightly easier expression)
- Although the mathematics to get here was horrible the expression is extremely easy to compute
- One of the advantages of Bayesian inference is it provides an estimate of its own uncertainty (in this case $\mathbb{V}\mathrm{ar}[f(\boldsymbol{x}^*)]$)

2.4 Learning Hyperparameters

- **Hyperparameters in the Bayesian Framework**
  - As I have said many times to get machine learning to work it is vital to choose the hyperparameter properly (we will see this is true for GPs)
  - In the Bayesian framework if $\boldsymbol{\phi}$ are hyperparameters then Bayes' rules says

$$p(\boldsymbol{x}|\mathcal{D}, \boldsymbol{\phi}) = \frac{p(\mathcal{D}|\boldsymbol{x}, \boldsymbol{\phi})\,p(\boldsymbol{x}|\boldsymbol{\phi})}{p(\mathcal{D}|\boldsymbol{\phi})}$$

    - this is just Bayes' rule with everything conditioned on the hyperparameters $\boldsymbol{\phi}$
  - We can select hyperparameters by considering the **evidence**, $p(\mathcal{D}|\boldsymbol{\phi})$
    - when the hyperparameters are viewed as different models this is known as **model selection**
    - this is also called the evidence framework
    - We can view the relatively likelihood of one model, $\boldsymbol{\phi}_1$ compare to a second, $\boldsymbol{\phi}_2$ by examining

$$\frac{p(\mathcal{D}|\boldsymbol{\phi}_1)}{p(\mathcal{D}|\boldsymbol{\phi}_2)}$$

  - If we want to be hyper-Bayesian then we can put a prior, $p(\boldsymbol{\phi})$, over our hyperparameters and the calculate the joint posterior for the parameters of the likelihood $\boldsymbol{\theta}$ and the hyper-parameters $\boldsymbol{\phi}$

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

    - to compute the posterior we are interested in we marginalise out the hyperparameters

$$p(\boldsymbol{\theta}|\mathcal{D}) = \int p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D})\,\mathrm{d}\boldsymbol{\phi}$$

    - Often this integral is not computable in closed form and we are forced to estimate it using Monte-Carlo techniques
- **Hyperparameter for Gaussian Processes**
  - For Gaussian Processes this means choosing the right kernel function

- Here we are in a better positions than in SVMs in that the kernel represents the covariance function
- Given data we could estimate the kernel function using

$$\mathbb{E}[y_i, y_j] = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sigma^2 \delta_{ij}$$

- It would be an inverse problem to estimate $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- Often we start with a guess of the form of the kernel
- A very common kernel function is the Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\ell^2}}$$

- Now ℓ is a *scale parameter* that has to be determined
- However, a great advantage of Gaussian Processes is that we can compute the evidence in closed form

$$\log(p(\mathcal{D}|\boldsymbol{\theta})) = -\frac{1}{2}\boldsymbol{y}^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\boldsymbol{y} - \frac{1}{2}\log\big(|\mathbf{K} + \sigma^2\mathbf{I}|\big) - \frac{m}{2}\log(2\,\pi)$$

  * This allows us to rapidly compare different hyperparameters (kernels and noise level)
  * We are not doing this in a fully Bayesian way (we are maximising our evidence rather than updating a distribution for the hyperparameters) therefore we could overfit
  * However, usually we have relatively few hyperparameters so overfitting is less severe

3 Exercises

3.1 Working with Gaussians

This is optional. Learning to work with Gaussians allows you to understand a lot of machine learning techniques, but it is not going to be examined.

1. Integrals involving Gaussians can be done in closed form. This means that they are heavily used in machine learning. It does, however, take a lot of practice to learn how to do these integrals. The starting point is the integral

$$I_1 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \sqrt{2\,\pi}.$$

   This integral isn't easy. The trick is to turn this into a two dimensional integral and perform the integral in polar coordinates. Have a go.

