Homework 2

Contents

References

- Lectures 4-8 (inclusive).

Instructions

- Type your name and email in the "Student details" section below.
- Develop the code and generate the figures you need to solve the problems using this notebook.
- For the answers that require a mathematical proof or derivation you should type them using latex. If you have never written latex before and you find it exceedingly difficult, we will likely accept handwritten solutions.
- The total homework points are 100. Please note that the problems are not weighed equally.

▶ Show code cell source

Student details

- **First Name:**
- **Last Name:**
- **Email:**
- **Used generative AI to complete this assignment (Yes/No):**
- **Which generative AI tool did you use (if applicable)?:**

Problem 1 - Joint probability mass function of two discrete random variables

Consider two random variables X and Y. X takes values $\{0,1,\dots,4\}$ and Y takes values $\{0,1,\dots,8\}$. Their joint probability mass function, can be described using a matrix:

```
P = np.array(
    [
        [0.03607908, 0.03760034, 0.00503184, 0.0205082 , 0.01051408,
         0.03776221, 0.00131325, 0.03760817, 0.01770659],
        [0.03750162, 0.04317351, 0.03869997, 0.03069872, 0.02176718,
         0.04778769, 0.01021053, 0.00324185, 0.02475319],
        [0.03770951, 0.01053285, 0.01227089, 0.0339596 , 0.02296711,
         0.02187814, 0.01925662, 0.0196836 , 0.01996279],
        [0.02845139, 0.01209429, 0.02450163, 0.00874645, 0.03612603,
         0.02352593, 0.00300314, 0.00103487, 0.04071951],
        [0.00940187, 0.04633153, 0.01094094, 0.00172007, 0.00092633,
         0.02032679, 0.02536328, 0.03552956, 0.01107725]
    ]
)
```

The rows of the matrix correspond to the values of X and the columns to the values of Y. So, if you wanted to find the probability of $p(X=2,Y=3)$ you would do:

```
print(f"p(X=2, Y=3) = {P[2, 3]:.3f}")
```

```
p(X=2, Y=3) = 0.034
```

A. Verify that all the elements of P sum to one, i.e., that $\sum_{x,y}p(X=x,Y=y)=1$.

Skip to main content

```
# Your code here
```

B. Find the marginal probability density of X:

$$p(x) = \sum_y p(x, y).$$

You can represent this as a 5-dimensional vector.

```
# Hint, you can do this in one line if you read this:
help(np.sum)
```

```
Help on function sum in module numpy:

sum(a, axis=None, dtype=None, out=None, keepdims=<no value>, initial=<no value>, where
    Sum of array elements over a given axis.

    Parameters
    ----------
    a : array_like
        Elements to sum.
    axis : None or int or tuple of ints, optional
        Axis or axes along which a sum is performed.  The default,
        axis=None, will sum all of the elements of the input array.  If
        axis is negative it counts from the last to the first axis.

        .. versionadded:: 1.7.0

        If axis is a tuple of ints, a sum is performed on all of the axes
        specified in the tuple instead of a single axis or all the axes as
        before.
    dtype : dtype, optional
        The type of the returned array and of the accumulator in which the
        elements are summed.  The dtype of `a` is used by default unless `a`
        has an integer dtype of less precision than the default platform
        integer.  In that case, if `a` is signed then the platform integer
        is used while if `a` is unsigned then an unsigned integer of the
        same precision as the platform integer is used.
    out : ndarray, optional
        Alternative output array in which to place the result. It must have
        the same shape as the expected output, but the type of the output
        values will be cast if necessary.
    keepdims : bool, optional
        If this is set to True, the axes which are reduced are left
        in the result as dimensions with size one. With this option,
        the result will broadcast correctly against the input array.
```

Skip to main content

```
        passed through to the `sum` method of sub-classes of
        `ndarray`, however any non-default value will be.  If the
        sub-class' method does not implement `keepdims` any
        exceptions will be raised.
    initial : scalar, optional
        Starting value for the sum. See `~numpy.ufunc.reduce` for details.

        .. versionadded:: 1.15.0

    where : array_like of bool, optional
        Elements to include in the sum. See `~numpy.ufunc.reduce` for details.

        .. versionadded:: 1.17.0

    Returns
    -------
    sum_along_axis : ndarray
        An array with the same shape as `a`, with the specified
        axis removed.   If `a` is a 0-d array, or if `axis` is None, a scalar
        is returned.  If an output array is specified, a reference to
        `out` is returned.

    See Also
    --------
    ndarray.sum : Equivalent method.

    add.reduce : Equivalent functionality of `add`.

    cumsum : Cumulative sum of array elements.

    trapz : Integration of array values using the composite trapezoidal rule.

    mean, average

    Notes
    -----
    Arithmetic is modular when using integer types, and no error is
    raised on overflow.

    The sum of an empty array is the neutral element 0:

    >>> np.sum([])
    0.0

    For floating point numbers the numerical precision of sum (and
    ``np.add.reduce``) is in general limited by directly adding each number
    individually to the result causing rounding errors in every step.
    However, often numpy will use a  numerically better approach (partial
    pairwise summation) leading to improved precision in many use-cases.
    This improved precision is always provided when no ``axis`` is given.
    When ``axis`` is given, it will depend on which axis is summed.
    Technically, to provide the best speed possible, the improved precision
    is only used when the summation is along the fast axis in memory.
```

