Mathematics for data assimilation revision

June 3, 2025

Part I
Linear algebra in data assimilation

Linear algebra is one of the fundamental tools used to describe data assimilation formulations.

1 Vector quantities

Vectors are one of the fundamental objects used in data assimilation because they are a convenient and compact notation to represent lots of quantities. The convention is to write vectors as lower case bold Roman symbols in print (e.g. $\mathbf{v}$). Let vector $\mathbf{v}$ be a vector with n components:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix}, \tag{1}$$

where the vector components are v_i $(i = 1, 2, \ldots, n)$. If the components are each real numbers (as they usually are in data assimilation applications), $\mathbf{v}$ can be said to be a member of the n-dimensional vector space of real numbers, abbreviated to $\mathbf{v} \in \mathbb{R}^n$.

- Vectors are normally written as columns by default, as in (1).
- In handwriting, bold face is difficult so an underline or an over-arrow is used instead (e.g. $\underline{v}$ or $\overrightarrow{v}$).
- Vectors exist in an n-dimensional vector space, made of n orthogonal axes, see Fig. 1. Given component i of $\mathbf{v}$ is v_i, this is the projection of the ith component along axis i.
  - When $n = 1$, the vector reduces to a scalar quantity, panel (a).
  - When $n = 2$, the vector exists in a two-dimensional vector space, panel (b).
  - When $n = 3$, the vector exists in a three-dimensional space, panel (c).
  - When $n > 3$ one cannot visualise the space, but one can think of the projection of the axes to the page, panel (d).
- Vectors can be used to store quantitative information. For example a *state vector* can be made of model quantities at each grid point in a model, see Fig. 2, or an observation vector can be made of values of each observation to be considered by a data assimilation system, see Fig. 3.
- A column vector can be made into a row vector (and vice versa) by the transpose operation:

$$\mathbf{v}^\mathsf{T} = \begin{pmatrix} v_1 & \cdots & v_i & \cdots & v_n \end{pmatrix}. \tag{2}$$

2 Addition of vectors

Vector addition follows the rule:

$$\mathbf{v} = \mathbf{a} + \mathbf{b}, \qquad v_i = a_i + b_i, \tag{3}$$

i.e. the addition is done component-by-component.

- Vector addition is commutative, $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$, and associative, $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$.
- Vectors added must have the same number of components and exist in the same vector space, $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$.

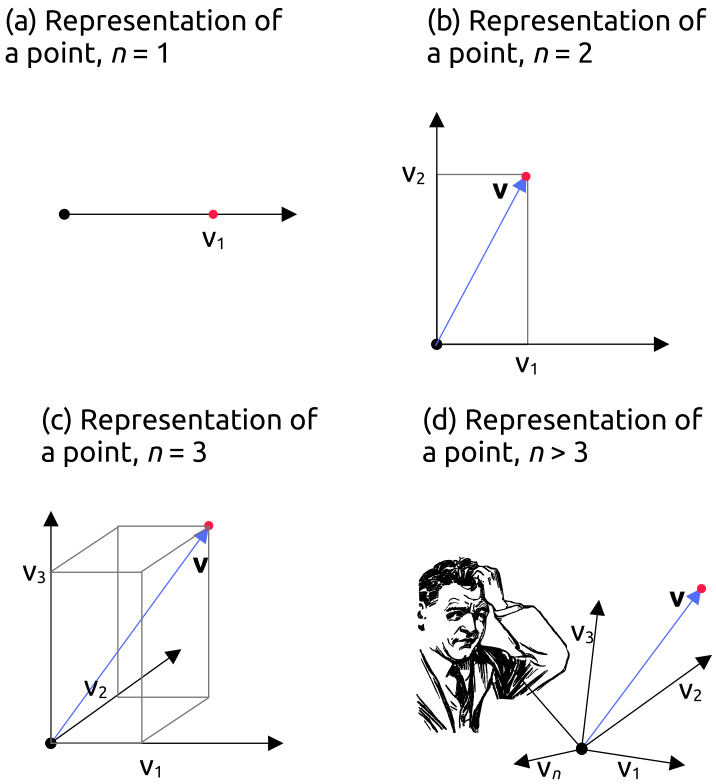

Figure 1: A vector represents a point in a vector space (red dot) of n dimensions. The origin is indicated by a black dot. Each dimension represented has an axis orthogonal (right angle) to all other axes. For $n > 3$ this is impossible to imagine (the axes have been projected onto the page), but is still a mathematically sound concept.

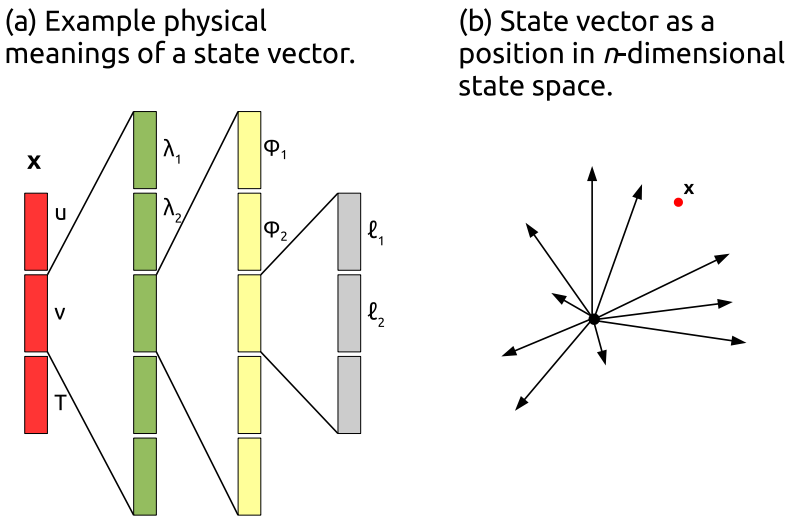

Figure 2: A state vector, $\mathbf{x}$, represents the state of the system under study, usually within the framework of a model. This example state vector represents model quantities in three spatial dimensions (3D). This example, panel (a) represents three model variables, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{T}$, ($n_\mathrm{v} = 3$). Each model variable is on a grid of n_λ longitudes, $\lambda_1, \lambda_2, \ldots$, n_ϕ latitudes, $\phi_1, \phi_2, \ldots$, and n_L vertical levels, $\ell_1, \ell_2, \ldots$. Even though the physical problem is in 3D space the vector space has $n = n_\mathrm{v} \times n_\lambda \times n_\phi$ dimensions (panel b). The value of n can be very large, e.g. $n = 10^9$. For the purposes of representing this collection of fields in mathematical formulae, imagine *flattening* the data so it exists as a linear 'string', as illustrated in (a). This is just a representation of the state so it can appear in matrix and vector equations. In the computer code however, the equivalent state might be best stored as 3D arrays for easy indexing.

