Inference

Joaquín Rapela

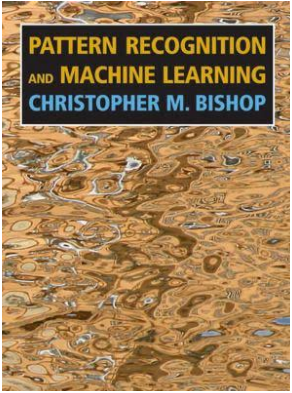
Gatsby Computational Neuroscience Unit
University College London

July 18, 2023

Contents

Main reference

I will mainly follow chapters two *Probability distributions* and three *Linear models for regression* from Bishop (2016).



Contents

The Gaussian distribution

- One-dimensional

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

- D-dimensional

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}\Sigma^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The Gaussian is the maximum entropy distribution (Cover and Thomas, 1991)

Definition 1 (Differential entropy)

The differential entropy $h(X)$ of a continuous random variable X with a density $f(x)$ is defined as

$$h(X) = -\int_S f(X) \log f(x)\ dx$$

where S is the support set of the random variable.

Theorem 1 (The Gaussian is the maximum entropy distribution)

Let the random vector $X \in \mathbb{R}^n$ have zero mean and covariance K. Then $h(X) \leq \frac{1}{2}\log(2\pi e)^n|K|$, with equality if $X \sim \mathcal{N}(0, K)$.

The central limit theorem (Papoulis and Pillai, 2002)

Theorem 2 (The central limit theorem)

Given n independent and identically distributed random vectors $\mathbf{X}_i$, with mean vector $\boldsymbol{\mu} = E\{\mathbf{X}_i\}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \to \mathcal{N}(0, \Sigma)$$

with convergence in distribution.

Very useful properties of the Gaussian distribution (Bishop, 2016)

Theorem 3 (Marginals and conditionals of Gaussians are Gaussians)

Given $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$ *such that*

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \middle| \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right) \\ &= \mathcal{N}\left(\mathbf{x} \middle| \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}^{-1}\right) \end{aligned}$$

Then

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a \middle| \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b), \Lambda_{aa}^{-1}\right) \quad (1)$$

$$= \mathcal{N}\left(\mathbf{x}_a \middle| \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right) \quad (2)$$

$$p(\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_b \middle| \boldsymbol{\mu}_b, \Sigma_{bb}\right) \quad (3)$$

Very useful properties of the Gaussian distribution (Bishop, 2016)

Theorem 4 (Marginals and conditionals of the linear Gaussian model)

Given the linear Gaussian model

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Lambda^{-1})$$
$$p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|A\boldsymbol{\mu} + \mathbf{b}, L^{-1})$$

Then

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|A\boldsymbol{\mu} + \mathbf{b}, L^{-1} + A\Lambda^{-1}A^{\intercal}) \tag{4}$$
$$p(\mathbf{x}|\mathbf{t}) = \mathcal{N}(\mathbf{x}|\Sigma\{A^{\intercal}L(\mathbf{t} - \mathbf{b}) + \Sigma\boldsymbol{\mu}\}, \Sigma) \tag{5}$$

where

$$\Sigma = (\Lambda + A^{\intercal}LA)^{-1}$$

Very useful properties of the Gaussian distribution (Bishop, 2016)

The conditional, $p(\mathbf{x}|\mathbf{t})$, of the linear Gaussian model is the fundamental result used in the derivation of

1. Bayesian linear regression (Bishop, 2016),
2. Gaussian process regression (Williams and Rasmussen, 2006),
3. Gaussian process factor analysis (Yu et al., 2009),
4. linear dynamical systems (Durbin and Koopman, 2012).

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Claim 1 (Quadratic form of Gaussian log pdf)

$p(\mathbf{x})$ is a Gaussian pdf with mean $\boldsymbol{\mu}$ and precision matrix Λ if and only if $\int p(\mathbf{x})d\mathbf{x} = 1$ and

$$\log p(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K \tag{6}$$

where K is a constant that does not depend on $\mathbf{x}$.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Claim 1.

$\rightarrow)$

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}\Lambda^{-\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \Lambda (\mathbf{x}-\boldsymbol{\mu})\right\} \\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \Lambda (\mathbf{x}-\boldsymbol{\mu}) - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K
\end{aligned}
$$

with $K = -\frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})$.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Claim 1.