Skip to main content

```
    Especially when summing a large number of lower precision floating point
    numbers, such as ``float32``, numerical errors can become significant.
    In such cases it can be advisable to use `dtype="float64"` to use a higher
    precision for the output.

    Examples
    --------
    >>> np.sum([0.5, 1.5])
    2.0
    >>> np.sum([0.5, 0.7, 0.2, 1.5], dtype=np.int32)
    1
    >>> np.sum([[0, 1], [0, 5]])
    6
    >>> np.sum([[0, 1], [0, 5]], axis=0)
    array([0, 6])
    >>> np.sum([[0, 1], [0, 5]], axis=1)
    array([1, 5])
    >>> np.sum([[0, 1], [np.nan, 5]], where=[False, True], axis=1)
    array([1., 5.])

    If the accumulator is too small, overflow occurs:

    >>> np.ones(128, dtype=np.int8).sum(dtype=np.int8)
    -128

    You can also start the sum with a value other than zero:

    >>> np.sum([10], initial=5)
    15
```

C. Find the marginal probability density of Y. This is a 9-dimensional vector.

```
# Your code here
```

D. Find the expectation and variance of X and Y.

```
# Your code here
```

E. Find the expectation of $E[X+Y]$.

```
# Your code here
```

F. Find the covariance of X and Y. Are the two variable correlated? If yes, are they positively or

Skip to main content

```
# Your code here
```

G. Find the variance of $X+Y$.

```
# Your code here
```

J. Find the probability that $X+Y$ is less than or equal to 5. That is, find $p(X+Y\le 5)$. Hint: Use two for loops to go over all the combinations of X and Y values, check if $X+Y\le 5$, and sum up the probabilities.

```
# Your code here
```

Problem 2 - Zero correlation does not imply independence

The purpose of this problem is to show that zero correlation does not imply independence. Consider the random variable X and Y following a standard normal distribution. Define the random variable as $Z=X^2+0.01\cdot Y$. You will show that the correlation between X and Z is zero even though they are not independent.

A. Take 100 samples of X and Z using numpy or scipy. Hint: First sample X and Y and use the samples to get Z.

```
# Your code here
```

B. Do the scatter plot between X and Z.

```
# Your code here
```

C. Use the scatter plot to argue that X and Z are not independent.

Answer:

Skip to main content

```
# Your code here
```

E. Use the samples you took to estimate the covariance between X and Z.

```
# Your code here
```

F. Use the results above to find the correlation between X and Z.

```
# Your code here
```

G. The correlation coefficient you get may not be very close to zero. This is due to the fact that we estimate it with Monte Carlo averaging. To get a better estimate, we can increase the number of samples. Try increasing the number of samples to 1000 and see if the correlation coefficient gets closer to zero.

```
# Your code here
```

H. Let's do a more serious estimation of Monte Carlo convergence. Take 100,000 samples of X and Z. Write code that estimates the correlation between X and Z using the first n samples for $n = 1, 2, \dots, 100,000$. Plot the estimates as a function of n. What do you observe? How many samples do you need to get a good estimate of the correlation?

```
# Your code here
```

Problem 3 - Creating a stochastic model for the magnetic properties of steel

The magnetic properties of steel are captured in the so-called $B-H$ curve, which connects the magnetic field H to the magnetic flux density B. The $B-H$ curve is a nonlinear function typically measured in the lab. It appears in Maxwell's equations and is, therefore, crucial in the design of electrical machines.

The shape of the $B-H$ curve depends on the manufacturing process of the steel. As a result,

Skip to main content

of this problem is to guide you through the process of creating a stochastic model for the $B-H$ curve using real data. Such a model is the first step when we do uncertainty quantification for the design of electrical machines. Once constructed, the stochastic model can generate random samples of the $B-H$ curve. We can then propagate the uncertainty in the $B-H$ curve through Maxwell's equations to quantify the uncertainty in the performance of the electrical machine.

