Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

3E3 PROBABILITY AND STATISTICS LABORATORY

Department of Electronic and Electrical Engineering

(MATLAB e-Report submission)

Assignments:

1. Homogeneous Markov chain modelling of a human gene DNA base sequence
2. Empirical and Exponential modelling of survival data in a large mechanical assembly
3. Empirical and normal modelling of multidimensional biosensor data

Lab Date: 15/10/2019

Submitted by:

Rohan Taneja
19323238

Assignment 1:
Homogeneous Markov chain modelling of a human gene DNA base sequence

Assignment stats are provided as follows:

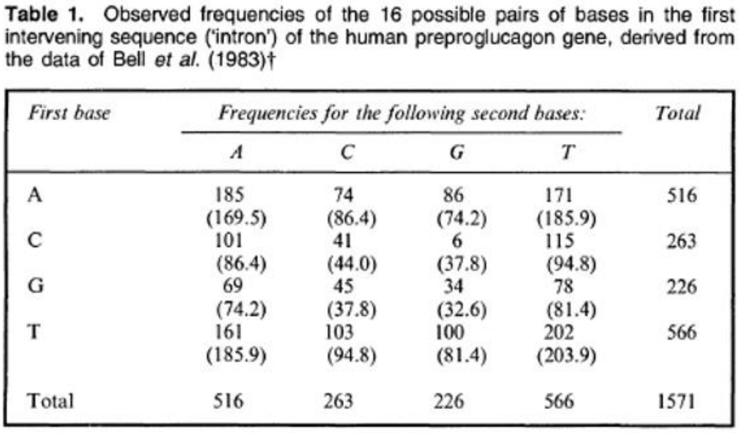
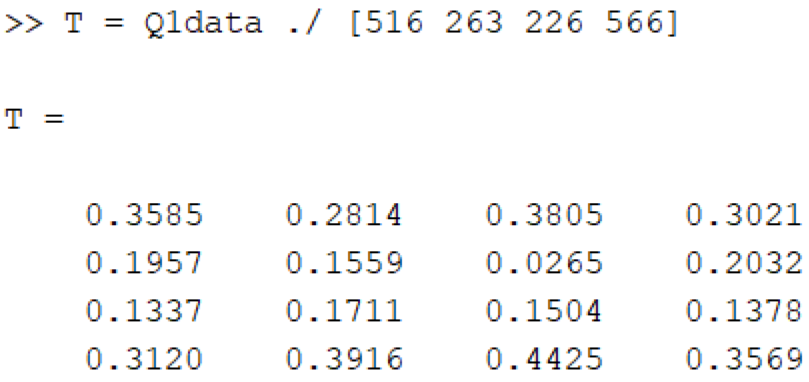


Table 1. Observed frequencies of the 16 possible pairs of bases in the first intervening sequence ('intron') of the human preproglucagon gene, derived from the data of Bell *et al.* (1983)†

First base	*Frequencies for the following second bases:*				*Total*
	A	*C*	*G*	*T*	
A	185 (169.5)	74 (86.4)	86 (74.2)	171 (185.9)	516
C	101 (86.4)	41 (44.0)	6 (37.8)	115 (94.8)	263
G	69 (74.2)	45 (37.8)	34 (32.6)	78 (81.4)	226
T	161 (185.9)	103 (94.8)	100 (81.4)	202 (203.9)	566
Total	516	263	226	566	1571

Figure 1: Cooccurrence statistics in a human gene DNA base sequence [1].

i) In this task, with given total. We get the count N = total + 1, i.e. N = 1571 + 1 = 1572.
The transition probability matrix generated is as follows:



```
>> T = Q1data ./ [516 263 226 566]

T =

    0.3585    0.2814    0.3805    0.3021
    0.1957    0.1559    0.0265    0.2032
    0.1337    0.1711    0.1504    0.1378
    0.3120    0.3916    0.4425    0.3569
```

Figure 1.1 – Transition Probability Matrix (T) of the HMC

Here, the total probability vector i.e. $\mathbf{p}_i$ **= [516 263 226 566]**, used for generation of Transition Probability Matrix.

The initial probability vector i.e. $\mathbf{p}_1\mathbf{X}$ (where X = {A, C, G, T}) can be taken at random as we are not specified with selection. Hence, after selection it's useful for modelling dynamics of the sequence using HMC.

ii) In this task, we modelled the transition probability matrix T as a graphical model. The plotted direct graph is as follows:

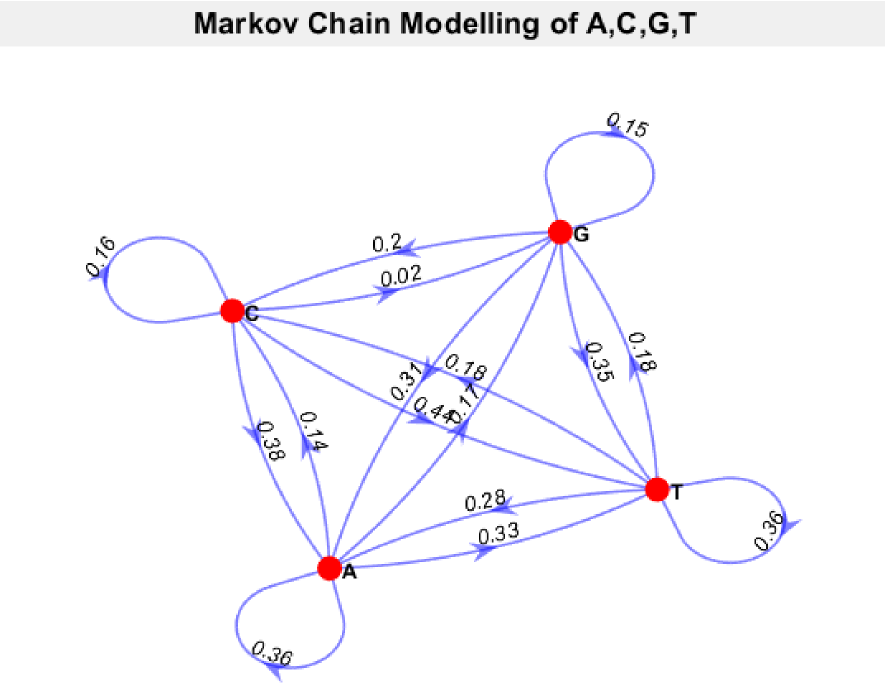


Figure 1.2 – Markov Chain – Directed Graphical Plot

After verification of data, the plotted directed graph is accurate up to 2 decimal places.

iii) In this task we generated the sequence using simulate function and then using stairs created the DNA sequence as follows. The simulate() function used generated random sequence based on the prior matrix selected and the generated markovChain of the Q1data.
stair() used to project the sequence gives step graph of the output.