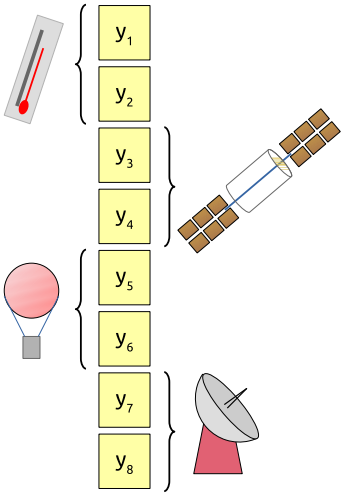

Figure 3: An example of a vector of observed values (normally given the symbol $\mathbf{y}$). Illustrated is the possible diversity of observation types (e.g. temperature measurements from an in-situ thermometer, remotely sensed radiance measurements from a satellite, in-situ humidity measurements from a radiosonde, and remotely sensed reflectivity measurements from a weather radar). Although $\mathbf{y}$ contains the measurements themselves (or their model's counterparts), meta data associated with the data (such as geographical position, time, etc.) needs to be stored alongside.

3 Product of vectors

Vectors can multiply in a number of ways, as illustrated in Fig. 4.

- The *inner (or scalar) product* of vectors $\mathbf{a} \in \mathbb{R}^n$ with $\mathbf{b} \in \mathbb{R}^n$ (panel a) is the projection of $\mathbf{a}$ onto $\mathbf{b}$. The result is a scalar:

$$\mathbf{a}^\intercal \mathbf{b} = \mathbf{a} \bullet \mathbf{b} = \sum_{i=1}^{n} a_i b_i. \tag{4}$$

  The "$\mathbf{a}^\intercal\mathbf{b}$" notation is consistent with the matrix multiplication convention by considering $\mathbf{a}^\intercal$ as a $1 \times n$ matrix and $\mathbf{b}$ as an $n \times 1$ matrix (see Sect. 9). For real vectors, the inner product satisfies $\mathbf{a}^\intercal\mathbf{b} = \mathbf{b}^\intercal\mathbf{a}$.

- The *weighted inner product* of vectors $\mathbf{a} \in \mathbb{R}^n$ with $\mathbf{b} \in \mathbb{R}^m$ (panel b) is the projection of $\mathbf{a}$ onto $\mathbf{Ab}$ (where $\mathbf{A}$ is a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, see Sect. 7). The result is a scalar:

$$\mathbf{a}^\intercal \mathbf{A} \mathbf{b} = \mathbf{a} \bullet (\mathbf{A}\mathbf{b}) = \sum_{i=1}^{n} a_i \sum_{j=1}^{m} A_{ij} b_j. \tag{5}$$

  The weighted inner product is associative, $(\mathbf{a}^\intercal\mathbf{A})\,\mathbf{b} = \mathbf{a}^\intercal\,(\mathbf{A}\mathbf{b})$, and for real vectors and matrices obeys the property $\mathbf{a}^\intercal\mathbf{A}\mathbf{b} = \mathbf{b}^\intercal\mathbf{A}^\intercal\mathbf{a}$. This product is commonly seen when specifying Gaussian probability density functions (Sect. 17), cost functions, and energy norms (Sect. 4).

- The *outer product* of vectors $\mathbf{a} \in \mathbb{R}^n$ with $\mathbf{b} \in \mathbb{R}^m$ (panel c) forms a matrix $\mathbf{a}\mathbf{b}^\intercal \in \mathbb{R}^{n\times m}$:

$$\mathbf{a}\mathbf{b}^\intercal = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_m \\ a_2b_1 & a_2b_2 & \cdots & a_2b_m \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_m \end{pmatrix}. \tag{6}$$

  The "$\mathbf{a}\mathbf{b}^\intercal$" notation is consistent with the matrix multiplication convention by considering $\mathbf{a}$ as an $n \times 1$ matrix and $\mathbf{b}^\intercal$ as a $1 \times m$ matrix (see Sect. 9). The outer product is not commutative, i.e. in general $\mathbf{a}\mathbf{b}^\intercal \neq \mathbf{b}\mathbf{a}^\intercal$. In fact $\mathbf{a}\mathbf{b}^\intercal = (\mathbf{b}\mathbf{a}^\intercal)^\intercal$.

- The *Schur product* (also known as a Hadamard product, panel d) follows the following rule:

$$\mathbf{v} = \mathbf{a} \circ \mathbf{b}, \qquad v_i = a_i b_i, \tag{7}$$

  i.e. the multiplication is done component-by-component. The Schur product is commutative, $\mathbf{a}\circ\mathbf{b} = \mathbf{b}\circ\mathbf{a}$, and associative, $(\mathbf{a}\circ\mathbf{b})\circ\mathbf{c} = \mathbf{a}\circ(\mathbf{b}\circ\mathbf{c})$. Vectors involved in the Schur product must have the same number of components and exist in the same vector space, $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$.

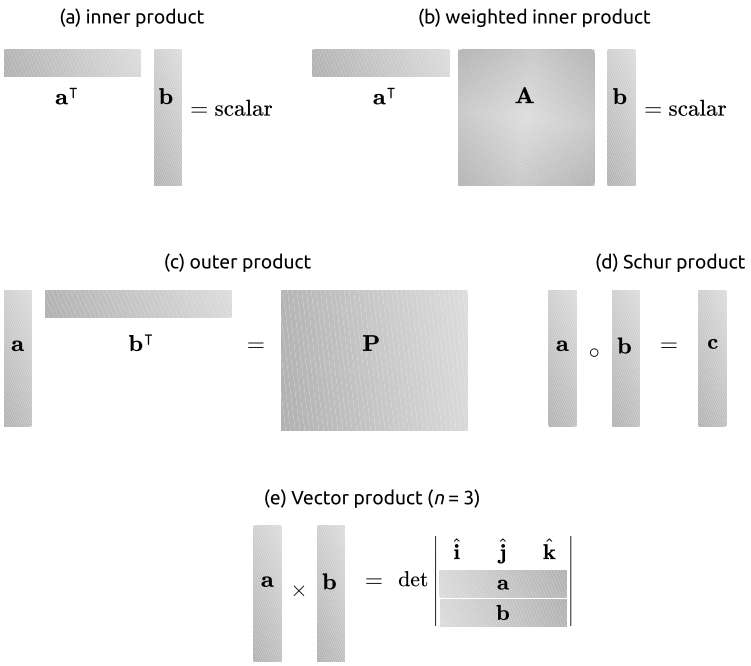

Figure 4: Visualisation of (a) the inner product (or scalar product) between two vectors $\mathbf{a}$ and $\mathbf{b}$, (b) the weighted inner product, (c) the outer product, (d) the Schur product, and (e) the vector product (or cross product).