$\leftarrow$)

$$
\begin{aligned}
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) + K \\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}^\intercal \Lambda \mathbf{x} - 2\mathbf{x}^\intercal \Lambda \boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&\qquad + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \Lambda (\mathbf{x} - \boldsymbol{\mu}) - \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&\qquad + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
&= \log N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) + K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}}) \\
p(\mathbf{x}) &= N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) \exp\left(K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)
\end{aligned}
\tag{7}
$$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Claim 1.

$\leftarrow$) cont

$$
\begin{aligned}
1 &= \int p(\mathbf{x}) d\mathbf{x} \\
&= \int N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) \exp\left(K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right) d\mathbf{x} \\
&= \exp\left(K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right) \int N(\mathbf{x}|\boldsymbol{\mu}, \Lambda) d\mathbf{x} \\
&= \exp\left(K + \frac{1}{2}\boldsymbol{\mu}^\intercal \Lambda \boldsymbol{\mu} + \log((2\pi)^{D/2}\Lambda^{-\frac{1}{2}})\right)
\end{aligned}
$$

From Eq. 7 then $p(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu}, \Lambda)$.

□

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Theorem 3, Eq. 1.

$$p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_b)} = \frac{p(\mathbf{x})}{p(\mathbf{x}_b)}$$
$$\log p(\mathbf{x}_a|\mathbf{x}_b) = \log p(\mathbf{x}) - \log p(\mathbf{x}_b) = \log p(\mathbf{x}) + K$$

Therefore, the terms of $\log p(\mathbf{x}_a|\mathbf{x}_b)$ that depend on $\mathbf{x}_a$ are those of $\log p(\mathbf{x})$.
Steps for the proof:

1. isolate the terms of $\log p(\mathbf{x})$ that depend on $\mathbf{x}_a$,
2. notice that these term has the quadratic form of Claim 1, therefore $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian,
3. identify $\boldsymbol{\mu}$ and Λ in this quadratic form.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 1)

Proof of Theorem 3, Eq. 1.

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu})\right) \\
\log p(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\Lambda(\mathbf{x}-\boldsymbol{\mu}) + K_1 \\
&= -\frac{1}{2}[(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal, (\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal] \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{x}_a-\boldsymbol{\mu}_a \\ \mathbf{x}_b-\boldsymbol{\mu}_b \end{bmatrix} + K_1 \\
&= -\frac{1}{2}\left\{(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\Lambda_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) + 2(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right. \\
&\quad \left. +(\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal\Lambda_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right\} + K_1 \\
&= -\frac{1}{2}\left\{\mathbf{x}_a^\intercal\Lambda_{aa}\mathbf{x}_a - 2\mathbf{x}_a^\intercal(\Lambda_{aa}\boldsymbol{\mu}_a - \Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b))\right\} + K_2 \\
&= -\frac{1}{2}\left\{\mathbf{x}_a^\intercal\Lambda_{aa}\mathbf{x}_a - 2\mathbf{x}_a^\intercal\Lambda_{aa}(\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b))\right\} + K_2
\end{aligned}
$$

Comparing the last equation with Eq. 6 we see that $\Lambda = \Lambda_{aa}$, $\boldsymbol{\mu} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$ and conclude that $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b), \Lambda_{aa})$ $\square$

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 2)

Claim 2 (Inverse of a partitioned matrix)

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \tag{8}$$

where

$$M = (A - BD^{-1}C)^{-1}$$

Proof.

Exercise. Hint: verify that the multiplication of the inverse of the matrix in the right hand side of Eq. 8 with the matrix in the left hand side of the same equation is the identity matrix.

Proof: the conditional of a Gaussian is a Gaussian (Theorem 3, Eq. 2)

Proof of Theorem 3, Eq. 2.