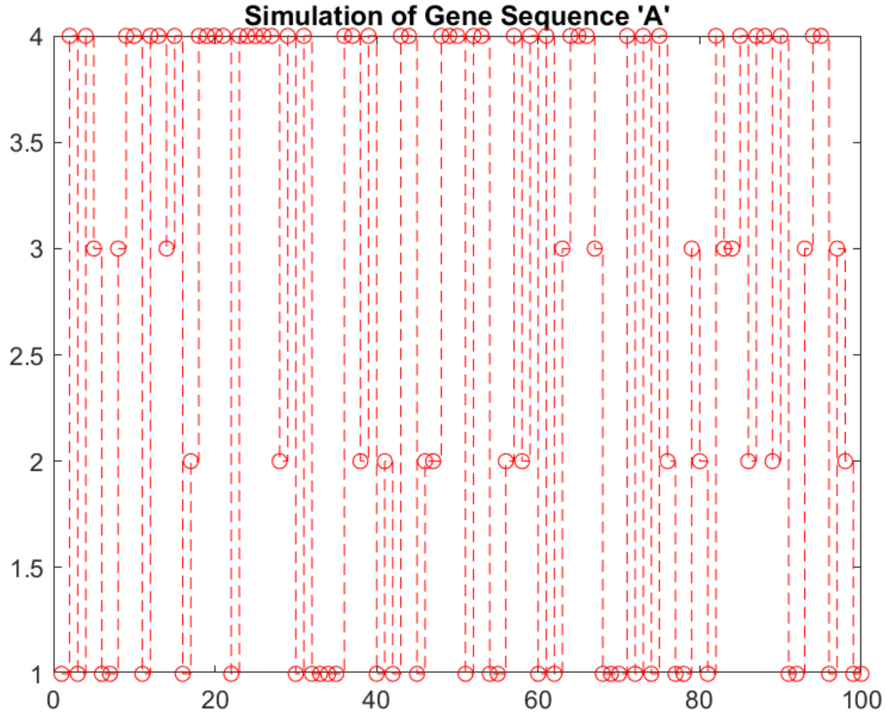


Figure 1.3.1 – Simulation of DNA Sequence ‘A’

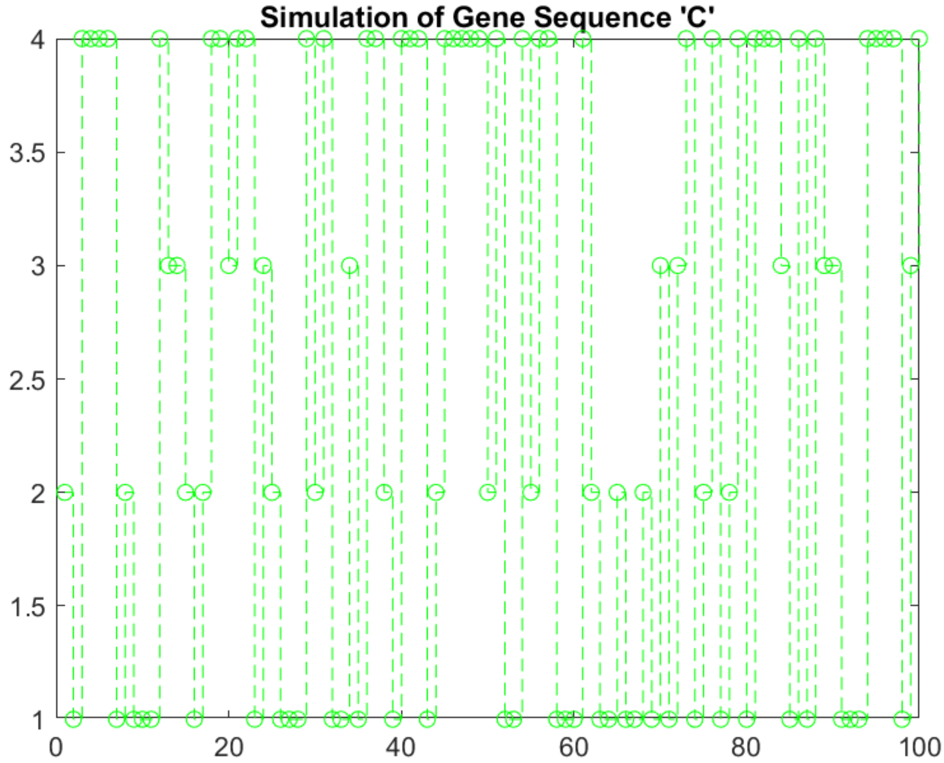


Figure 1.3.2 – Simulation of DNA Sequence ‘C’

iv) In this task, Marginal Probability for i^{th} bases was generated as following graphical representation:

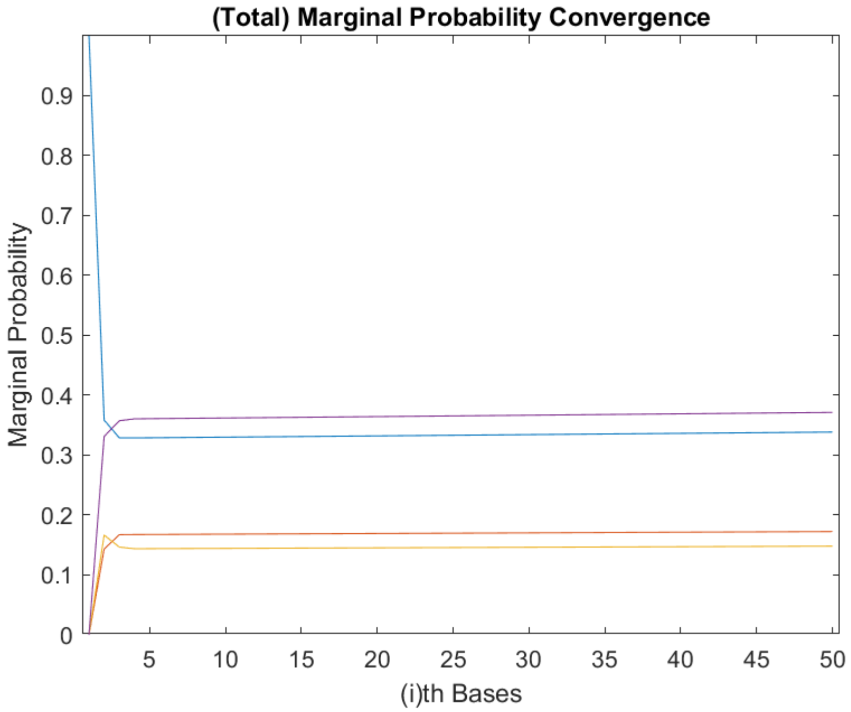


Figure 1.4 – *Marginal Probability v/s* i^{th} *bases* plot depicting HMC behaviour

The marginal probability converges to stable marginal probability over long run of HMC. Also, for the task prior probability matrix assumed was **p1A** i.e. = **[1 0 0 0]**.

v) In this task, k-step-ahead T for taken HMC model was evaluated. The results are plotted as follows in graphical representation. (taken upto, k = 6).

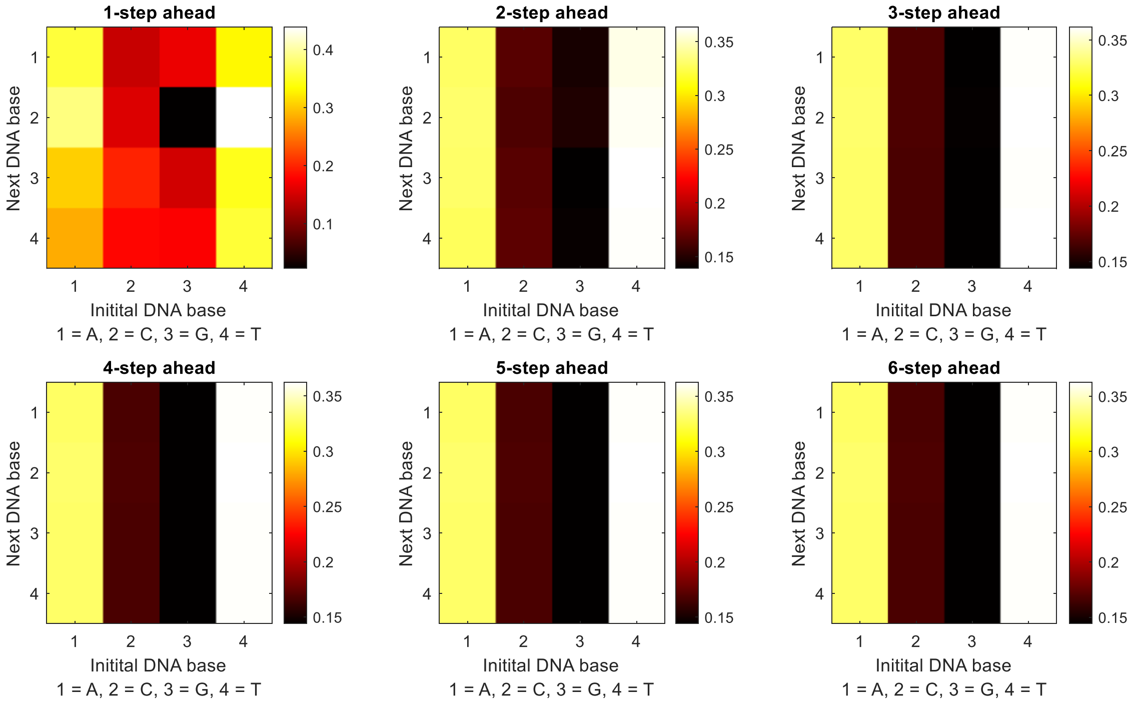


Figure 1.5 – *colormap(hot())* plot for k-step ahead transition

Compared to task (iv), the convergence rate of this task is higher than previous one. As seen in the colormap, generated at 3-step ahead the colours are uniform from this stage forward.

vi) In this task, the probability evaluated is as follows. First, we find the probability of next 4 bases are going to be all A.

i.e. **Pr** $[B_{i+4} = A \mid B_{i=1,2,3} = A]$ = **0.016523**

vii) In this task, using the above probability we found out the marginal probability vector. By using **p1A** as the prior probability vector with its product with **4th T** matrix. And we get,

Pr $[B_i = A \mid B_{i+4} = A]$ = **0.3285**

Assignment 2:
Empirical and Exponential modelling of survival data in a large mechanical assembly

i) In this task, we plotted bar chart for the provided data in the Q2stats.mat file. With inclusion of the data in the assignment. The graphical representation is as follows:

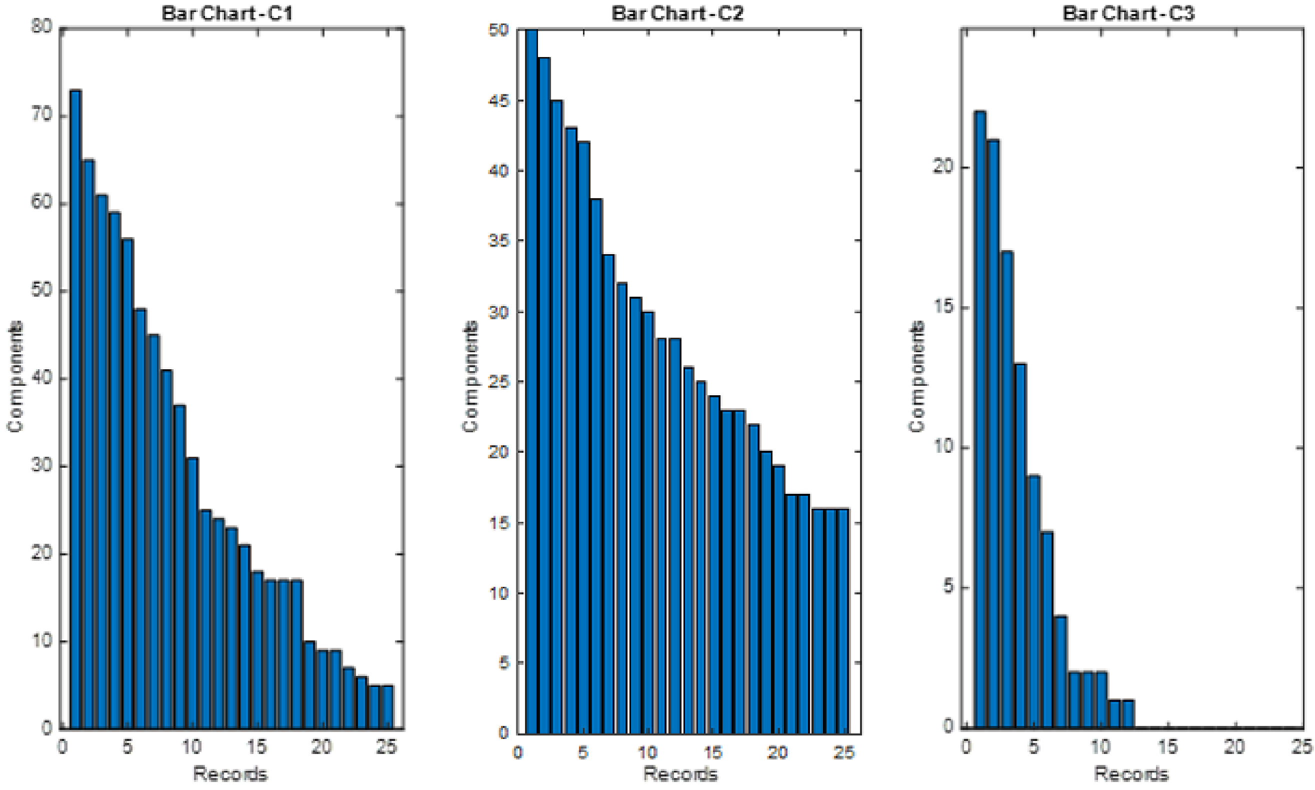


Figure 2.1 – *Components v/s Record* plot for each component

From the **Components v/s Record** chart, we can clearly observe that C2 has the most success rate amongst the three components. And the worst success rate is shown by component C3.

ii) In this task, failed components are plotted in bar chart and scatter plot. The visualization is as follows:

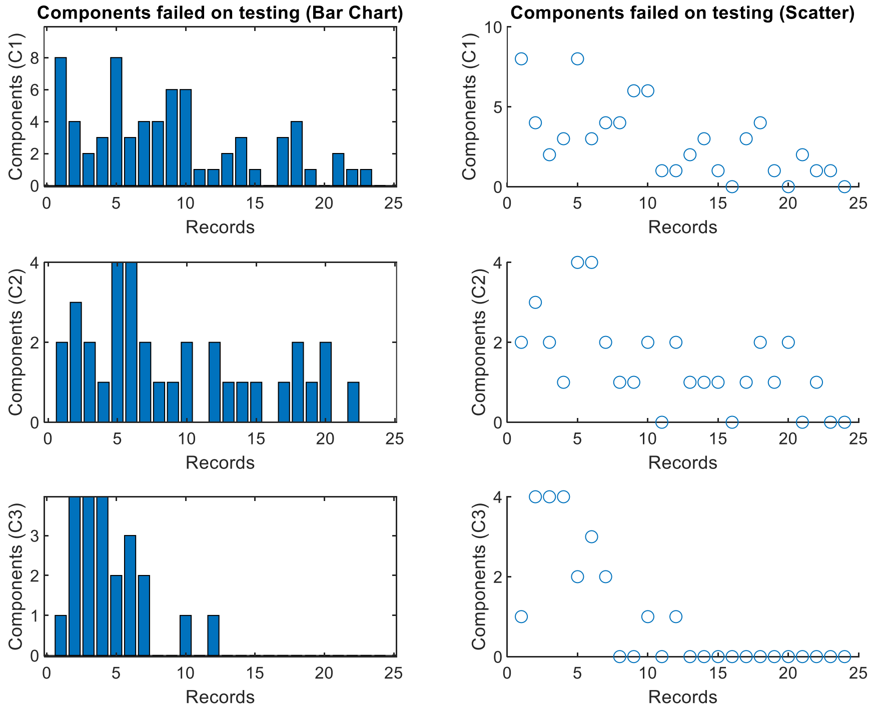


Figure 2.2 – *Components v/s Record* plot for each component

iii) In this task, **T** denotes unknown time to failure. And **I** denote interval of 6 months as natural numbers. So, the relation can be taken as:

$$I = floor(T/6) + 1$$

Here, T have units of month which is cancelled out in the floor(T/6). Hence, giving a constant which belongs to set of I.

iv) In this task, marginal probability mass function of all the three components, **C** ∈ **{C1, C2, C3}**.