- The *vector (or cross) product* of vectors $\mathbf{a} \in \mathbb{R}^3$ with $\mathbf{b} \in \mathbb{R}^3$ (panel e) produces a vector as defined by

$$\mathbf{a} \times \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \sin\theta \hat{\mathbf{n}}, \tag{8}$$

  where $\|\mathbf{a}\|$ is the norm or length of vector $\mathbf{a}$ (see Sect. 4), θ is the angle between $\mathbf{a}$ and $\mathbf{b}$, and $\hat{\mathbf{n}}$ is the unit vector orthogonal to $\mathbf{a}$ and $\mathbf{b}$ according to the right-hand-rule. This product is normally defined for vectors with $n = 3$. It is not often used in data assimilation theory, but is included here for completeness.

4 Norm of a vector

The *norm* of a vector $\mathbf{v} \in \mathbb{R}^n$, is its length (distance of the point represented by the vector from the origin, see Fig. 1). There are different ways of defining the norm of a vector, where the following are two examples often found in data assimilation.

- The *2-norm* is the L^2-norm (Euclidean norm):

$$L^2\text{-norm of } \mathbf{v} = \|\mathbf{v}\|_2 = (\mathbf{v}^\mathsf{T}\mathbf{v})^{1/2} = \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2}. \tag{9}$$

- The *energy norm* with respect to the symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is

$$\text{energy norm} = \|\mathbf{v}\|_\mathbf{A} = (\mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v})^{1/2} = \left(\sum_{i=1}^{n} v_i \sum_{j=1}^{n} A_{ij} v_j \right)^{1/2}. \tag{10}$$

  The 2-norm is a special case of the energy norm when $\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix (see Sect. 12). $\|\mathbf{v}\|_\mathbf{A}$ still has the interpretation of a length, but with respect to the metric $\mathbf{A}$. Energy norms are often used in data assimilation, especially when defining a cost function.

5 The vector derivative

Data assimilation often involves derivatives with respect to a vector. The derivative of a scalar function $f(\mathbf{x})$ with respect to its vector argument $\mathbf{x} \in \mathbb{R}^n$ is itself a vector $\in \mathbb{R}^n$ of the *partial derivative* with respect to each

component of $\mathbf{x}$:

$$\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)^\intercal = \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \partial f(\mathbf{x})/\partial x_1 \\ \partial f(\mathbf{x})/\partial x_2 \\ \vdots \\ \partial f(\mathbf{x})/\partial x_n \end{pmatrix} = \hat{\mathbf{x}}. \tag{11}$$

- Strictly, the derivative $\partial f(\mathbf{x})/\partial \mathbf{x}$ is a row matrix, so the transpose on the leftmost term makes this into a column vector (sometimes this convention is not observed).
- An example of the derivative of a nonlinear vector function with respect to its vector argument is done in Sect. 16.
- Sometimes the vector of derivatives with respect to a state $\mathbf{x}$ is called an *adjoint state*, $\hat{\mathbf{x}}$ (as indicated in Eq. (11)).

6 Matrix quantities

Matrices are another of the fundamental objects used in data assimilation. Matrices generally transform one vector to another vector. The convention is to write matrices as upper case bold Roman symbols (e.g. $\mathbf{A}$). Let vector $\mathbf{A}$ have n rows and m columns:

$$\mathbf{A} = \begin{pmatrix} A_{11} & \cdots & A_{1j} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots & & \vdots \\ A_{i1} & \cdots & A_{ij} & \cdots & A_{im} \\ \vdots & & \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nj} & \cdots & A_{nm} \end{pmatrix}, \tag{12}$$

where the matrix elements are A_{ij} $(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m)$. The index i specifies the row (how far 'down' the matrix from the top) and index j specifies the column (how far 'along' the matrix from the left). If the matrix elements are each real numbers (as they usually are in data assimilation applications), $\mathbf{A}$ can be said to be a member of the set $\mathbf{A} \in \mathbb{R}^{n \times m}$.

- If $n = m$ the matrix is *square*.
- The transpose of a matrix, $\mathbf{A}^\intercal$, swaps rows with columns. $\mathbf{A}^\intercal$ has matrix elements A_{ji} (swapped indices) and $\mathbf{A}^\intercal \in \mathbb{R}^{m \times n}$. Note that $(\mathbf{A}^\intercal)^\intercal = \mathbf{A}$. See Sect. 10.
- If $n = m$ and $A_{ij} = A_{ji}$ the matrix is said to be *symmetric*. Symmetric matrices satisfy $\mathbf{A} = \mathbf{A}^\intercal$.
- Matrices are used to hold a wide range of information, including a linear model, a linear observation operator, a transform from one set of variables to another, or a set of covariances or correlations between variables (see Sects. 7, 20, and 21).

7 Multiplication of a matrix with a vector

Matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ can multiply a vector $\mathbf{a} \in \mathbb{R}^m$ to yield another vector $\mathbf{b} \in \mathbb{R}^n$ (Fig. 5a):

$$\mathbf{b} = \mathbf{A}\mathbf{a}, \qquad \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & \cdots & A_{1m} \\ A_{11} & A_{12} & \cdots & \cdots & A_{1m} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & \cdots & A_{nm} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_m \end{pmatrix}, \qquad b_i = \sum_{j=1}^{m} A_{ij} a_j. \tag{13}$$

- The dimensionalities of $\mathbf{a}$ and $\mathbf{b}$ (m and n respectively) do not have to be the same. In the example (13) $m > n$, but this is just an illustration. The number of elements in $\mathbf{b}$ must equal the number of rows of $\mathbf{A}$ and the number of columns of $\mathbf{A}$ must equal the number of elements in $\mathbf{a}$. An easy way of seeing if such a multiplication is valid: write down the dimensionalities of each object in (13) as follows

$$\underset{\uparrow}{[n]} \overset{\checkmark}{=} \underset{\uparrow}{[n} \times \overset{\downarrow}{m]} \ \overset{\downarrow}{[m]}. \tag{14}$$

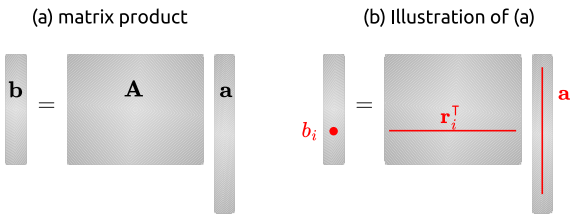

Figure 5: Visualisation of (a) product of matrix $\mathbf{A}$ on the vector $\mathbf{b}$. In (b) it is shown how component b_i is found by the inner product (Sect. 3) of the row vector $\mathbf{r}_i^\intercal$ with $\mathbf{a}$, i.e. $\mathbf{r}_i^\intercal\mathbf{a}$.