2. Show by a change of variables

$$I_2 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/(2\,\sigma^2)}\,\mathrm{d}x = \sqrt{2\,\pi}\,\sigma.$$

3. Show that

$$I_3 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2 + a\,x}\,\frac{\mathrm{d}x}{\sqrt{2\,\pi}} = \mathrm{e}^{a^2/2}$$

   This involves completing the square and using a change of variables. It is probably the most frequently used trick working with Gaussian integrals.

4. Show that

$$I_4 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}\|^2} \mathrm{d}\boldsymbol{x} = (2\,\pi)^{\frac{n}{2}}$$

   Often we don't bother writing all the integral signs (one will do)

5. By writing $\mathbf{M} = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{\mathsf{T}}$ show that

$$I_5 = \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\mathbf{M}\boldsymbol{x}} \mathrm{d}\boldsymbol{x} = (2\,\pi)^{\frac{n}{2}}\,\frac{1}{\sqrt{|\mathbf{M}|}}$$

   where $|\mathbf{M}|$ is the determinant of $\mathbf{M}$. Note that $|\mathbf{M}| = |\mathbf{\Lambda}| = \prod_i \lambda_i$.

6. Show that

$$I_6 = \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{x}} \mathrm{d}\boldsymbol{x} = (2\,\pi)^{\frac{n}{2}}\,\sqrt{|\mathbf{M}|}$$

3.2 Working with Matrix Inverses

- Modelling using normal (aka Gaussian) distributions is very powerful because we can usually computer everything in closed form—integrals involving Gaussians are always doable
- However, there is a twist which makes it extremely painful
  - The normal distribution uses the inverse of the covariance matrix
  - The inverse of matrices are a pain to work with
- To get nice results we often consider partition matrices into four blocks—the inverse of the full matrix can be written in terms of the inverse of each block
- Let's look at a very simple example
  - Show that

$$\begin{pmatrix} \mathbf{K} & \boldsymbol{\ell} \\ \boldsymbol{\ell}^{\mathsf{T}} & m \end{pmatrix}^{-1} = \begin{pmatrix} \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^{\mathsf{T}}}{m}\right)^{-1} & -\frac{\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} \\ -\frac{\boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} & \frac{1}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} \end{pmatrix}$$

    - $\mathbf{K}$ is a symmetric matrix, ℓ is a vector and m a scalar
    - Similar identities exist when ℓ and m are matrices, but let's keep thing simple
  - Now we can show this by multiply both sides by $\begin{pmatrix} \mathbf{K} & \boldsymbol{\ell} \\ \boldsymbol{\ell}^{\mathsf{T}} & m \end{pmatrix}$ (the algebra is easier if we multiply on the right)
  - This is still very fiddly to show, so break it down
    - Show that

$$\begin{pmatrix} \mathbf{A} & \boldsymbol{b} \\ \boldsymbol{b}^{\mathsf{T}} & c \end{pmatrix} \begin{pmatrix} \mathbf{K} & \boldsymbol{\ell} \\ \boldsymbol{\ell}^{\mathsf{T}} & m \end{pmatrix} = \begin{pmatrix} \mathbf{A}\,\mathbf{K} + \boldsymbol{b}\,\boldsymbol{\ell}^{\mathsf{T}} & \mathbf{A}\,\boldsymbol{\ell} + m\,\boldsymbol{b} \\ \boldsymbol{b}^{\mathsf{T}}\mathbf{K} + c\,\boldsymbol{\ell}^{\mathsf{T}} & \boldsymbol{b}^{\mathsf{T}}\boldsymbol{\ell} + c\,m \end{pmatrix}$$