Let's use some actual manufacturer data to visualize the differences in the $B-H$ curve across different suppliers. The data are here. Explaining how to upload data on Google Colab will take a while. We will do it in the next homework set. You should know that the data file `B_data.csv` needs to be in the same working directory as this Jupyter Notebook. I have written some code that allows you to put the data file in the right place without too much trouble. Run the following:

```
url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/lectureboo
download(url)
```

If everything worked well, then the following will work:

```
B_data = np.loadtxt('B_data.csv')
B_data
```

```
array([[0.        , 0.00490631, 0.01913362, ..., 1.79321352, 1.79337681,
        1.79354006],
       [0.        , 0.00360282, 0.01426636, ..., 1.8367998 , 1.83697627,
        1.83715271],
       [0.        , 0.00365133, 0.01433438, ..., 1.77555287, 1.77570402,
        1.77585514],
       ...,
       [0.        , 0.00289346, 0.01154411, ..., 1.7668308 , 1.76697657,
        1.76712232],
       [0.        , 0.00809884, 0.03108513, ..., 1.7774044 , 1.77756225,
        1.77772007],
       [0.        , 0.00349638, 0.0139246 , ..., 1.76460358, 1.76474439,
        1.76488516]])
```

The shape of this dataset is:

```
B_data.shape
```

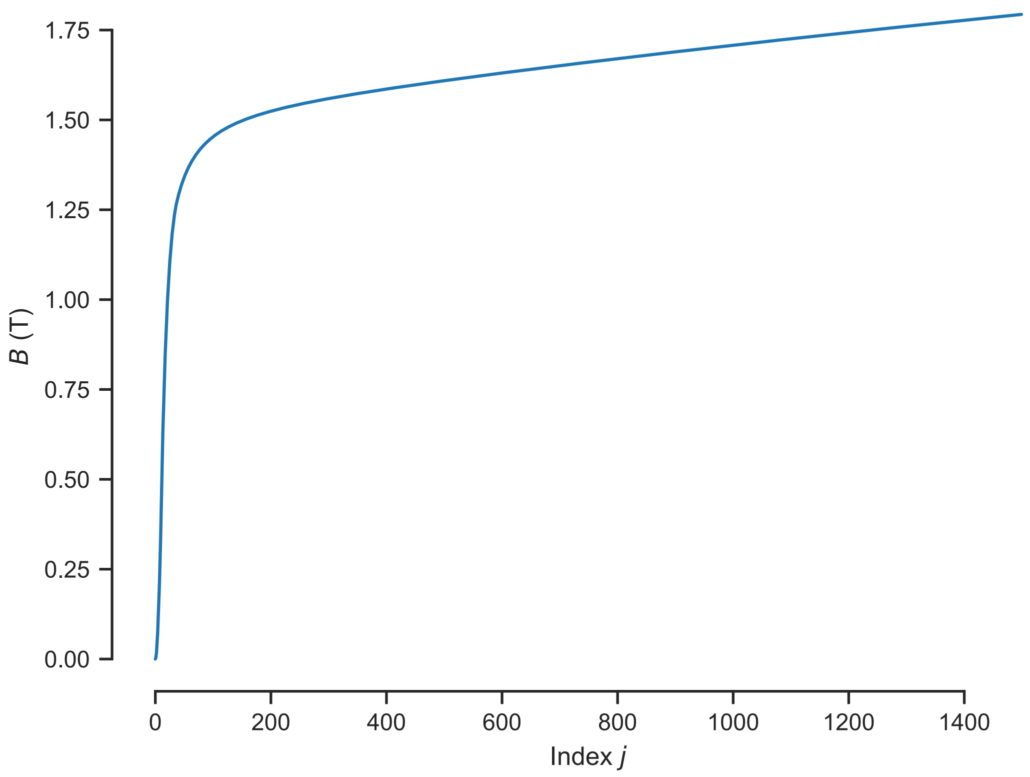
Skip to main content

```
(200, 1500)
```

The rows (200) correspond to different samples of the $B-H$ curves (suppliers and times). The columns (1500) correspond to different values of H. That is, the i, j element is the value of B at the specific value of H, say H_j. The values of H are equidistant and identical; we will ignore them in this analysis. Let's visualize some of the samples.

Here is one sample:

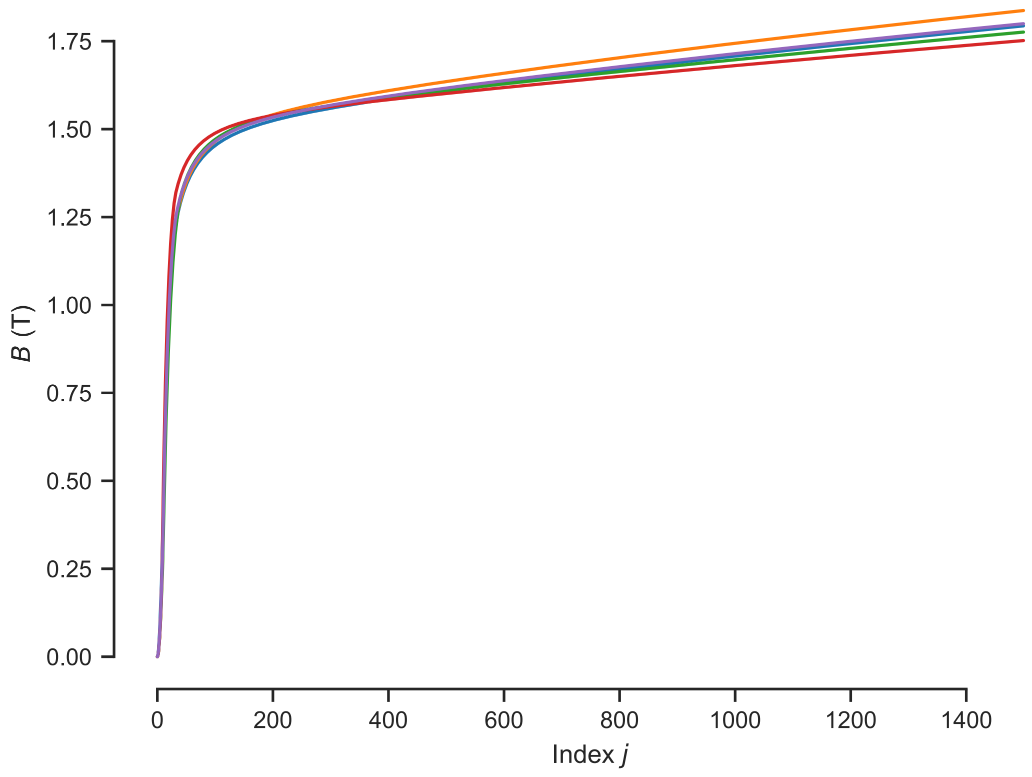
▶ Show code cell source



Here are five samples:

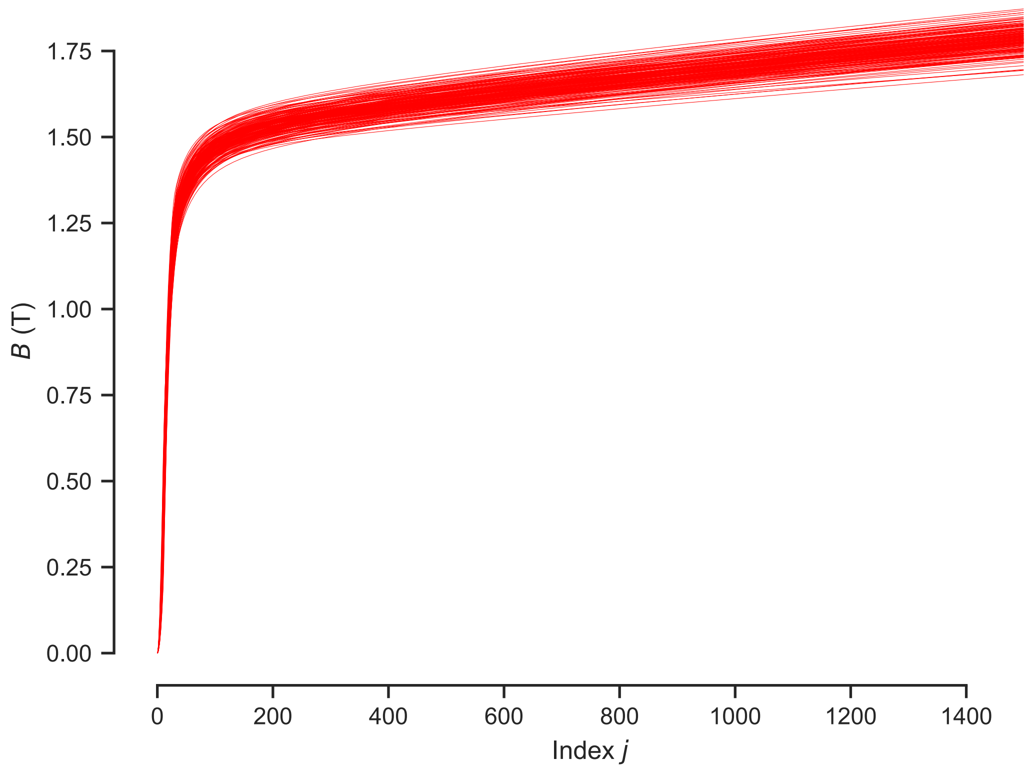
▶ Show code cell source

Skip to main content



Here are all the samples:

▶ Show code cell source



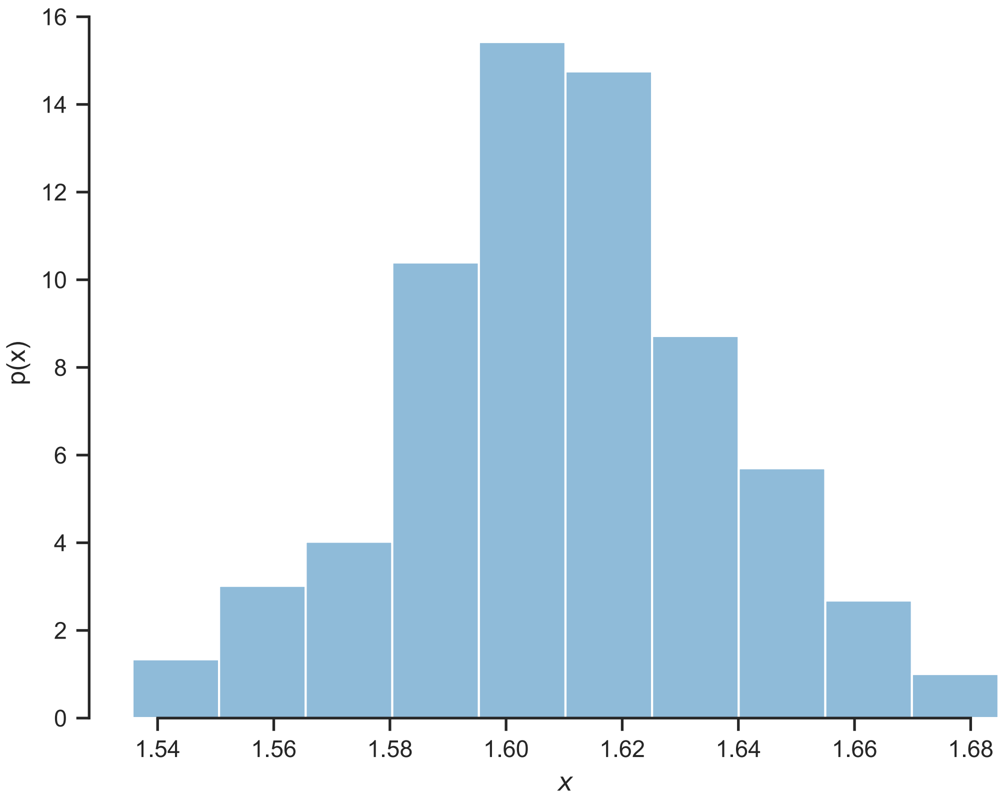
Skip to main content

A. We are going to start by studying the data at only one index. Say index $j = 500$. Let's define a random variable

$$X = B(H_{500}),$$

for this reason. Extract and do a histogram of the data for X:

▶ Show code cell source



This looks like a Gaussian $N(\mu_{500}, \sigma_{500}^2)$. Let's try to find a mean and variance for that Gaussian. A good choice for the mean is the empirical average of the data:

$$\mu_j = \frac{1}{N}\sum_{i=1}^N B_{ij}.$$

By the law of large numbers, this is a good approximation of the true mean as $N \to \infty$. Later we will learn that this is also the *maximum likelihood* estimate of the mean.

So, the mean is:

Skip to main content

```
print(f"mu_500 = {mu_500:.2f}")
```

```
mu_500 = 1.61
```

Similarly, for the variance a good choice is the empirical variance defined by:

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^N (B_{ij}-\mu_j)^2.$$

This also converges to the true variance as $N\rightarrow\infty$. Here it is:

```
sigma2_500 = np.var(X_data)
print(f"sigma_500 = {sigma2_500:.2e}")
```

```
sigma_500 = 7.42e-04
```

Repeat the plot of the histogram of X along with the PDF of the normal variable we have just identified using the functionality of `scipy.stats`.

```
# Your code here
```

B. Using your normal approximation to the PDF of X, find the probability that $X = B(H_{500})$ is greater than 1.66 T.

```
# Your code here
```

C. Let us now consider another random variable

$$Y = B(H_{1000}).$$

Isolate the data for this as well:

```
Y_data = B_data[:, 1000]
```

Skip to main content

```python
# Your code here
```

D. From the scatter plot, it looks like the random vector

$$\mathbf{X} = (X, Y),$$

follows a multivariate normal distribution. What would be the mean and covariance of the distribution? First, organize the samples of X and Y in a matrix with the number of rows being the number of samples and two columns (one corresponding to X and one to Y).

```python
XY_data = np.hstack([X_data[:, None], Y_data[:, None]])
```

In case you are wondering, the code above takes two 1D numpy arrays of the same size and puts them in a two-column numpy array. The first column is the first array, the second column is the second array. The result is a 2D numpy array. We take sampling averages over the first axis of the array.