For the multiplication of a vector by a matrix to be valid, the two values indicated by the downward arrows (blue) must be the same, and the two values indicated by the upward arrows (red) must also be the same.

- Element b_i may be thought of as the inner product between the vector comprised of row i of $\mathbf{A}$, $\mathbf{r}_i^\intercal$ with $\mathbf{a}$ (Fig. 5b).
- There is another very useful interpretation of the operator of $\mathbf{A}$ in $\mathbf{b} = \mathbf{A}\mathbf{a}$. The columns of $\mathbf{A}$ can be regarded as basis vectors. Let column j of $\mathbf{A}$ be $\boldsymbol{\alpha}_j$. $\mathbf{b}$ can then be interpreted as the vector comprised of the linear combination of $\boldsymbol{\alpha}_j$, where the coefficients in the linear combination are components of $\mathbf{a}$:

$$\mathbf{b} = \sum_{j=1}^{m} \boldsymbol{\alpha}_j a_j. \tag{15}$$

  This is exactly equivalent to Eq. (13).
- The purposes of multiplying a matrix with a vector are many.
  - The matrix may represent a 'physical process'. One example in data assimilation is the linear model propagation. Given that a state vector $\mathbf{x}_t \in \mathbb{R}^n$ may represent the state of a system at a time t (Fig. 2), the action of the square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ may represent the linear propagation (Sect. 15) of $\mathbf{x}_t$ to the new time $t + \delta t$, $\mathbf{x}_{t+\delta t} = \mathbf{M}\mathbf{x}_t$. Another example is the observation operator. For the same $\mathbf{x}_t$ vector just mentioned, the action of the rectangular matrix $\mathbf{H} \in \mathbb{R}^{p\times n}$ may represent a linear observation operator to give the model's estimation of the observations, $\mathbf{y}_t \in \mathbb{R}^p$ (p observations), given state $\mathbf{x}_t$, $\mathbf{y}_t = \mathbf{H}\mathbf{x}_t$. While $\mathbf{x}_t$ exists in n-dimensional state space (Figs. 1 and 2), $\mathbf{y}_t$ exists in p-dimensional observation space (Fig. 3).
  - The matrix may represent a change of variables, which is sometimes convenient in data assimilation. A change of variables may, for instance, comprise of a rotation of the axes and a scaling of the axes.
  - The matrix may represent a metric to measure a 'distance' as in the weighted inner product in Sect. 3, Eq. (5).

8 Addition of matrices

Matrix addition follows the rule:

$$\mathbf{C} = \mathbf{A} + \mathbf{B}, \qquad C_{ij} = A_{ij} + B_{ij}, \tag{16}$$

i.e. the addition is done matrix-element-by-matrix-element.

- Matrix addition is commutative, $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, and associative, $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.
- Matrices added must have the same shape, $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n\times m}$.

9 Multiplication of matrices

Matrices can multiply in a couple of ways, as illustrated in Fig. 6.

- The *matrix product* of $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $\mathbf{B} \in \mathbb{R}^{m\times p}$ (panel a) gives $\mathbf{C} \in \mathbb{R}^{n\times p}$:

$$\mathbf{C} = \mathbf{A}\mathbf{B}, \qquad C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}. \tag{17}$$

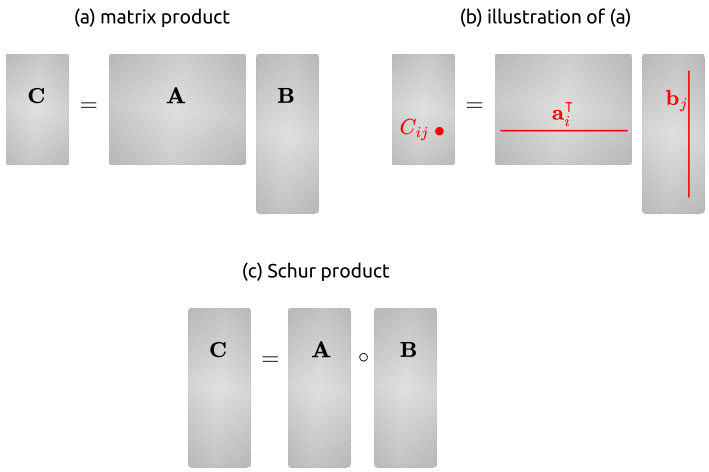

Figure 6: Visualisation of (a) the standard matrix product between two matrices $\mathbf{A}$ and $\mathbf{B}$ to give $\mathbf{C}$. In (b) it is shown how C_{ij} is found by the inner product (Sect. 3) of the row i of matrix $\mathbf{A}$,$\mathbf{a}_i^\intercal$, with column j of matrix $\mathbf{B}$, $\mathbf{b}_j$, i.e. $\mathbf{a}_i^\intercal \mathbf{b}_j$. Visualisation of (c) the Schur product.

To find component i, j of the product $\mathbf{C}$, consider row i of $\mathbf{A}$ as vector $\mathbf{a}_i^\intercal$ and column j of $\mathbf{B}$ as vector $\mathbf{b}_j$, and compute the inner product $\mathbf{a}_i^\intercal \mathbf{b}_j$ (see panel b and Sect. 3). This gives the result (17).
For the matrix product to be valid, the number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$. An easy way of seeing if such a multiplication is valid, write down the dimensionalities of each object in (17) as follows

$$\underset{\uparrow}{[n} \times \overset{\downarrow}{p}] \overset{\swarrow}{=} \underset{\uparrow}{[n} \times \overset{\searrow}{m}] \quad [\overset{\swarrow}{m} \times \overset{\downarrow}{p}] \,. \tag{18}$$

For the multiplication of two matrices to be valid, the two values indicated by the slanted arrows (red) must be the same. The number of rows of the matrix product will be the same as the number of rows of the first matrix, indicated by the upward arrows (blue); and the number of columns of the matrix product will be the same as the number of columns of the second matrix, indicated by the downward arrows (green).

- The matrix product is not commutative, i.e. in general $\mathbf{AB} \neq \mathbf{BA}$, but it is associative, i.e. $(\mathbf{AB})\,\mathbf{C} = \mathbf{A}\,(\mathbf{BC})$.
- The algebra involved for the multiplication of a matrix with a vector (Sect. 7) is the same as for matrix multiplication where the vectors are each considered matrices of width 1 ($n \times 1$ and $m \times 1$ matrices in the example in Eq. (13)).
- The *Schur product* (also known as a Hadamard product, panel c) follows the rule:

$$\mathbf{C} = \mathbf{A} \circ \mathbf{B}, \qquad C_{ij} = A_{ij} B_{ij}, \tag{19}$$

  i.e. the multiplication is done matrix-element-by-matrix-element. The Schur product is commutative, $\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$, and associative, $(\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C} = \mathbf{A} \circ (\mathbf{B} \circ \mathbf{C})$. Matrices involved in the Schur product must have the same number of components and exist in the same vector space, $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times m}$.
- The algebra involved for the Schur product of two vectors (Sect. 3) is the same as for Schur product of two matrices where the vectors are each considered matrices of width 1 ($n \times 1$ matrices in the example in Eq. (7)).