    - Now you have to show that when $\mathbf{A} = \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^{\mathsf{T}}}{m}\right)^{-1}$, $\boldsymbol{b} = -\frac{\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}$ and $c = \frac{1}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}$ that
      1. $\mathbf{A}\,\mathbf{K} + \boldsymbol{b}\,\boldsymbol{\ell}^{\mathsf{T}} = \mathbf{I}$
      2. $\mathbf{A}\,\boldsymbol{\ell} + m\,\boldsymbol{b} = \mathbf{0}$
      3. $\boldsymbol{b}^{\mathsf{T}}\mathbf{K} + c\,\boldsymbol{\ell}^{\mathsf{T}} = \mathbf{0}^{\mathsf{T}}$
      4. $\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\ell} + c\,m = 1$
    - To show 1. and 2. take out $\mathbf{A}$ as a common factor using $\mathbf{I} = \mathbf{A}\,\mathbf{A}^{-1}$ (answer below)

4 Experiments

4.1 Generate Samples from a Gaussian Processes

- To generate a Gaussian Process, $\mathcal{GP}(0, k)$, with $m(\boldsymbol{x}) = 0$ and covariance

$$k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\,\ell^2}}$$

  a points $\{\boldsymbol{x}_i | i = 1, 2, \ldots, n\}$

  1. We use first compute the covariance matrix $\mathbf{K}$ with elements $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$
  2. Compute the *Cholesky decompostion* $\mathbf{L}$ such that $\mathbf{K} = \mathbf{L}\,\mathbf{L}^\mathsf{T}$
     - For any positive definite matrices we can always compute a Cholesky decomposition (although beware, due to numerical rounding some matrices that should be positive definite, appear not to have Cholesky decompositions—you can $\epsilon\,\mathbf{I}$ to your matrix to help)
     - The Cholesky decomposition is a lower diagonal matrix
     - It is used to efficiently solve linear problems involving positive definite matrices (it is more efficient and more stable than Gaussian elimination)
  3. Now let $\boldsymbol{f} = \mathbf{L}\,\boldsymbol{\eta}$ where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
     - We note that

$$\mathbb{E}\left[\boldsymbol{f}\,\boldsymbol{f}^\mathsf{T}\right] = \mathbb{E}\left[\mathbf{L}\,\boldsymbol{\eta}\,\boldsymbol{\eta}^\mathsf{T}\mathbf{L}^\mathsf{T}\right] = \mathbf{L}\,\mathbb{E}\left[\boldsymbol{\eta}\,\boldsymbol{\eta}^\mathsf{T}\right]\mathbf{L}^\mathsf{T} = \mathbf{L}\,\mathbf{I}\,\mathbf{L}^\mathsf{T} = \mathbf{L}\,\mathbf{L}^\mathsf{T} = \mathbf{K}$$

     - Thus $\boldsymbol{f}$ can be viewed as samples from a Gaussian Process
     - Draw many different samples and plot them
  4. Experiment with different values of ℓ

1-d Gaussian Processes

```
x = [-10:0.2:10];    % define some points
n = length(x)
l = 1.0                      % define length scale
K = zeros(n,n);          % define holder
for i = 1:n
  K(i,i) = 1;
  for j = 1:i
    K(j,i) = K(i,j) = exp(-(x(i)-x(j))^2/(2*l)); % define covariance matrix
  endfor
endfor
L = chol(K+0.0001*eye(n));  % cheat to persuade octave K is positive definite
f1 = L*randn(n,1);
f2 = L*randn(n,1);
f3 = L*randn(n,1);
plot(x,f1,x,f2,x,f3)
```

2-d Gaussian Process

```
range = [-4:0.2:4];    % define some points
n = length(range)
X = zeros(n*n,2);
for i = 1:n
  for j = 1:n
     X((i-1)*n+j,1) = range(i);
```

```
    X((i-1)*n+j,2) = range(j);
  endfor
endfor
l = 1.0                          % define length scale
N =length(X)
K = zeros(N,N);
for i = 1:N
  K(i,i) = 1;
  for j = 1:i
    K(j,i) = K(i,j) = exp(-norm(X(i,:)-X(j,:))^2/(2*l)); % define covariance matrix
  endfor
endfor

L = chol(K+0.0001*eye(N));  % cheat to persuade octave K is positive definite
f = L*randn(N,1);
fm = reshape(f,n,n);
[xx, yy] = meshgrid(range, range);
mesh(xx, yy, fm);
xlabel ("x");
ylabel ("y");
zlabel ("f(x,y)");
title ("2-d Gaussian Process");
```