The mean vector is:

```python
mu_XY = np.mean(XY_data, axis=0)
print(f"mu_XY = {mu_XY}")
```

```
mu_XY = [1.61041566 1.70263681]
```

The covariance matrix is trickier. We have already discussed how to find the diagonals of the covariance matrix (it is simply the variance). For the off-diagonal terms, this is the formula that is being used:

$$C_{jk} = \frac{1}{N}\sum_{i=1}^{N}(B_{ij} - \mu_j)(B_{ik} - \mu_k).$$

This formula converges as $N \to \infty$. Here is the implementation:

```python
# Careful with np.cov because it requires you to transpose the matrix we defined in cl
C_XY = np.cov(XY_data.T)
```

Skip to main content

```
print(f"C_XY =")
print(C_XY)
```

```
C_XY =
[[0.00074572 0.00082435]
 [0.00082435 0.00096729]]
```

Use the covariance matrix `C_XY` to find the correlation coefficient between X and Y.

```
# Your code here
```

Are the two variables X and Y positively or negatively correlated?

Answer:

E. Use `np.linalg.eigh` to check that the matrix `C_XY` is indeed positive definite.

```
# Your code here
```

F. Use the functionality of `scipy.stats.multivariate_normal` to plot the joint probability function of the samples of X and Y in the same plot as the scatter plot of X and Y.

```
# Your code here
```

G. Now, consider each $B-H$ curve a random vector. That is, the random vector $\mathbf{B}$ corresponds to the magnetic flux density values at a fixed number of H-values. It is:

$$\mathbf{B} = (B(H_1), \dots, B(H_{1500})).$$

It is like $\mathbf{X} = (X, Y)$ only now we have 1,500 dimensions instead of 2.

First, let's find the mean of this random vector:

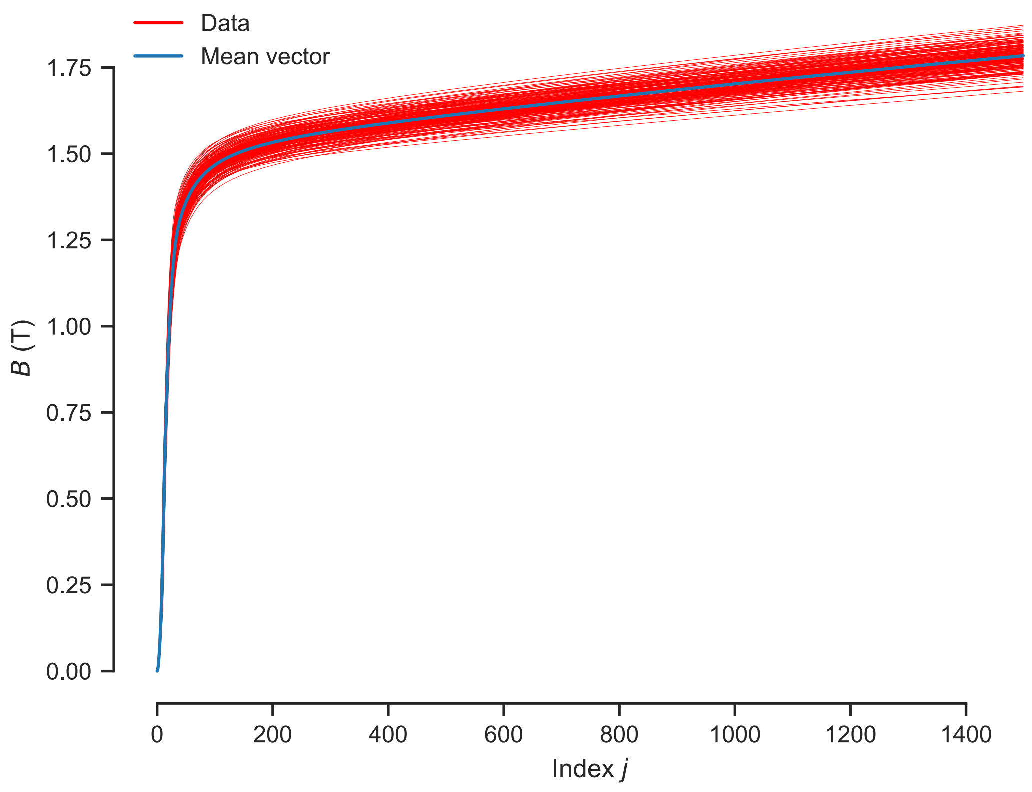
```
B_mu = np.mean(B_data, axis=0)
```

Skip to main content

```
array([0.        , 0.00385192, 0.01517452, ..., 1.78373703, 1.78389267,
       1.78404828])
```

Let’s plot the mean on top of all the data we have:

▶ Show code cell source



It looks good. Now, find the covariance matrix of $\mathbf{B}$. This is going to be a 1500x1500 matrix.

```
B_cov = np.cov(B_data.T)
B_cov
```