10 Transpose of a matrix

Just as a column vector can be made into a row vector (and vice versa) in Eq. (2) by the transpose operation, a matrix can be transposed, making its rows into columns and columns into rows. Consider matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$

in Eq. (12). The transpose gives the matrix $\mathbf{A}^\intercal \in \mathbb{R}^{m\times n}$:

$$\mathbf{A}^\intercal = \begin{pmatrix} A_{11} & \cdots & A_{j1} & \cdots & A_{n1} \\ \vdots & \ddots & \vdots & & \vdots \\ A_{1i} & \cdots & A_{ji} & \cdots & A_{ni} \\ \vdots & & \vdots & \ddots & \vdots \\ A_{1m} & \cdots & A_{jm} & \cdots & A_{nm} \end{pmatrix}, \qquad A^\intercal_{ij} = A_{ji}. \tag{20}$$

- The transpose of a product of matrices is the transpose of each individual matrix, but in reverse order. For instance if $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$, then

  $$\mathbf{AB} \in \mathbb{R}^{n\times p}, \quad (\mathbf{AB})^\intercal = \mathbf{B}^\intercal\mathbf{A}^\intercal \in \mathbb{R}^{p\times n}.$$

  As a check use Eq. (18) to see if the dimensions of this works:

  $$\underset{\uparrow}{[p} \times \overset{\downarrow}{n}] \overset{\checkmark}{=} \underset{\uparrow}{[p} \times \overset{\searrow}{m}] \; [\overset{\swarrow}{m} \times \overset{\downarrow}{n}] \; . \tag{21}$$

- The transpose of a matrix is sometimes called the *adjoint* of the matrix.

11 Diagonal matrices

Square diagonal matrices can have nonzero values only in the diagonal elements. An example diagonal matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is

$$\mathbf{A} = \begin{pmatrix} A_{11} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & A_{ii} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & A_{nn} \end{pmatrix}. \tag{22}$$

12 The identity matrix

There is a special square matrix called the identity matrix, $\mathbf{I}$. This is a diagonal matrix with all diagonal matrix elements unity. For $\mathbf{I} \in \mathbb{R}^{n\times n}$:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{23}$$

Multiplying a vector or matrix with the identity matrix does not change the vector or matrix.

13 Rank of a matrix

The rank of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is an integer quantity describing the number of independent 'directions' that are relevant to the matrix. Consider column i of $\mathbf{A}$ as the vector $\boldsymbol{\alpha}_i$. The rank of $\mathbf{A}$ is the dimensionality of the *space spanned by all the* $\boldsymbol{\alpha}_i$. The rank is not necessarily the number of columns. If two columns point in the same direction for instance (or in exactly opposite directions), they count as only one independent direction. More generally if $\boldsymbol{\alpha}_i$ can be described entirely as a linear combination of other columns, then it does not count towards the rank. Furthermore, the rank cannot be greater than the minimum of n and m.

As an example, consider the following 3×4 matrix:

$$\mathbf{A} = \begin{pmatrix} 7 & -4 & 2 & -1 \\ -3 & 1 & -3 & -2 \\ 1 & 2 & 8 & 5 \end{pmatrix}.$$

This has rank 3. Even though there are four column vectors, they span a space of only three dimensions. The third column, $\boldsymbol{\alpha}_3$ is not linearly independent; it is related to the first two columns via $\boldsymbol{\alpha}_3 = 2\boldsymbol{\alpha}_1 + 3\boldsymbol{\alpha}_2$, and so therefore does not represent a new direction.

Consider a square matrix, $\mathbf{A} \in \mathbb{R}^{n\times n}$. If $\mathbf{A}$ has rank n it is said to be of *full rank*.

14 Inverse of a matrix

The inverse of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is denoted $\mathbf{A}^{-1} \in \mathbb{R}^{n\times n}$. When multiplied with the original matrix, the identity matrix is produced:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \qquad \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{24}$$

Only square full-rank matrices have true inverses.

- Often an equation of the form $\mathbf{b} = \mathbf{A}\mathbf{a}$ (as in Eq. (13), but where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{n\times n}$) emerges from some analysis, where $\mathbf{A}$ is known, $\mathbf{b}$ is given, and $\mathbf{a}$ is unknown. Vector $\mathbf{a}$ can be found by operating with $\mathbf{A}^{-1}$ from the left on each side:

$$\begin{aligned} \mathbf{b} &= \mathbf{A}\mathbf{a} && (25)\\ \mathbf{A}^{-1}\mathbf{b} &= \mathbf{A}^{-1}\mathbf{A}\mathbf{a} \\ \mathbf{A}^{-1}\mathbf{b} &= \mathbf{I}\mathbf{a} \\ \mathbf{A}^{-1}\mathbf{b} &= \mathbf{a}. && (26) \end{aligned}$$

  (In practice, one wouldn't solve for $\mathbf{a}$ by finding the inverse, but by using a solver algorithm by specifying Eq. (25). The steps leading to (26) are still useful though when manipulating matrix equations.)

- In general the inverse is not comprised of the separate inverses of each matrix element of $\mathbf{A}$, i.e. in general $(A^{-1})_{ij} \neq (1/A_{ij})$. The exception to this is when $\mathbf{A}$ is diagonal. In that case, the inverse of (22) is

$$\mathbf{A}^{-1} = \begin{pmatrix} 1/A_{11} & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1/A_{ii} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1/A_{nn} \end{pmatrix}. \tag{27}$$

- The inverse of a product of square matrices is the inverse of each individual matrix, but in reverse order. For instance if $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$, then

$$\mathbf{A}\mathbf{B} \in \mathbb{R}^{n\times n}, \quad (\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \in \mathbb{R}^{n\times n}.$$