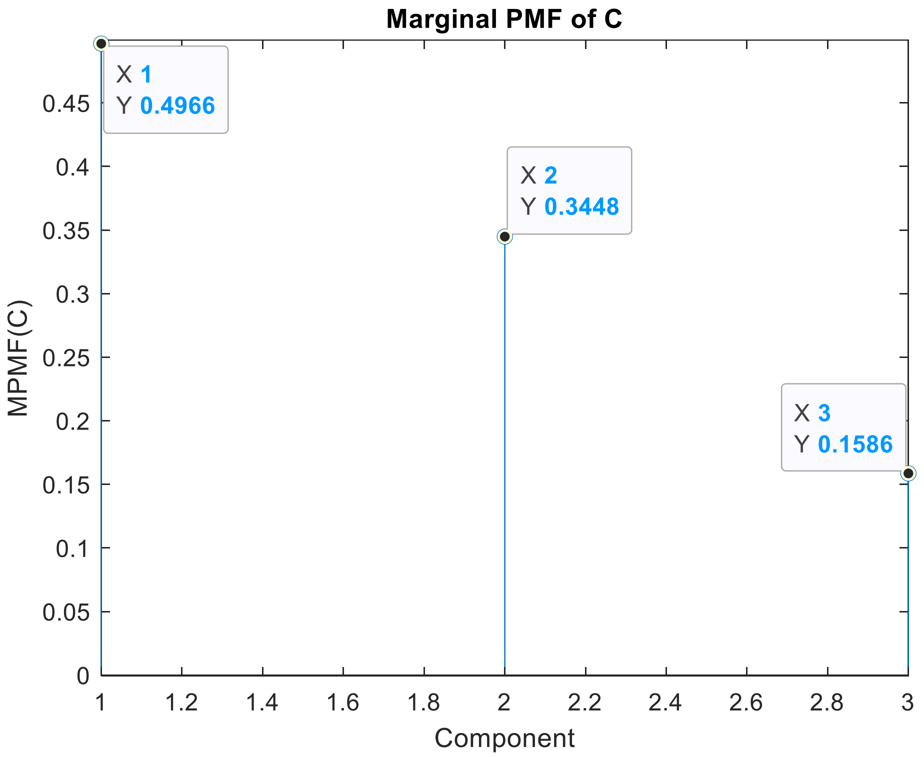


Figure 2.4 – *Marginal PMF(C) v/s each component* plot for each component

v) In this task, we projected the Joint Probability distribution of **Q2failtimes**, which consist of data of failed components over 24 records. The components working beyond 12 years won't contribute to this PMF because the record set provided depicts only the working components tested over the time period. Where only active components contribute in modelling the probability for fail times, which difference of initial components with the currently active ones.

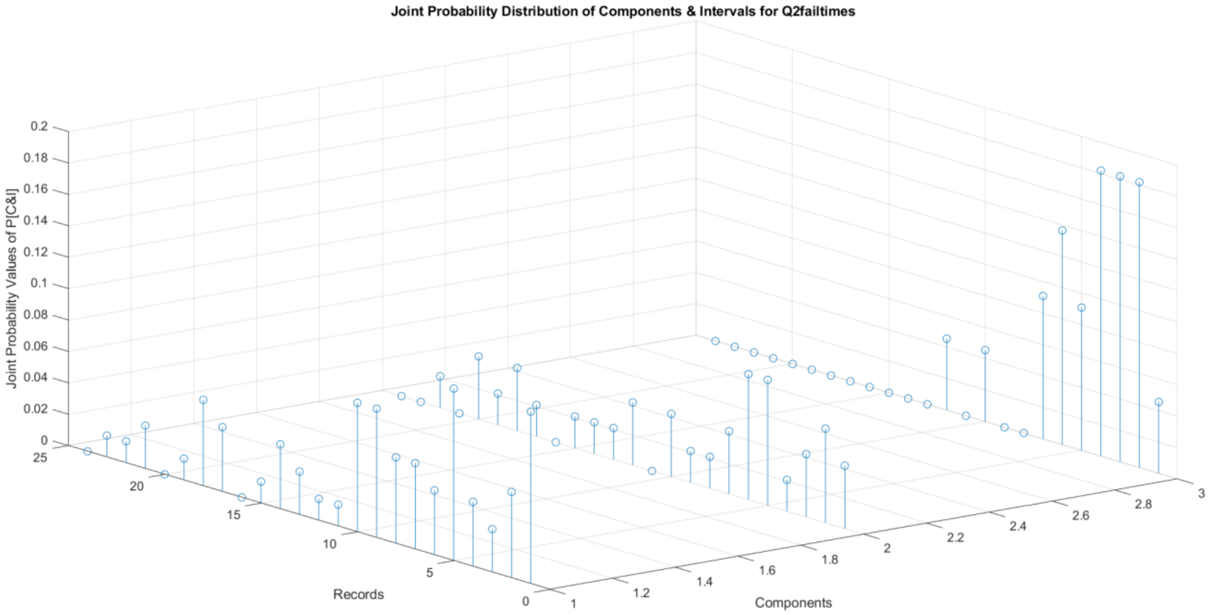


Figure 2.5 – *Joint PMF(C) of each component's v/s record* stem3 plot

vi) In this task, conditional probability of intervals conditioned to each component is plotted as follows. The Conditional Probability v/s Record stem graph depicts the comparison in one figure.