Using the definition

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

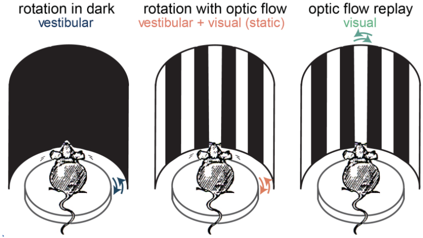
and using Eq. 8, we obtain

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$
$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

Replacing the above equations in Eq. 1 we obtain Eq. 2. $\square$

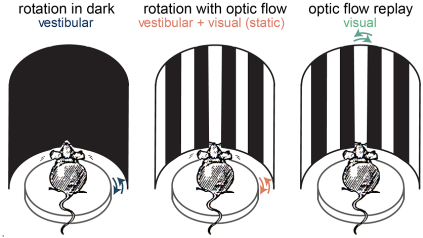
Contents

Linear regression example

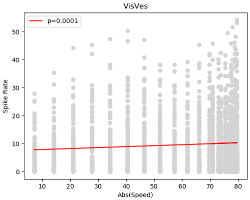


Keshavarzi et al., 2021

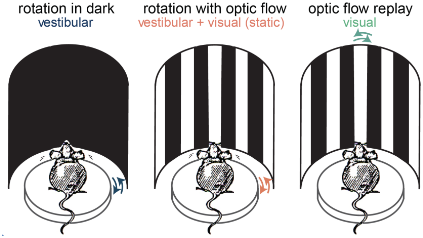
Linear regression example



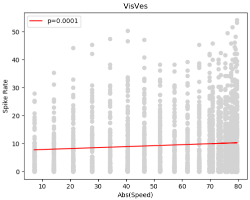
Keshavarzi et al., 2021



Linear regression example



Keshavarzi et al., 2021



Is there a linear relation between the speed of rotation and the firing rate of visual cells?

Linear regression model

simple linear regression model

$$\begin{aligned} y(x_i, \mathbf{w}) &= w_0 + w_1 x_i = [1, x_i] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = [\phi_0(x_i), \phi_1(x_i)] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \\ &= \phi(x_i)^\mathsf{T} \mathbf{w} \end{aligned}$$

polynomial regression model

$$\begin{aligned} y(x_i, \mathbf{w}) &= w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 = [1, x_i, x_i^2, x_i^3] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} \\ &= [\phi_0(x_i), \phi_1(x_i), \phi_2(x_i), \phi_3(x_i)] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix} = \phi(x_i)^\mathsf{T} \mathbf{w} \end{aligned}$$

basis functions linear regression model

$$y(x_i, \mathbf{w}) = \phi(x_i)^\mathsf{T} \mathbf{w} = \sum_{j=1}^{M} w_j \phi_j(x_i)$$

Linear regression model

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \begin{bmatrix} y(x_1, \mathbf{w}) \\ y(x_2, \mathbf{w}) \\ \dots \\ y(x_N, \mathbf{w}) \end{bmatrix} = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_M(x_2) \\ \vdots & \vdots & \dots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \dots & \phi_M(x_N) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix}$$

$$= \mathbf{\Phi w}$$

where $\mathbf{y}(\mathbf{x}, \mathbf{w}) \in \mathbb{R}^N, \mathbf{\Phi} \in \mathbb{R}^{N \times M}, \mathbf{w} \in \mathbb{R}^M$.

Basis functions for regression

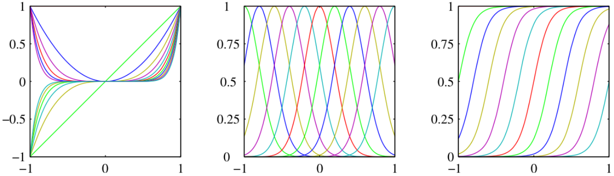


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

Bishop (2016))

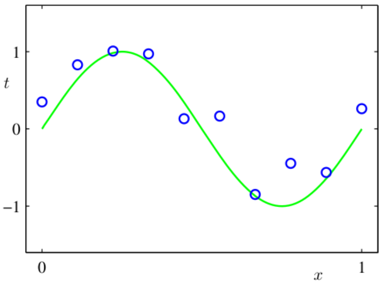
polynomial $\phi_i(x) = x^i$

Gaussian $\phi_i(x) = \exp(-\frac{(x-\mu_i)^2}{2\sigma^2})$

sigmoidal $\phi_i(x) = \frac{1}{1+\exp(-\frac{x-\mu_i}{\sigma^2})}$

Example dataset

Figure 1.2 Plot of a training data set of $N = 10$ points, shown as blue circles, each comprising an observation of the input variable x along with the corresponding target variable t. The green curve shows the function $\sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve.