15 Linear and nonlinear functions

Consider a general function, $\boldsymbol{\mathcal{A}}$, that has a vector $\mathbf{a} \in \mathbb{R}^n$ as its input and vector $\mathbf{b} \in \mathbb{R}^m$ as its output:

$$\mathbf{b} = \boldsymbol{\mathcal{A}}(\mathbf{a}). \tag{28}$$

If $\boldsymbol{\mathcal{A}}$ is a *linear function*, $\boldsymbol{\mathcal{A}}^{\text{linear}}$, then the following rule is satisfied:

$$\boldsymbol{\mathcal{A}}^{\text{linear}}(\alpha\mathbf{a}_1 + \beta\mathbf{a}_2) = \alpha\boldsymbol{\mathcal{A}}^{\text{linear}}(\mathbf{a}_1) + \beta\boldsymbol{\mathcal{A}}^{\text{linear}}(\mathbf{a}_2), \tag{29}$$

otherwise $\boldsymbol{\mathcal{A}}$ is a *nonlinear function*. A linear function is equivalent to multiplication by a matrix, so the notation can be used where $\mathbf{A}$ represents the matrix action of $\boldsymbol{\mathcal{A}}^{\text{linear}}$, and the brackets to enclose arguments are no longer required: $\mathbf{A}(\alpha\mathbf{a}_1 + \beta\mathbf{a}_2) = \alpha\mathbf{A}\mathbf{a}_1 + \beta\mathbf{A}\mathbf{a}_2$ (brackets are still needed to factorise terms, as in the left hand side). Given the input vector has n elements and the output vector has m elements, then the matrix must be an $m \times n$ matrix.

16 Taylor expansions

Provided a nonlinear function, $\boldsymbol{\mathcal{A}}(\mathbf{x})$, is smooth, etc., the following expansion can be written:

$$\boldsymbol{\mathcal{A}}(\mathbf{x} + \delta\mathbf{x}) = \boldsymbol{\mathcal{A}}(\mathbf{x}) + \left.\frac{\partial\boldsymbol{\mathcal{A}}(\mathbf{x})}{\partial\mathbf{x}}\right|_{\mathbf{x}} \delta\mathbf{x} + \text{h.o.t.}, \tag{30}$$

where $\boldsymbol{\mathcal{A}}(\mathbf{x})$ is a (potentially nonlinear) function that has a vector $\in \mathbb{R}^n$ as its input and vector $\in \mathbb{R}^m$ as its output vector, $\delta\mathbf{x} \in \mathbb{R}^n$ is a perturbation to $\mathbf{x}$, $\partial\boldsymbol{\mathcal{A}}(\mathbf{x})/\partial\mathbf{x}|_{\mathbf{x}}$ is the $\in \mathbb{R}^{m\times n}$ matrix called the *linearised version* of $\boldsymbol{\mathcal{A}}$, the *Jacobian* of $\boldsymbol{\mathcal{A}}$, or the *tangent linear* of $\boldsymbol{\mathcal{A}}$ (evaluated at $\mathbf{x}$), and h.o.t. stands for "higher order terms".

- Matrix element i, j of the Jacobian is the partial derivative of the ith output element of $\boldsymbol{\mathcal{A}}(\mathbf{x})$ with respect to the jth input element of $\mathbf{x}$, namely $[\partial \boldsymbol{\mathcal{A}}(\mathbf{x})/\partial \mathbf{x}|_{\mathbf{x}}]_{ij} = \partial \mathcal{A}_i/\partial x_j$. Expanding out the matrix algebra for this term in (30) (a.k.a. the *first order term*) reveals the chain rule.
- Equation (30) is the multi-dimensional or multivariate Taylor expansion of $\boldsymbol{\mathcal{A}}(\mathbf{x})$ and shows (to first order) how perturbations to the input affect the output.

Part II
Basic probability and statistics in data assimilation

Basic statistical concepts are another fundamental tool used to describe data assimilation methods.

17 Probability density functions

It is useful to regard data from observations and models as random draws from a distribution. This allows these quantities to be regarded as fundamentally probabilistic, which is a convenient way of quantifying uncertainty. Consider a single (scalar) random continuous variable, x, e.g. a measured or modelled quantity. Associated with x is a probability density function, $P(x)$ (see Fig. 7). Suppose a value is drawn from the PDF (i.e. a measurement is made or a model is run). The chance of drawing values is proportional to the size of the PDF. Since x is continuous, the probability of drawing a particular value of x is infinitesimally small, but the probability of drawing a value of x between two specified values x_1 and x_2 can be computed as the area under the PDF (see the equations in the Fig.).

- $P(x)$ itself is not a probability but a *probability density* (probability per unit x).
- The area under the whole PDF must be equal to unity.
- A commonly used PDF is the *Gaussian* or *normal* density:

$$P(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}, \tag{31}$$

  where μ is the mean (Sect. 18), σ^2 is the variance (Sect. 19), and σ is the standard deviation of the Gaussian PDF (see Fig. 8). The exponential part describes the shape of the PDF, and the factor that appears at the front is for normalisation (see previous bullet point). The multivariate Gaussian density is:

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{COV})}} \exp -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^{\intercal} \mathbf{COV}^{-1} (\mathbf{x}-\boldsymbol{\mu}), \tag{32}$$

  where $\mathbf{x} \in \mathbb{R}^n$ is a state vector, $\boldsymbol{\mu} \in \mathbb{R}^n$ is the PDF's mean, and $\mathbf{COV} \in \mathbb{R}^{n \times n}$ is the covariance (see Sect. 20).
- Every quantity in the real world is only known to some precision. Quantifying this precision is done by specifying a PDF.
- Often a quantity (model output or observation) is specified as $y \pm \delta y$ (i.e. with error bars). This is often shorthand for saying that the quantity is represented by a Gaussian distribution, Eq. (31), with mean y and standard deviation δy (assuming the 'error' is represented by one standard deviation).

18 Averaging

There are various ways of specifying the average (or first moment) of a PDF.

- The *mean* is a common way of specifying an average. A PDF, $P(x)$, has a mean value of x, μ_{mean}, found from

$$\mu_{\text{mean}} = \int_{x=-\infty}^{\infty} x P(x) dx. \tag{33}$$

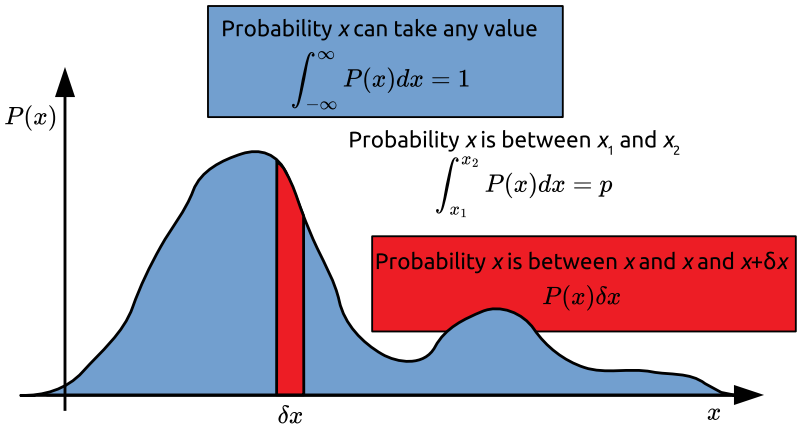

Figure 7: Essentials of a probability density function. x is the random variable and $P(x)$ is its probability density.