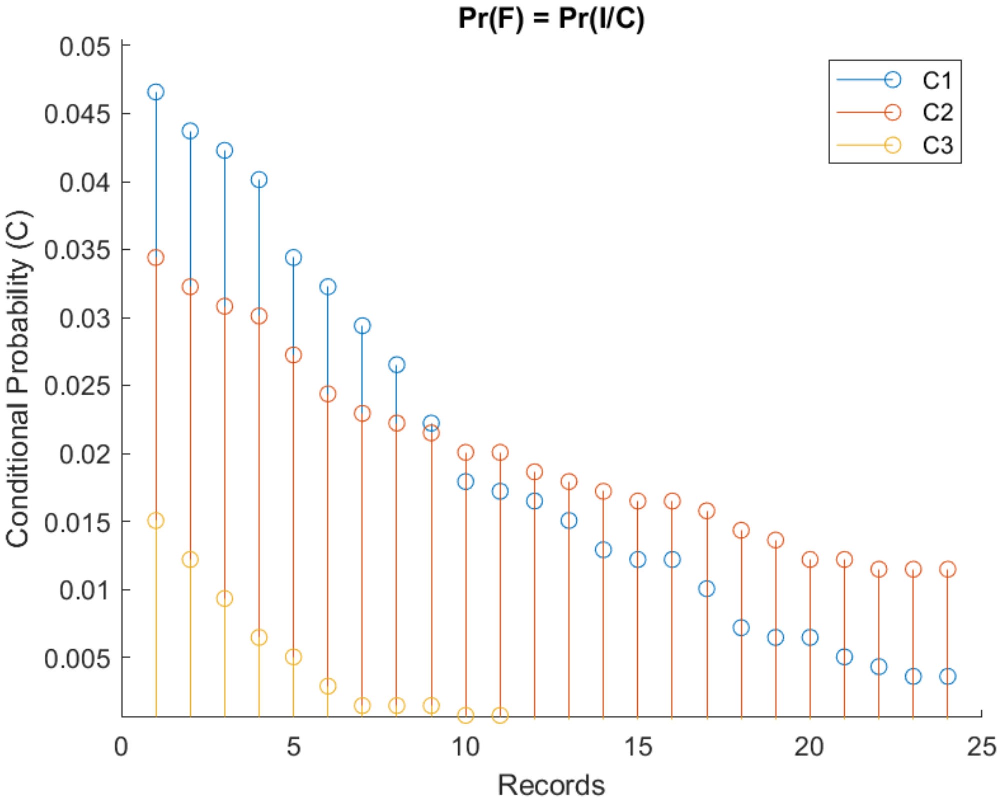


Figure 2.6 – *Conditional PMF(C) of each component's v/s record* stem plot

vii) In this task, the marginal PMF of Interval v/s Record is plotted as follows:

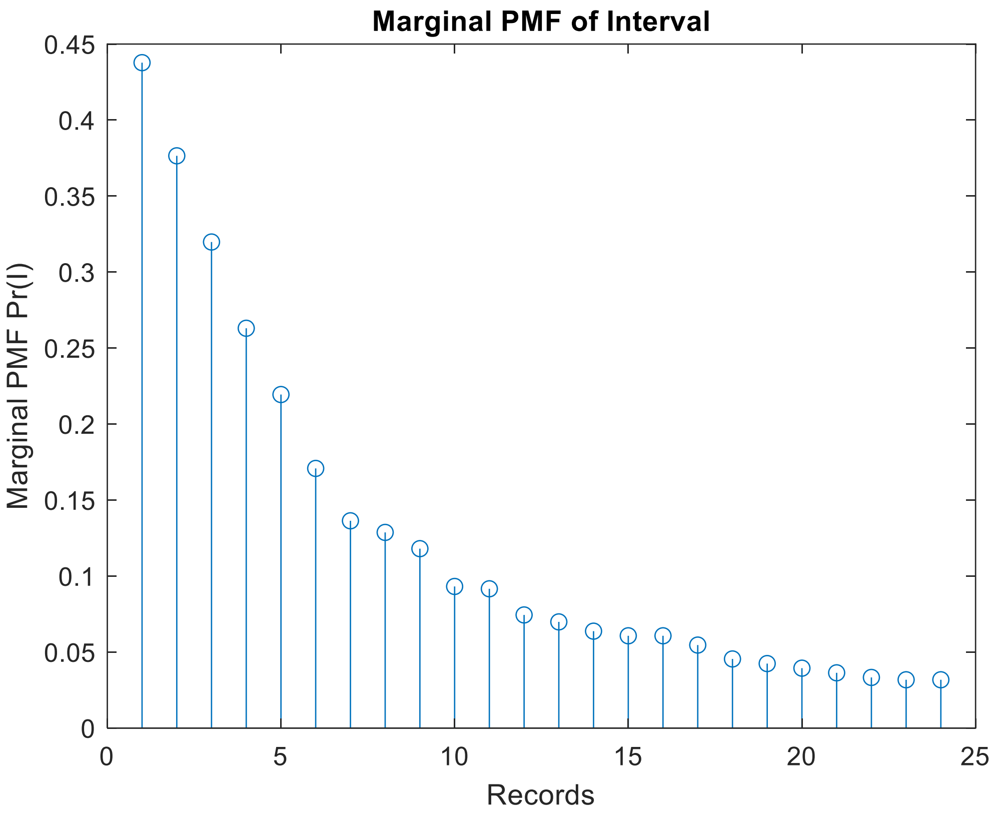


Figure 2.7 – *Marginal PMF(I) v/s record* stem plot

viii) In this task, taking PMF from (vii), and the joint PMF from (v). We get the following result.