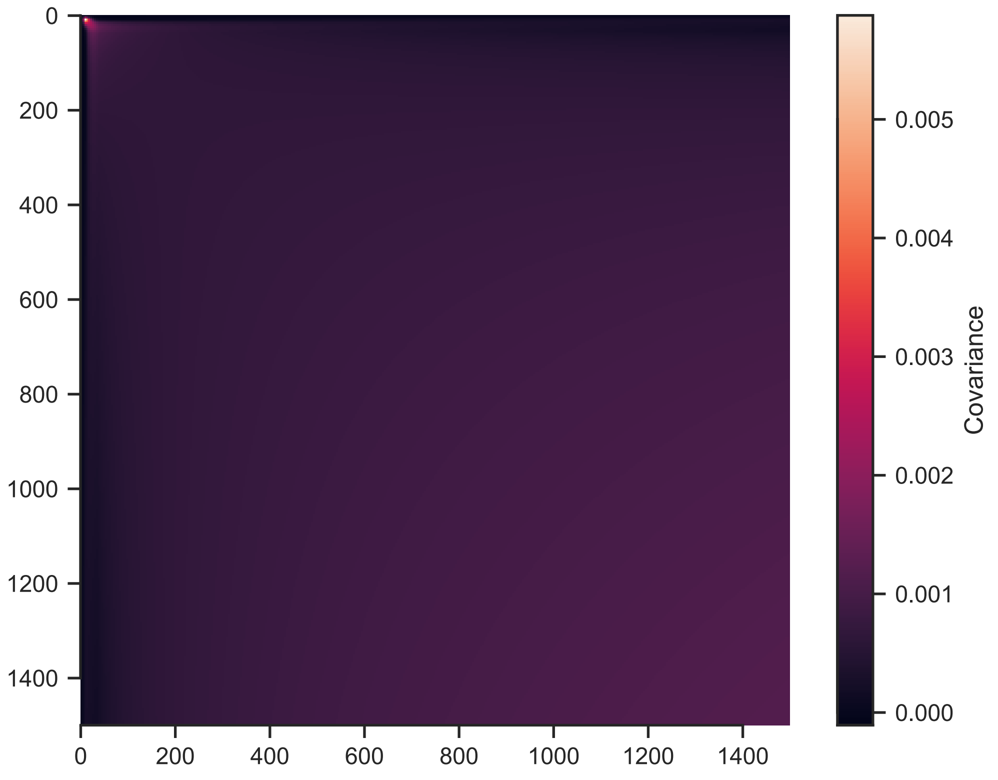
```
array([[0.00000000e+00, 0.00000000e+00, 0.00000000e+00, ...,
        0.00000000e+00, 0.00000000e+00, 0.00000000e+00],
       [0.00000000e+00, 1.16277948e-06, 4.41977479e-06, ...,
        3.18233676e-06, 3.18391580e-06, 3.18549316e-06],
       [0.00000000e+00, 4.41977479e-06, 1.68041482e-05, ...,
        1.22832828e-05, 1.22890907e-05, 1.22948922e-05],
       ...,
       [0.00000000e+00, 3.18233676e-06, 1.22832828e-05, ...,
        1.20268920e-03, 1.20293022e-03, 1.20317114e-03],
       [0.00000000e+00, 3.18391580e-06, 1.22890907e-05, ...,
        1.20293022e-03, 1.20317134e-03, 1.20341237e-03],
```

Skip to main content

```
        [0.00000000e+00, 3.18549316e-06, 1.22948922e-05, ...,
         1.20317114e-03, 1.20341237e-03, 1.20365351e-03]])
```

Let’s plot this matrix:

▶ Show code cell source



The numbers are very small. This is because the covariance depends on the units of the variables. We need to do the same thing we did with the correlation coefficient: divide by the standard deviations of the variables. Here is how you can get the correlation coefficients:

```
# Note that I have to remove the first point because it is always zero
# and it has zero variance.
B_corr = np.corrcoef(B_data[:,1:].T)
B_corr
```