Bishop (2016)

Outline

Least-squares estimation of model parameters (Trefethen and Bau III, 1997)

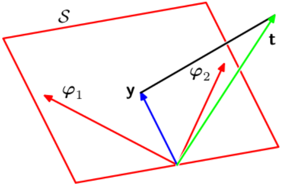
Definition 2 (Least-squares problem)

Given $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, $N \geq M$, $\mathbf{t} \in \mathbb{R}^N$, find $\mathbf{w} \in \mathbb{R}^M$ such that $E_{LS}(\mathbf{w}) = ||\mathbf{t} - \mathbf{\Phi}\boldsymbol{w}||_2$ is minimised.

Theorem 5 (Least-squares solution)

Let $\mathbf{\Phi} \in \mathbb{R}^{N \times M}(N \geq M)$ and $\mathbf{t} \in \mathbb{R}^N$ be given. A vector $\mathbf{w} \in \mathbb{R}^M$ minimises $||\mathbf{r}||_2 = ||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||_2$, thereby solving the least-squares problem, if and only if $\mathbf{r} \perp range(\mathbf{\Phi})$, that is, $\mathbf{\Phi}^\intercal \mathbf{r} = 0$, or equivalently, $\mathbf{\Phi}^\intercal \mathbf{\Phi} \mathbf{w} = \mathbf{\Phi}^\intercal \mathbf{t}$, or again equivalently, $P\mathbf{t} = \mathbf{\Phi}\mathbf{w}$.

Figure 3.2 Geometrical interpretation of the least-squares solution, in an N-dimensional space whose axes are the values of $t_1, \ldots, t_N$. The least-squares regression function is obtained by finding the orthogonal projection of the data vector $\mathbf{t}$ onto the subspace spanned by the basis functions $\phi_j(\mathbf{x})$ in which each basis function is viewed as a vector $\boldsymbol{\varphi}_j$ of length N with elements $\phi_j(\mathbf{x}_n)$.



Bishop (2016)

Instruction to run notebooks in Google Colab

1. open a notebook from here
2. replace **github.com** by **githubtocolab.com** in the URL
3. insert a cell at the beginning of the notebook with the following content

```
!git clone https://github.com/joacorapela/gcnuBridging2023.git
%cd gcnuBridging2023
!pip install -e .
```

4. from the menu **Runtime** select **Run all**.

Code for least-squares estimation of model parameters

- overfitting
- cross validation
- larger datasets allow more complex models

Regularised least-squares estimation of model parameters

To cope with the overfitting of least squares, we can add to the least squares optimisation criterion a term that enforces coefficients to be zero. The regularised least-squares optimisation criterion becomes:

$$E_{RLS}(\mathbf{w}) = ||\mathbf{t} - \mathbf{\Phi w}||_2^2 + \lambda ||\mathbf{w}||_2^2$$

where λ is the regularisation parameter that weights the strength of the regularisation.

Regularised least-squares estimation of model parameters

Claim 3 (Regularized least-squares estimate)

$$\mathbf{w}_{RLS} = \arg\min_{\mathbf{w}} E_{RLS}(\mathbf{w}) = \arg\min_{\mathbf{w}} ||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2 = (\lambda\mathbf{I} + \mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal\mathbf{t}$$

Proof.

Since $E_{RLS}(\mathbf{w})$ is a polynomial of order two on the elements of $\mathbf{w}$ (i.e., a quadratic form), we can use the *Completing the Squares* technique below to find its minimum.

$$\begin{aligned}
\boldsymbol{\mu} &= \arg\max_{\mathbf{w}} \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\Sigma) = \arg\max_{\mathbf{w}} \log\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\Sigma) \\
&= \arg\max_{\mathbf{w}}\{K - \frac{1}{2}(-2\boldsymbol{\mu}^\intercal\Sigma^{-1}\mathbf{w} + \mathbf{w}\Sigma^{-1}\mathbf{w})\} \qquad (9)\\
&= \arg\min_{\mathbf{w}}\{-K + \frac{1}{2}(-2\boldsymbol{\mu}^\intercal\Sigma^{-1}\mathbf{w} + \mathbf{w}\Sigma^{-1}\mathbf{w})\} \\
&= \arg\min_{\mathbf{w}}\{K_1 - 2\boldsymbol{\mu}^\intercal\Sigma^{-1}\mathbf{w} + \mathbf{w}\Sigma^{-1}\mathbf{w}\} \qquad (10)
\end{aligned}$$

Note: Eq. 9 uses Eq. 6.
To find the minimum of a quadratic form, we write it in the form of the terms inside the curly brackets of Eq. 10, and the term corresponding to $\boldsymbol{\mu}$ will be the minimum.