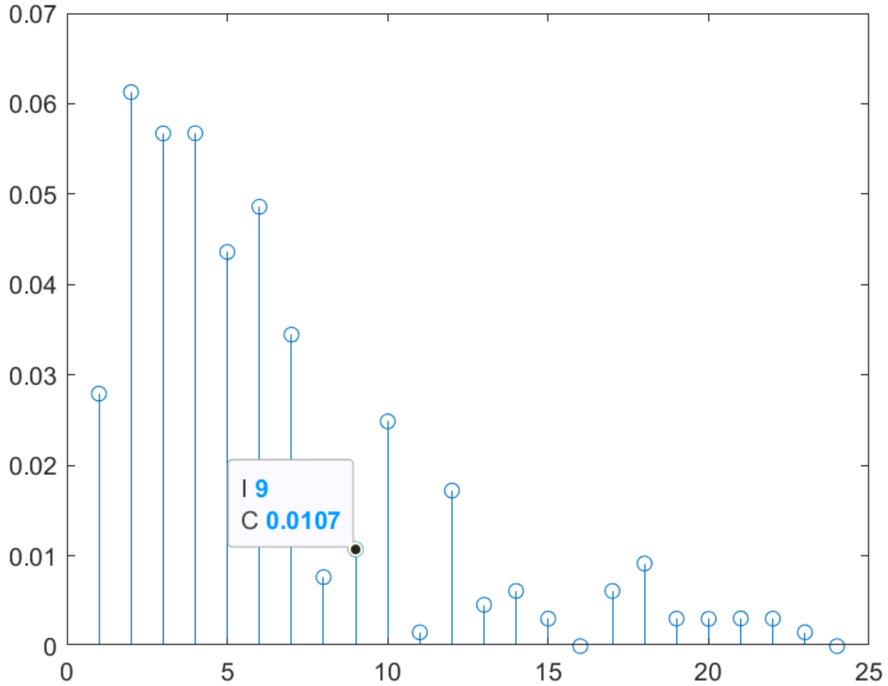


Figure 2.8 – Probability of *C given I over given records*

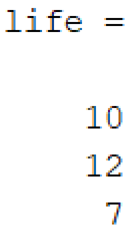
P$_{\mathbf{r}}$ [C | I $_{= 9}$] = **0.0107** (Joint Probability value)

For approximation of **C** given in *50 < T ≤ 60*, can be found out as follows:
T being unknown time to failure, with I denoted as interval of 6 months. We can translate the T by relationship as in (iii).

We get interval as, *9 < I ≤ 10*

Components	Probability Value
C$_1$ (I=9)	0.0092
C$_2$ (I=9)	0.0015
C$_3$ (I=9)	0
C$_1$ (I=10)	0.0092
C$_2$ (I=10)	0.0030
C$_3$ (I=10)	0.0127

ix) In this task, we find out the mean of data provided to us using expfit() function which is pre-defined in MATLAB. After that we compare the data with the mean life with respect to the collection of records (i.e. 24). When the data converges with expfit() mean of data, the point of convergence is taken as the expected lifetime of that component. *(units of Interval i.e. = 6 months)*



```
life =


    10
    12
     7
```

Figure 2.9.1 – *Mean Time of Failure* of each components

Using **exppdf()**, a predefined function we generated the following graph for probability density function of T, expressing the data in exponential terms. Plot for all 3 components' $\mathbf{P_r}$ **[I | C]** v/s **T** is as follows:

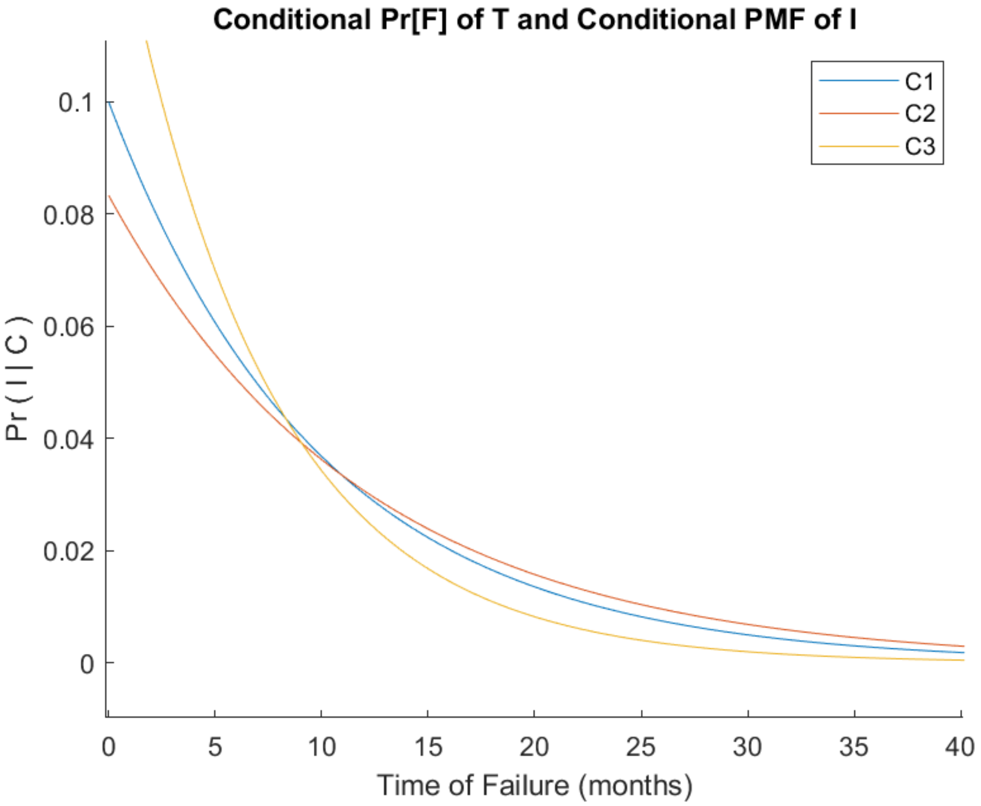


Figure 2.9.2 – *Conditional Probability of I give C v/s Time of Failure* line plot

x) In this task, we found out the lifetime of failure to be 12 by records. Hence, evaluating the data set using **exppdf()**, we are able to generate the following parametric marginal probability density of T v/s Interval graph.