5 Answers

5.1 Working with Gaussians

1. The Gaussian Integral.

   - We consider the two dimensional integral

$$I_1^2 = \left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x\right)^2 \overset{\text{(a)}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2-y^2/2}\,\mathrm{d}x\,\mathrm{d}y$$
$$\overset{\text{(b)}}{=} \int_0^{\infty}\int_0^{2\pi} r\,\mathrm{e}^{-r^2}\mathrm{d}r\,\mathrm{d}\theta \overset{\text{(c)}}{=} \int_0^{\infty} r\,\mathrm{e}^{-r^2}\mathrm{d}r \int_0^{2\pi}\mathrm{d}\theta = 2\,\pi$$

     (a) Changing the name of the dummy index from x to y
     (b) Using the change of variable $x = r\,\cos(\theta)$ and $y = r\,\sin(\theta)$ so that $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\,\mathrm{d}\theta$
     (c) Reordering the integrals and using $\int_0^{\infty} r\,\mathrm{e}^{-r^2}\mathrm{d}r = 1$ (which you can prove by a change of variables $u = r^2/2$ so that $r\,\mathrm{d}r = \mathrm{d}u$)

     Therefore $I_1 = \sqrt{2\,\pi}$.

2.
$$I_2 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/(2\,\sigma^2)}\,\mathrm{d}x \overset{\text{(a)}}{=} \sigma \int_{-\infty}^{\infty} \mathrm{e}^{-u^2/2}\,\mathrm{d}u \overset{\text{(b)}}{=} \sqrt{2\,\pi}\,\sigma.$$

   (a) Using $u = x/\sigma$ so that $\mathrm{d}x = \sigma\,\mathrm{d}u$
   (b) Using I_1

3.
$$I_2 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2+a\,x}\,\frac{\mathrm{d}x}{\sqrt{2\,\pi}} \overset{\text{(a)}}{=} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(x-a)^2+\frac{a^2}{2}}\,\frac{\mathrm{d}x}{\sqrt{2\,\pi}}$$
$$\overset{\text{(b)}}{=} \mathrm{e}^{a^2/2}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(x-a)^2}\,\frac{\mathrm{d}x}{\sqrt{2\,\pi}} \overset{\text{(c)}}{=} \mathrm{e}^{a^2/2}$$

(a) Using $-x^2/2 + a\,x = -\frac{1}{2}(x-a)^2 + \frac{a^2}{2}$

(b) Factoring out $\mathrm{e}^{a^2/2}$

(c) Making a change of variables $u = x - a$ and using integral I_1

4.

$$I_4 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}\|^2} \mathrm{d}\boldsymbol{x} \overset{\text{(a)}}{=} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \prod_{i=1}^{n} \mathrm{d}x_i$$
$$\overset{\text{(b)}}{=} \prod_{i=1}^{n} \left(\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x_i^2}{2}} \, \mathrm{d}x_i \right) \overset{\text{(c)}}{=} (2\,\pi)^{\frac{n}{2}}$$

(a) Using $\|\boldsymbol{x}\|^2 = \sum_{i=1}^{n} x_i^2$

(b) Using $\exp(\sum_i a_i) = \prod_i \mathrm{e}^{a_i}$ and factoring out the different terms

(c) Using integral I_1

5.

$$I_5 = \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\mathbf{M}\boldsymbol{x}} \mathrm{d}\boldsymbol{x} \overset{\text{(a)}}{=} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}\boldsymbol{x}} \mathrm{d}\boldsymbol{x}$$
$$\overset{\text{(b)}}{=} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{u}} \, \mathrm{d}\boldsymbol{u} \overset{\text{(c)}}{=} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\sum_i \lambda_i u_i^2} \, \mathrm{d}\boldsymbol{u}$$
$$\overset{\text{(d)}}{=} \prod_i \left(\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\lambda_i u_i^2} \, \mathrm{d}u_i \right) \overset{\text{(e)}}{=} \prod_i \left(\sqrt{\frac{2\pi}{\lambda_i}} \right) \overset{\text{(f)}}{=} (2\,\pi)^{n/2} \frac{1}{\sqrt{\prod_i \lambda_i}} \overset{\text{(g)}}{=} \frac{(2\,\pi)^{\frac{n}{2}}}{\sqrt{|\mathbf{M}|}}$$