Regularised least-squares estimation of model parameters

Proof.

Let's write E_{RLS} in the form of the terms inside the curly brackets of Eq. 10.

$$\begin{aligned} E_{RLS} &= ||\mathbf{t} - \mathbf{\Phi w}||_2^2 + \lambda ||\mathbf{w}||_2^2 = (\mathbf{t} - \mathbf{\Phi w})^\intercal (\mathbf{t} - \mathbf{\Phi w}) + \lambda \mathbf{w}^\intercal \mathbf{w} \\ &= \mathbf{t}^\intercal \mathbf{t} - 2\mathbf{t}^\intercal \mathbf{\Phi w} + \mathbf{w}^\intercal \mathbf{\Phi}^\intercal \mathbf{\Phi w} + \lambda \mathbf{w}^\intercal \mathbf{w} \\ &= \mathbf{t}^\intercal \mathbf{t} - 2\mathbf{t}^\intercal \mathbf{\Phi w} + \mathbf{w}^\intercal (\mathbf{\Phi}^\intercal \mathbf{\Phi}^\intercal + \lambda \mathbf{I}_M) \mathbf{w} \end{aligned}$$

Calling

$$\begin{aligned} \Sigma^{-1} &= \mathbf{\Phi}^\intercal \mathbf{\Phi}^\intercal + \lambda \mathbf{I}_M \\ \boldsymbol{\mu}^\intercal \Sigma^{-1} &= \mathbf{t}^\intercal \mathbf{\Phi} \text{ or } \boldsymbol{\mu}^\intercal = \mathbf{t}^\intercal \mathbf{\Phi} \Sigma \text{ or } \boldsymbol{\mu} = \Sigma \mathbf{\Phi}^\intercal \mathbf{t} = \left(\mathbf{\Phi}^\intercal \mathbf{\Phi}^\intercal + \lambda \mathbf{I}_M\right)^{-1} \mathbf{\Phi}^\intercal \mathbf{t} \end{aligned}$$

we can express

$$E_{RLS} = K + 2\boldsymbol{\mu}^\intercal \Sigma^{-1} \mathbf{w} + \mathbf{w} \Sigma^{-1} \mathbf{w}$$

Then

$$\mathbf{w}_{RLS} = \arg\min_{\mathbf{w}} E_{RLS}(\mathbf{w}) = \boldsymbol{\mu} = \left(\mathbf{\Phi}^\intercal \mathbf{\Phi}^\intercal + \lambda \mathbf{I}_M\right)^{-1} \mathbf{\Phi}^\intercal \mathbf{t}$$

□

Code for regularised least-squares estimation of model parameters

- control of overfitting

Outline

Maximum-likelihood estimation of model parameters

Definition 3 (Likelihood function)

For a statistical model characterised by a probability density function $p(\mathbf{x}|\theta)$ *(or probability mass function* $P_\theta(X = \mathbf{x})$*) the likelihood function is a function of the parameters* θ, $\mathcal{L}(\theta) = p(\mathbf{x}|\theta)$ *(or* $\mathcal{L}(\theta) = P_\theta(\mathbf{x})$*).*

Definition 4 (Maximum likelihood parameters estimates)

The maximum likelihood parameters estimates are the parameters that maximise the likelihood function.

$$\theta_{ML} = \arg\max_{\theta} \mathcal{L}(\theta)$$

Maximum-likelihood estimation for the basis function linear regression model

We seek the parameter $\mathbf{w}_{ML}$ and β_{ML} that maximised the following likelihood function

$$\mathcal{L}(\mathbf{w}, \beta) = p(\mathbf{t}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1} I_N) = \prod_{n=1}^{N} \mathcal{N}(t_n|\phi^\intercal(x_n)\mathbf{w}, \beta^{-1}) \quad (11)$$

They are

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^\intercal \mathbf{\Phi})^{-1} \mathbf{\Phi}^\intercal \mathbf{t} \quad (12)$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \phi(\mathbf{x}_n)^\intercal \mathbf{w}_{ML})^2 \quad (13)$$

- first regression method that assumes random observations
- if the likelihood function is assumed to be Normal, maximum-likelihood and least-squares coefficients estimates are equal.