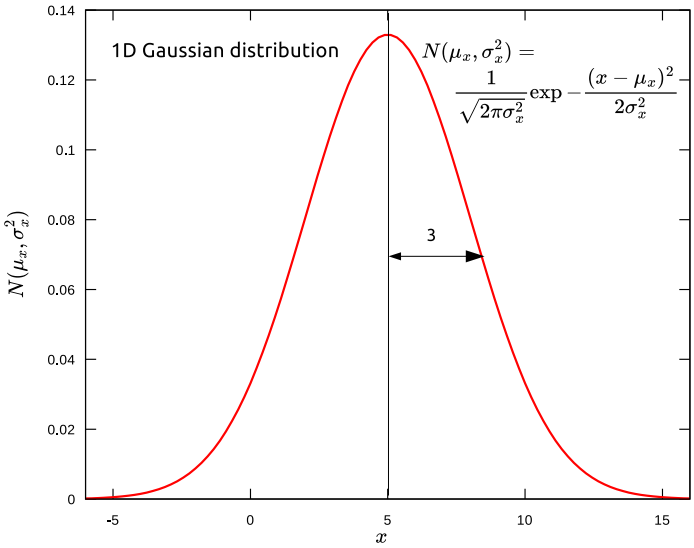

Figure 8: Gaussian distribution for the scalar x. The general form of a 1D Gaussian is given and the plot is for the example case with mean $\mu_x = 5$ and standard deviation $\sigma_x = 3$.

Given a sample of quantities drawn from the PDF, $x^{(1)}, x^{(2)}_{,} \ldots, x^{(N)}$ (e.g. repeated measurements or repeated model outputs for instance from an ensemble), the mean can be estimated to be $\mu_{\text{mean}} \approx (1/N) \sum_{k=1}^{N} x^{(k)}$. Be aware that it is possible for μ_{mean} to lie where the probability density is zero.

- The *median* of $P(x)$, μ_{median} is the value of x that has a equal probability $(1/2)$ of drawing a number less than μ_{median} and drawing a number greater than μ_{median}:

$$\int_{-\infty}^{\mu_{\text{median}}} P(x)dx = \int_{\mu_{\text{median}}}^{\infty} P(x)dx = \frac{1}{2}. \tag{34}$$

- The *mode* of $P(x)$, μ_{mode}, is the most likely value of x (where $P(x)$ is maximum).
- For the Gaussian PDF, Eq. (31), all these averages are equal: $\mu_{\text{mean}} = \mu_{\text{median}} = \mu_{\text{mode}}$.

19 Variance and standard deviation

The *variance*, σ^2 is the *second moment* of a PDF. It is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu_{\text{mean}})^2 P(x)dx, \tag{35}$$

where the mean, μ_{mean} is found from Eq. (33). The square-root of the variance, σ, is called the *standard deviation*. Given a sample of quantities drawn from the PDF, $x^{(1)}, x^{(2)}_{,} \ldots, x^{(N)}$, the variance can be estimated as $\sigma^2 \approx (1/N) \sum_{k=1}^{N} \left(x^{(k)} - \mu_{\text{mean}}\right)^2$ if μ_{mean} is known already, or $\sigma^2 \approx 1/(N-1) \sum_{k=1}^{N} \left(x^{(k)} - \mu_{\text{mean}}\right)^2$ if μ_{mean} is itself estimated from the sample as given at the end of the first bullet point in Sect. 18.

20 Mean and co-variance

The mean and variances in Sects. 18 and 19 are for a scalar random variable x. The multivariate (or vector) versions of Eqs. (33) and (35) are

$$\boldsymbol{\mu} = \int P(\mathbf{x})d\mathbf{x} \tag{36}$$

$$\mathbf{COV} = \int (\mathbf{x} - \boldsymbol{\mu}) (\mathbf{x} - \boldsymbol{\mu})^{\intercal} P(\mathbf{x})d\mathbf{x}, \tag{37}$$

where the "mean" annotation is dropped from $\boldsymbol{\mu}$ as it is implicit, and the integral is over the state space. Notice that the variance, σ^2, in (35) has become a *co-variance*, **COV**, in (37). Equations (36) and (37) are the first and second moments of $P(\mathbf{x})$ respectively.

- Just as $\mathbf{x}$ is a vector quantity, so is $\boldsymbol{\mu}$.
- The co-variance formula involves the outer product of vectors (Sect. 3), which gives a matrix (the *co-variance matrix*). **COV** is a matrix because it describes relationships between pairs of variables.
- Matrix element i, j of **COV**, COV_{ij} is the co-variance between elements i and j of $\mathbf{x}$. A positive co-variance means that when x_i increases (decreases), x_j also increases (decreases) on average. A negative co-variance means that when x_i increases (decreases), x_j decreases (increases) on average. The magnitude of COV_{ij} is a measure of the 'strength' of the relationship **and** the variability of the two variables. A related quantity is the *correlation* between elements i and j, which is a measure only of the 'strength' of the relationship (see Sect. 21).
- Diagonal element i, i of **COV**, COV_{ii}, is the variance of element i.
- Explicit formulae for mean and co-variance in terms of samples drawn from the multivariate PDF,

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \ldots, \mathbf{x}^{(N)}$:

$$\boldsymbol{\mu} \approx \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}^{(k)} \tag{38}$$

$$\text{where element } i \text{ is } \mu_i \approx \frac{1}{N}\sum_{k=1}^{N}x_i^{(k)} \tag{39}$$

$$\mathbf{COV} \approx \frac{1}{N-1}\sum_{k=1}^{N}\left(\mathbf{x}^{(k)}-\boldsymbol{\mu}\right)\left(\mathbf{x}^{(k)}-\boldsymbol{\mu}\right)^{\intercal} \tag{40}$$

$$\text{where matrix element } i,j \text{ is } COV_{ij} \approx \frac{1}{N-1}\sum_{k=1}^{N}\left(x_i^{(k)}-\mu_i\right)\left(x_j^{(k)}-\mu_j\right). \tag{41}$$

- There are also more compact expressions for Eqs. (38) and (40) (see below), which are very convenient for some derivations in ensemble data assimilation.
  - Construct a matrix $\mathbf{X} \in \mathbb{R}^{n\times N}$, where n is the dimensionality of the state space and N is the number of samples. The samples make up the columns of $\mathbf{X}$ (this is often done in ensemble data assimilation):

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(N)} \end{pmatrix}. \tag{42}$$