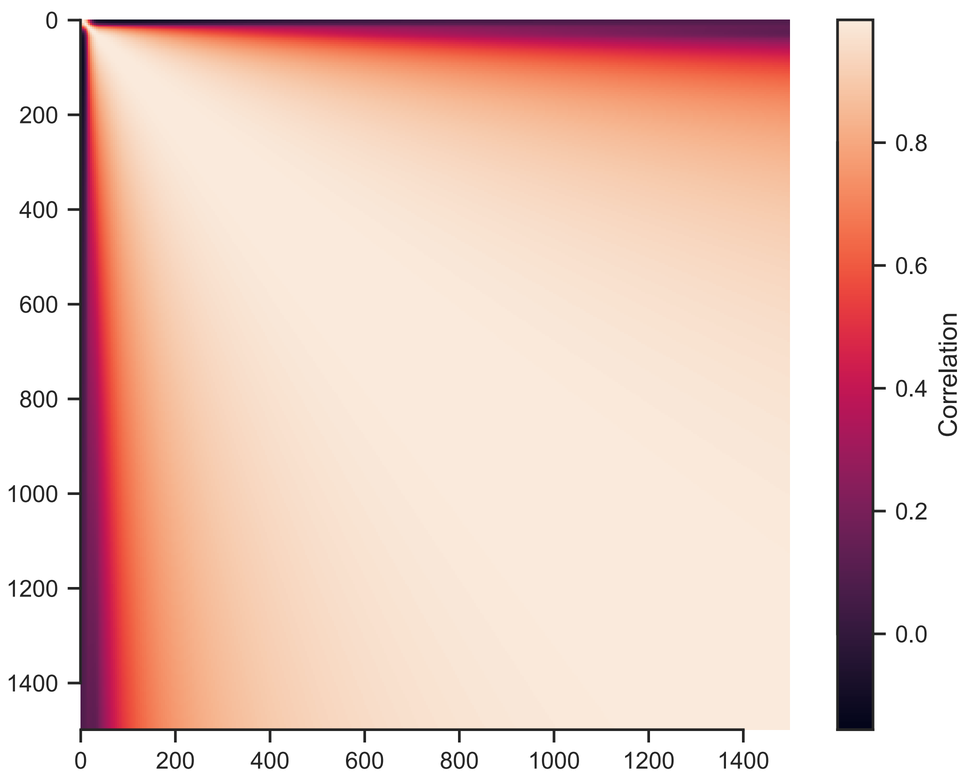
```
array([[1.        , 0.99986924, 0.99941799, ..., 0.08509827, 0.08512344,
         0.08514855],
       [0.99986924, 1.        , 0.99983894, ..., 0.08640313, 0.08642667,
         0.08645015],
       [0.99941799, 0.99983894, 1.        , ..., 0.08782484, 0.08784655,
         0.08786822],
       ...,
```

Skip to main content

```
       [0.08512344, 0.08642667, 0.08784655, ..., 0.99999998, 1.        ,
        0.99999998],
       [0.08514855, 0.08645015, 0.08786822, ..., 0.9999999 , 0.99999998,
        1.        ]])
```

Here is the correlation visualized:

```
fig, ax = plt.subplots()
c = ax.imshow(B_corr, interpolation='nearest')
plt.colorbar(c, label="Correlation")
sns.despine(trim=True);
```



The values are quite a bit correlated. This makes sense because the curves are all very smooth and look very much alike.

Let's check if the covariance is indeed positive definite:

```
print("Eigenvalues of B_cov:")
print(np.linalg.eigh(B_cov)[0])
```

```
Eigenvalues of B_cov:
[-3.28311404e-16 -1.83517775e-16 -9.47916665e-17 ...  4.66244763e-02
```

Skip to main content

Notice that several eigenvalues are negative, but they are too small. Very close to zero. This happens often in practice when you are finding the covariance of large random vectors. It arises from the fact that we use floating-point arithmetic instead of real numbers. It is a numerical artifact. If you tried to use this covariance to make a multivariate average random vector using `scipy.stats` it would fail. Try this:

```
B = st.multivariate_normal(mean=B_mu, cov=B_cov)
```

▶ Show code cell output

The way to overcome this problem is to add a small positive number to the diagonal. This needs to be very small so that the distribution stays mostly the same. It must be the smallest possible number that makes the covariance matrix behave well. This is known as the *jitter* or the *nugget*. Find the nugget playing with the code below. Every time you try, multiply the nugget by ten.

```
# Pick the nugget here
nugget = 1e-12
# This is the modified covariance matrix
B_cov_w_nugget = B_cov + nugget * np.eye(B_cov.shape[0])
# Try building the distribution:
try:
    B = st.multivariate_normal(mean=B_mu, cov=B_cov_w_nugget)
    print('It worked! Move on.')
except:
    print('It did not work. Increase nugget by 10.')
```

```
It did not work. Increase nugget by 10.
```

H. Now, you have created your first stochastic model of a complicated physical quantity. By sampling from your newly constructed random vector $\mathbf{B}$, you have essentially quantified your uncertainty about the $B-H$ curve as induced by the inability to control steel production perfectly. Take ten samples of this random vector and plot them.

```
# Your code here
```

Congratulations! You have made your first stochastic model of a physical field quantity. You can now sample $B-H$ curves in a way that honors the manufacturing uncertainties. This is the first step in uncertainty quantification studies. The next step would be to propagate these samples

Skip to main content

machine. If you want to see how that looks, look at [Sahu *et al.*, 2020] and [Beltrán-Pulido *et al.*, 2020].

Previous
Homework 1

Next
Homework 3