Exercise

Exercise 1

Derive the formulas for the maximum likelihood estimates of the coefficients, $\mathbf{w}$*, and noise precision,* β*, of the basis functions linear regression model given in Eqs. 12 and 13.*

Outline

Bayesian linear regression: motivation

- elegant,
- naturally allows online regression,
- does not require cross-validation for model selection,
- it is the first step to more complex Bayesian modelling.

Batch Bayesian linear regression: posterior distribution of parameters

In Bayesian linear regression we seek the posterior distribution of the weights of the linear regression model, $\mathbf{w}$, given the observations, which is proportional to the product of the likelihood function, $p(\mathbf{t}|\mathbf{w})$, and the prior, $p(\mathbf{w})$; i.e.,

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \tag{14}$$

To calculate this posterior below we use the likelihood function defined in Eq. 11 and the following prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

Using the expression of the conditional of the Linear Gaussian model, Eq. 5, we obtain

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^\intercal\mathbf{t} \tag{15}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^\intercal\mathbf{\Phi} \tag{16}$$

Batch Bayesian linear regression: exercise

Exercise 2

Derive the formulas for the Bayesian posterior mean (Eq. 15) and covariance (Eq. 16) of the basis function linear regression model.

Exercise 3

Show that

$$log \log p(\mathbf{w}|\boldsymbol{t}) = K - \frac{\beta}{2}||\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}||_2^2 - \frac{\alpha}{2}||\mathbf{w}||_2^2 \tag{17}$$

Therefore, the maximum-a-posteriori parameters of the basis function linear regression model are the solution of the regularised least-squares problem with $\lambda = \alpha/\beta$.
Note that, as we will show next, Bayesian linear regression uses the full posterior of the parameters to make predictions or to do model selection, and not just the maximum-a-posteriori parameters.

Batch Bayesian linear regression: demo code

Available here

Online Bayesian linear regression: recursive update of posterior distribution of parameters

Claim 4 (recursive update)

If the observations, $\{\mathbf{t}_1, \ldots, \mathbf{t}_n, \ldots\}$, are linearly independent when conditioned on the model parameters, $\boldsymbol{\theta}$, then for any $n \in \mathbb{N}$

$$p(\boldsymbol{\theta}|\mathbf{t}_1, \ldots, \mathbf{t}_n) = K \; p(\mathbf{t}_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{t}_1, \ldots, \mathbf{t}_{n-1}) \tag{18}$$

where K is a quantity that does not depend on $\boldsymbol{\theta}$.

Online Bayesian linear regression: recursive update of posterior distribution of parameters

Proof.

By induction on $H_n : p(\boldsymbol{\theta}|\mathbf{t}_1, \dots, \mathbf{t}_n) = K\ p(\mathbf{t}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{t}_1, \dots, \mathbf{t}_{n-1})$.

H_1

$$p(\boldsymbol{\theta}|\mathbf{t}_1) = \frac{p(\boldsymbol{\theta}, \mathbf{t}_1)}{p(\mathbf{t}_1)} = \frac{p(\mathbf{t}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{t}_1)} = K\ p(\mathbf{t}_1|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$H_n \rightarrow H_{n+1}$

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{t}_1, \dots, \mathbf{t}_{n+1}) &= \frac{p(\boldsymbol{\theta}, \mathbf{t}_1, \dots, \mathbf{t}_{n+1})}{p(\mathbf{t}_1, \dots, \mathbf{t}_{n+1})} \\
&= \frac{p(\mathbf{t}_{n+1}|\boldsymbol{\theta}, \mathbf{t}_1, \dots, \mathbf{t}_n)p(\boldsymbol{\theta}, \mathbf{t}_1, \dots, \mathbf{t}_n)}{p(\mathbf{t}_1 \dots, \mathbf{t}_{n+1})} \\
&= \frac{p(\mathbf{t}_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}, \mathbf{t}_1, \dots, \mathbf{t}_n)}{p(\mathbf{t}_1 \dots, \mathbf{t}_{n+1})} \\
&= \frac{p(\mathbf{t}_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{t}_1, \dots, \mathbf{t}_n)p(\mathbf{t}_1, \dots, \mathbf{t}_n)}{p(\mathbf{t}_1 \dots, \mathbf{t}_{n+1})} \\
&= K\ p(\mathbf{t}_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{t}_1, \dots, \mathbf{t}_n)
\end{aligned}$$

Note: the third equality above holds because the observations are assumed to be conditional independent given the parameters.