  - The mean can be found from the following multiplication:

$$\boldsymbol{\mu} \approx \mathbf{X}\begin{pmatrix} 1/N \\ 1/N \\ \vdots \\ 1/N \end{pmatrix}, \tag{43}$$

    where the column matrix $\in \mathbb{R}^N$. Also a matrix $\mathbf{M} \in \mathbb{R}^{n\times N}$ can be formed as N repeated columns of $\boldsymbol{\mu}$ (one column for each sample):

$$\mathbf{M} = \underbrace{\begin{pmatrix} \boldsymbol{\mu} & \boldsymbol{\mu} & \cdots & \boldsymbol{\mu} \end{pmatrix}}_{\mathbb{R}^{n\times N}} = \overbrace{\mathbf{X}}^{\mathbb{R}^{n\times N}}\overbrace{\begin{pmatrix} 1/N & 1/N & \cdots & 1/N \\ 1/N & 1/N & \cdots & 1/N \\ \vdots & \vdots & \ddots & \vdots \\ 1/N & 1/N & \cdots & 1/N \end{pmatrix}}^{\mathbb{R}^{N\times N}}. \tag{44}$$

  - The matrix of perturbed samples (perturbed from the mean) is $\delta\mathbf{X} \in \mathbb{R}^{n\times N}$:

$$\delta\mathbf{X} = \mathbf{X} - \mathbf{M}. \tag{45}$$

  - The co-variance matrix, $\mathbf{COV} \in \mathbb{R}^{n\times n}$ is then

$$\mathbf{COV} \approx \frac{1}{N-1}\delta\mathbf{X}\delta\mathbf{X}^{\intercal}. \tag{46}$$

    This is a very useful compact notation of a sample covariance. It can be checked that this is correct. Consider matrix element i,j of $\mathbf{COV}$ in Eq. (46), which can be expanded by the usual rules of matrix multiplication, Eq. (17):

$$\begin{aligned} COV_{ij} &= \frac{1}{N-1}\sum_{k=1}^{N}\delta X_{ik}(\delta X^{\intercal})_{kj} \\ &= \frac{1}{N-1}\sum_{k=1}^{N}\delta X_{ik}\delta X_{jk} \\ &= \frac{1}{N-1}\sum_{k=1}^{N}\left(X_{ik}-M_{ik}\right)\left(X_{jk}-M_{jk}\right). \end{aligned}$$

Now, X_{ik} is the ith element of sample $\mathbf{x}^{(k)}$, i.e. $x_i^{(k)}$, and M_{ik} is the ith element of the mean μ_i, so

$$COV_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}\left(x_i^{(k)} - \mu_i\right)\left(x_j^{(k)} - \mu_j\right),$$

which is exactly the same as that found from the explicit calculation in Eq. (41).

21 Correlation

The correlation between two variables x_i and x_j, COR_{ij}, is a normalised measure of the 'strength' of the statistical relationship. COR_{ij} is related to the covariance, COV_{ij}, via $COV_{ij} = \sigma_i COR_{ij} \sigma_j$. The correlation matrix **COR** is the matrix of elements COR_{ij}. The matrix relationship between **COR** and **COV** is

$$\mathbf{COV} = \mathbf{\Sigma COR \Sigma}, \tag{47}$$

where $\mathbf{\Sigma}$ is not the summation symbol here, but is the diagonal matrix of standard deviations:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n \end{pmatrix}. \tag{48}$$

22 Bayes' Theorem

All modern data assimilation schemes have their roots in *Bayes' Theorem*. A Bayesian data assimilation method ideally describes all quantities as PDFs (which describe the quantities' uncertainties, Sect. 17), and allows those PDFs to be updated/revised with new information according to Bayes' Theorem.

A plausibility argument for Bayes' Theorem is as follows. Suppose there are two events, A and B. The "probability of A and B happening" is $p(\mathrm{A} \cap \mathrm{B})$. This can be written as $p(\mathrm{A} \cap \mathrm{B}) = p(\mathrm{A}|\mathrm{B})p(\mathrm{B})$, where $p(\mathrm{A}|\mathrm{B})$ is the "probability of A happening given that B has happened" (a *conditional probability*) and $p(\mathrm{B})$ is the "probability of B happening". This can also be written by swapping A and B: $p(\mathrm{B} \cap \mathrm{A}) = p(\mathrm{B}|\mathrm{A})p(\mathrm{A})$. Since $p(\mathrm{A} \cap \mathrm{B}) = p(\mathrm{B} \cap \mathrm{A})$, this leads to

$$p(\mathrm{A}|\mathrm{B})p(\mathrm{B}) = p(\mathrm{B}|\mathrm{A})p(\mathrm{A}), \tag{49}$$

which is essentially Bayes' Theorem.

Replace "A" with the "event that state vector $\mathbf{x} \in \mathbb{R}^n$ represents truth", and replace "B" with the "event that observations $\mathbf{y} \in \mathbb{R}^p$ are the true value being measured", slightly rearranging, and using probability densities, leads Eq. (49) to

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}. \tag{50}$$

The form of Eq. (50) is the more useful for data assimilation.

- Even though the same symbol $p(\bullet)$ is used in each term in Eq. (49) (and $P(\bullet)$ in Eq. (50)), it should not be regarded as the same function in each term. This is really an abuse of notation.
- On the left hand side, $P(\mathbf{x}|\mathbf{y})$ is called the *posterior density* (posterior = after observations are made). It is the "PDF of $\mathbf{x}$ given observations $\mathbf{y}$", or the PDF that "$\mathbf{x}$ represents the truth given information from observations $\mathbf{y}$ is made available". Obtaining $P(\mathbf{x}|\mathbf{y})$ (and/or its first or second moments, see Sect. 20) is a major objective of data assimilation.
- On the right hand side, $P(\mathbf{y}|\mathbf{x})$ is the "PDF of obtaining observed values $\mathbf{y}$ given $\mathbf{x}$ is the true state". Formally this is a PDF with $\mathbf{y}$ as the random variable and $\mathbf{x}$ fixed. As a starting point in data assimilation however, $\mathbf{x}$ is regarded as a variable and $\mathbf{y}$ as fixed (i.e. the given observations). When regarded in this way, $P(\mathbf{y}|\mathbf{x})$ is known as the *likelihood function*. Often this is assumed to have the form of a Gaussian, Eq. (32) (in observation space), with the mean set to the observations themselves and with a specified covariance.
- Also on the right hand side, $P(\mathbf{x})$ is the *prior density* (prior = before observations are considered since it does not depend on $\mathbf{y}$). Often this is assumed to have the form of a Gaussian, Eq. (32) (in model space), with mean set to a *first guess* or *background* state and with a specified covariance.
- $P(\mathbf{y})$ in the denominator is often not important as it does not depend on $\mathbf{x}$ (it does not affect the moments of the posterior density).

Errata from previous versions

- From May 19th 2025, indices in Eq. (20) corrected.