(a) Using $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}$ where

(b) Writing $\boldsymbol{u} = \mathbf{V}^\mathsf{T}\boldsymbol{x}$. Note that $\mathrm{d}\boldsymbol{u} = |\mathbf{V}^\mathsf{T}| \, \mathrm{d}\boldsymbol{x}$, but $\mathbf{V}$ is an orthogonal matrix so $|\mathbf{V}^\mathsf{T}| = 1$. So $\mathrm{d}\boldsymbol{u} = \mathrm{d}\boldsymbol{x}$.

(c) $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\Lambda_{ii} = \lambda_i$

(d) Rearranging

(e) Using integral I_2 with $\sigma = 1/\sqrt{\lambda_i}$

(f) Rearranging

(g) Using $\prod_i \lambda_i = |\mathbf{M}|$.

6. This follows using I_5 because $|\mathbf{M}^{-1}| = \frac{1}{|\mathbf{M}|}$.

5.2 Working with Matrix Inverses

- Using $\mathbf{A} = \left(\mathbf{K} - \frac{\boldsymbol{\ell}\boldsymbol{\ell}^\mathsf{T}}{m}\right)^{-1}$, $\boldsymbol{b} = -\frac{\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}$ and $c = \frac{1}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}$, we show the four identities

  1.

$$\begin{aligned}
\mathbf{A}\,\mathbf{K} + \boldsymbol{b}\,\boldsymbol{\ell}^\mathsf{T} &= \mathbf{A}\,\left(\mathbf{K} + \mathbf{A}^{-1}\boldsymbol{b}\,\boldsymbol{\ell}^\mathsf{T}\right) \\
&= \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right)^{-1} \left(\mathbf{K} + \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right) \frac{(-\mathbf{K}^{-1}\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T})}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}\right) \\
&= \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right)^{-1} \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}} + \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}/m}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}\right) \\
&= \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right)^{-1} \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m} \frac{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^\mathsf{T}\mathbf{K}^{-1}\boldsymbol{\ell}}\right) \\
&= \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right)^{-1} \left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^\mathsf{T}}{m}\right) = \mathbf{I}
\end{aligned}$$

2.

$$\begin{aligned}\mathbf{A}\,\boldsymbol{\ell} + m\,\boldsymbol{b} &= \mathbf{A}\left(\boldsymbol{\ell} + m\,\mathbf{A}^{-1}\boldsymbol{b}\right)\\ &= \mathbf{A}\left(\boldsymbol{\ell} - m\left(\mathbf{K} - \frac{\boldsymbol{\ell}\,\boldsymbol{\ell}^{\mathsf{T}}}{m}\right)\frac{\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}\right)\\ &= \mathbf{A}\left(\boldsymbol{\ell} - \frac{m\,\boldsymbol{\ell}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}(1 - \frac{\boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}{m})\right)\\ &= \mathbf{A}\,(\boldsymbol{\ell} - \boldsymbol{\ell}) = \mathbf{0}\end{aligned}$$

3.

$$\boldsymbol{b}^{\mathsf{T}}\mathbf{K} + c\,\boldsymbol{\ell}^{\mathsf{T}} = -\frac{\boldsymbol{\ell}^{\mathsf{T}}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} + \frac{\boldsymbol{\ell}^{\mathsf{T}}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} = \mathbf{0}^{\mathsf{T}}$$

4.

$$\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\ell} + c\,m = -\frac{\boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} + \frac{m}{m - \boldsymbol{\ell}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\ell}} = 1$$