□

Conjugate priors

Above we showed that, if observations are independent, for the basis functions linear regression model

$$p(\mathbf{w}|\mathbf{t}_1) \propto p(\mathbf{t}_1|\mathbf{w})p(\mathbf{w}) \quad (19)$$

$$p(\mathbf{w}|\mathbf{t}_1, \ldots, \mathbf{t}_{n+1}) \propto p(\mathbf{t}_{n+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_1, \ldots, \mathbf{t}_n) \quad (20)$$

It would be helpful to choose a prior $p(\mathbf{w})$ in Eq. 19 such that, for the likelihood $p(\mathbf{t}_1|\mathbf{w})$, the posterior $p(\mathbf{w}|\mathbf{t}_1)$ has the same functional form at the prior.
Then, the posterior in Eq. 20 will have the same functional form as the "prior" $p(\mathbf{w}|\mathbf{t}_1, \ldots, \mathbf{t}_n)$ in the same equation.
Thus, all posteriors will have the same functional form as the prior $p(\mathbf{w})$.

Definition 5 (Conjugate prior)

If the posterior distribution, $p(\theta|x)$, is in the same probability distribution family as the prior probability distribution, $p(\theta)$, the prior is called a conjugate prior for the likelihood function $p(x|\theta)$.

Below we prove that the prior we chose for the coefficients of the basis function linear regression model, Eq. 14, is a conjugate prior for the likelihood function of this model, Eq. 11.

Conjugate prior for the coefficients of the basis functions linear regression model

Claim 5

If

$$P(\mathbf{w}|\mathbf{t}_1, \ldots, \mathbf{t}_n) = \mathcal{N}(\mathbf{w}|\mathbf{m}_n, \mathbf{S}_n) \tag{21}$$

$$P(\mathbf{t}_{n+1}|\mathbf{w}) = \mathcal{N}(\mathbf{t}_{n+1}|\mathbf{\Phi w}, \beta^{-1}\mathbf{I}) \tag{22}$$

then

$$P(\mathbf{w}|\mathbf{t}_1, \ldots, \mathbf{t}_{n+1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{n+1}, \mathbf{S}_{n+1})$$

with

$$S_{n+1} = S_n - (\beta^{-1} + \phi(\mathbf{x}_{n+1})^\intercal S_n \phi(\mathbf{x}_{n+1}))^{-1} S_n \phi(\mathbf{x}_{n+1}) \phi(\mathbf{x}_{n+1})^\intercal S_n \tag{23}$$

$$\begin{aligned} m_{n+1} = {} & \beta t_{n+1} S_{n+1} \phi(\mathbf{x}_{n+1}) + \mathbf{m}_n - \\ & (\beta^{-1} + \phi(\mathbf{x}_{n+1})^\intercal S_n \phi(\mathbf{x}_{n+1}))^{-1} \phi(x_{n+1})^\intercal \mathbf{m}_n S_n \phi(x_{n+1}) \end{aligned} \tag{24}$$

Conjugate prior for the coefficients of the basis functions linear regression model

In the proof below we will use the following lemma.

Lemma 6 (Matrix inversion lemma)

If $A \in \mathbb{R}^{N\times N}$, $U, V \in \mathbf{R}^{N\times M}$ and $C \in \mathbb{R}^{M\times M}$ then

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

Proof for Claim 5.

Using the formula for the conditional of the linear Gaussian model, Eq. 5, with the expression of the prior, Eq. 21, and likelihood, Eq. 22, we obtain

$$S_{n+1} = (S_n^{-1} + \beta\phi(\mathbf{x}_{n+1})\phi(\mathbf{x}_{n+1}^\intercal))^{-1} \tag{25}$$

$$\mathbf{m}_{n+1} = S_{n+1}(\beta t_{n+1}\phi(\mathbf{x}_{n+1}) + S_n^{-1}\mathbf{m}_n) \tag{26}$$

Note that Eq. 25 requires the inversion and $N \times N$ matrix, which has a complexity of $\mathcal{O}(N^3)$. We can avoid this inversion by using the matrix inversion lemma (with $A = S_n^{-1}, U = V = \phi(\mathbf{x}_{n+1}), C = \beta$), yielding Eq. 23.

Conjugate prior for the coefficients of the basis functions linear regression model

Proof.

Eq. 26 also requires the inversion of an $N \times N$ matrix. We can avoid this inversion as follows. From Eq. 26

$$\mathbf{m}_{n+1} = \beta t_{n+1} S_{n+1} \phi(\mathbf{x}_{n+1}) + S_{n+1} S_n^{-1} \mathbf{m}_n \tag{27}$$

Now we can replace the expression of S_{n+1} given in Eq. 23 into Eq. 27

$$\begin{aligned}
\mathbf{m}_{n+1} &= \beta t_{n+1} S_{n+1} \phi(\mathbf{x}_{n+1}) + \\
&\quad (S_n - (\beta^{-1} + \phi(\mathbf{x}_{n+1})^\intercal S_n \phi(\mathbf{x}_{n+1}))^{-1} S_n \phi(\mathbf{x}_{n+1}) \phi(\mathbf{x}_{n+1})^\intercal) S_n^{-1} \mathbf{m}_n \\
&= \beta t_{n+1} S_{n+1} \phi(\mathbf{x}_{n+1}) + \\
&\quad (I_n - (\beta^{-1} + \phi(\mathbf{x}_{n+1})^\intercal \phi(\mathbf{x}_{n+1}))^{-1} S_n \phi(\mathbf{x}_{n+1}) \phi(\mathbf{x}_{n+1})^\intercal) \mathbf{m}_n \\
&= \beta t_{n+1} S_{n+1} \phi(\mathbf{x}_{n+1}) + \\
&\quad \mathbf{m}_n - (\beta^{-1} + \phi(\mathbf{x}_{n+1})^\intercal \phi(\mathbf{x}_{n+1}))^{-1} S_n \phi(\mathbf{x}_{n+1}) \phi(\mathbf{x}_{n+1})^\intercal \mathbf{m}_n
\end{aligned}$$

$\square$

Conjugate prior for the coefficients of the basis functions linear regression model

Note that Eqs. 25 and 26 both required the inversion of an $N \times N$ matrix, but Eqs. 23 and 24 only require the inversion of scalars.

Python code implementing online Bayesian regression can be found here.

Bayesian predictions

least squares

$$t_{new} = \phi(\mathbf{x}_{new})^{\mathsf{T}} \mathbf{w}_{LS}$$

Bayesian

$$\begin{aligned} p(t_{new}|\mathbf{t}, \alpha, \beta) &= \int p(t_{new}, \mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w} \\ &= \int p(t_{new}|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w} \end{aligned}$$

Exercise 4

Derive the close form solution of the Bayesian predictive distribution.

Code for Bayesian predictions

Available here.

Model comparison

We want to compare which of a set of basis function linear regression models $\{\mathcal{M}_1, \ldots, \mathcal{M}_Q\}$ best fits a given dataset, $\mathbf{t}$ without using cross validation. For this, we will compare the models evidences or marginal likelihoods:

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}, \mathbf{w}|\alpha, \beta) d\mathbf{w} = \int p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) d\mathbf{w} \tag{28}$$

with $p(\mathbf{t}|\mathbf{w}, \beta)$ and $p(\mathbf{w}|\alpha)$ given in Eqs. 11 and 14, respectively.

Exercise 5

Show that

$$\log p(\mathbf{t}|\alpha, \beta) = \frac{M}{2} \log \alpha + \frac{N}{2} \log \beta - E(\mathbf{m}_N) - \frac{1}{2} \log |\mathbf{A}| - \frac{N}{2} \log(2\pi)$$

where $E(\mathbf{m}_N) = \frac{\beta}{2} ||\mathbf{t} - \mathbf{\Phi} \mathbf{m}_n||^2 + \frac{\alpha}{2} \mathbf{m}_N^\intercal \mathbf{m}_N$, $A = \alpha \mathbf{I} + \beta \mathbf{\Phi}^\intercal \mathbf{\Phi}$ *and* $\mathbf{m}_N$ *is the mean of* $p(\mathbf{w}|\mathbf{t}, \alpha, \beta)$

Hint: Integrate Eq. 28 using Eq. 4, or by completing the squares.

Code for Bayesian model comparison

Available here.

References

Bishop, C. M. (2016). *Pattern recognition and machine learning*. Springer-Verlag New York.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables and stochastic processes*. Mc Graw Hill, fourth edition.

Trefethen, L. n. and Bau III, D. (1997). Numerical linear algebra.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of neurophysiology*, 102(1):614–635.