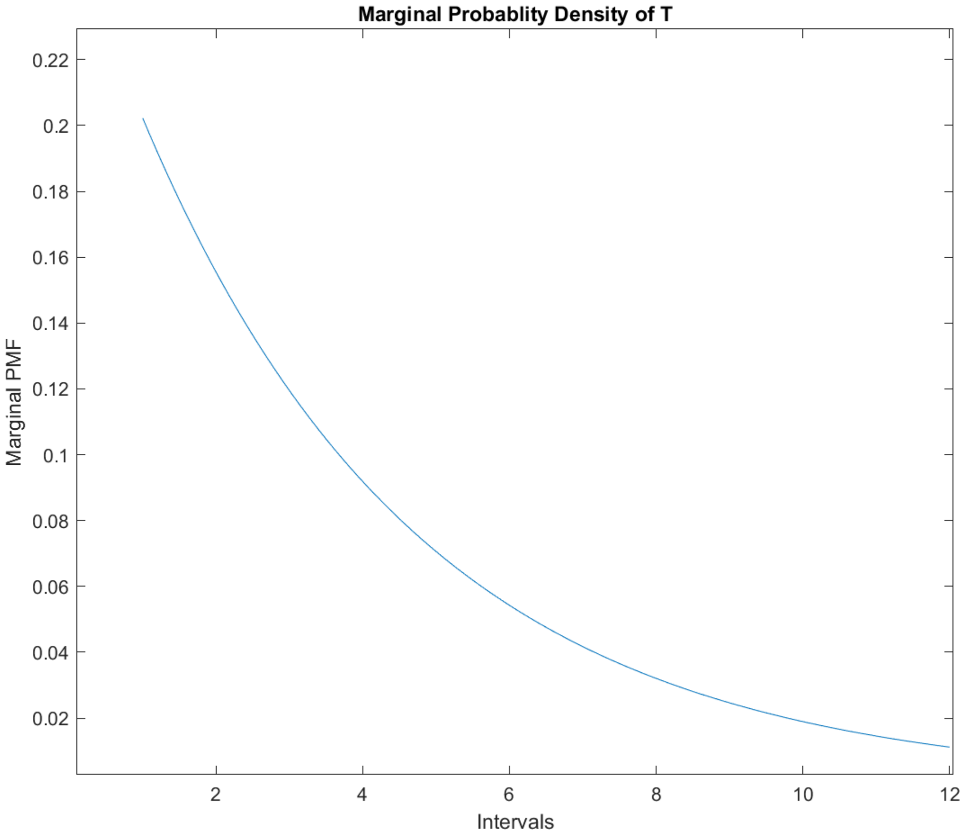


Figure 2.10 – *Marginal Probability of T* v/s *Intervals* line plot

Assignment 3:
Empirical and normal modelling of multidimensional biosensor data

i) In this task, first we plotted pairwise combination of all for **{V1,V2,V3,V4}** channels expressed as followed:

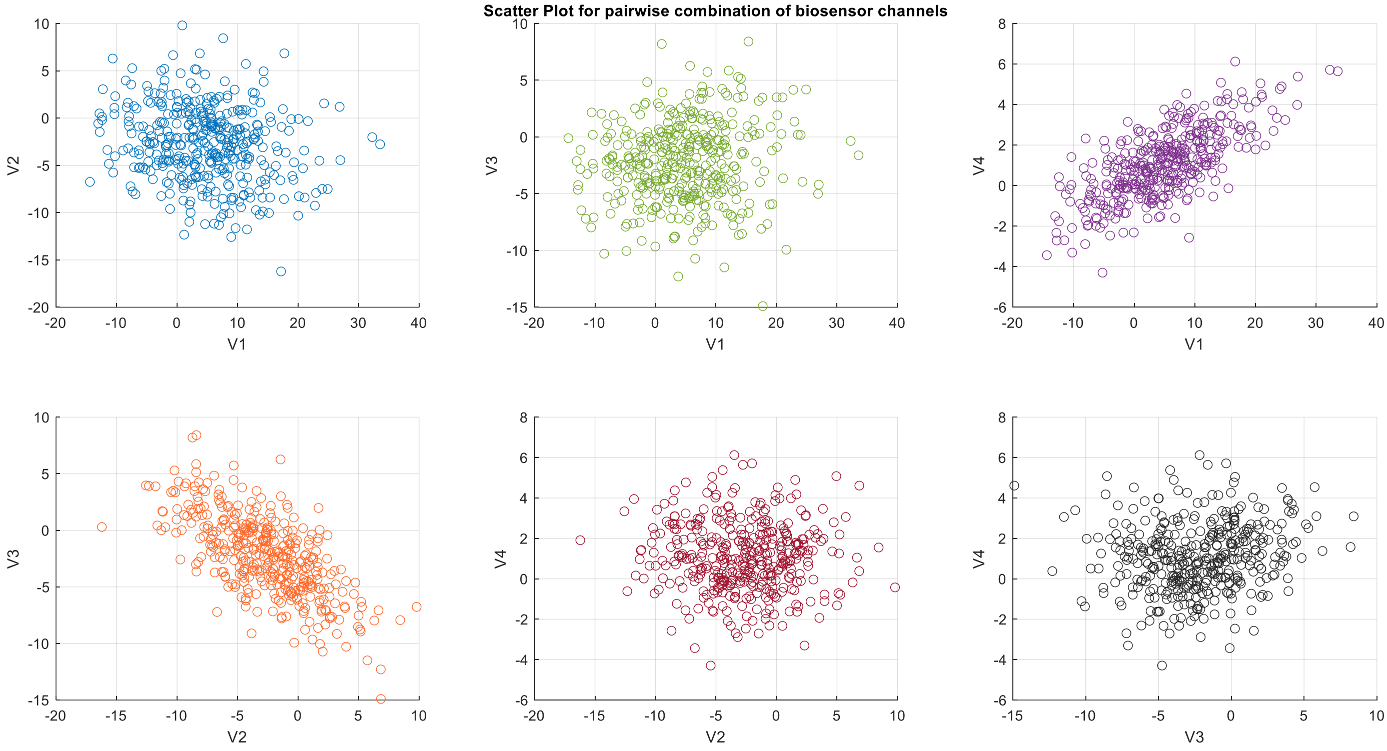


Figure 3.1.1 – Scatter plot for *pair-wise combination of biosensor channels*

Second, we plotted 3-Channel combination in scatter3 plot. Expressed the combinations in x,y,z-axis as in the plot below:

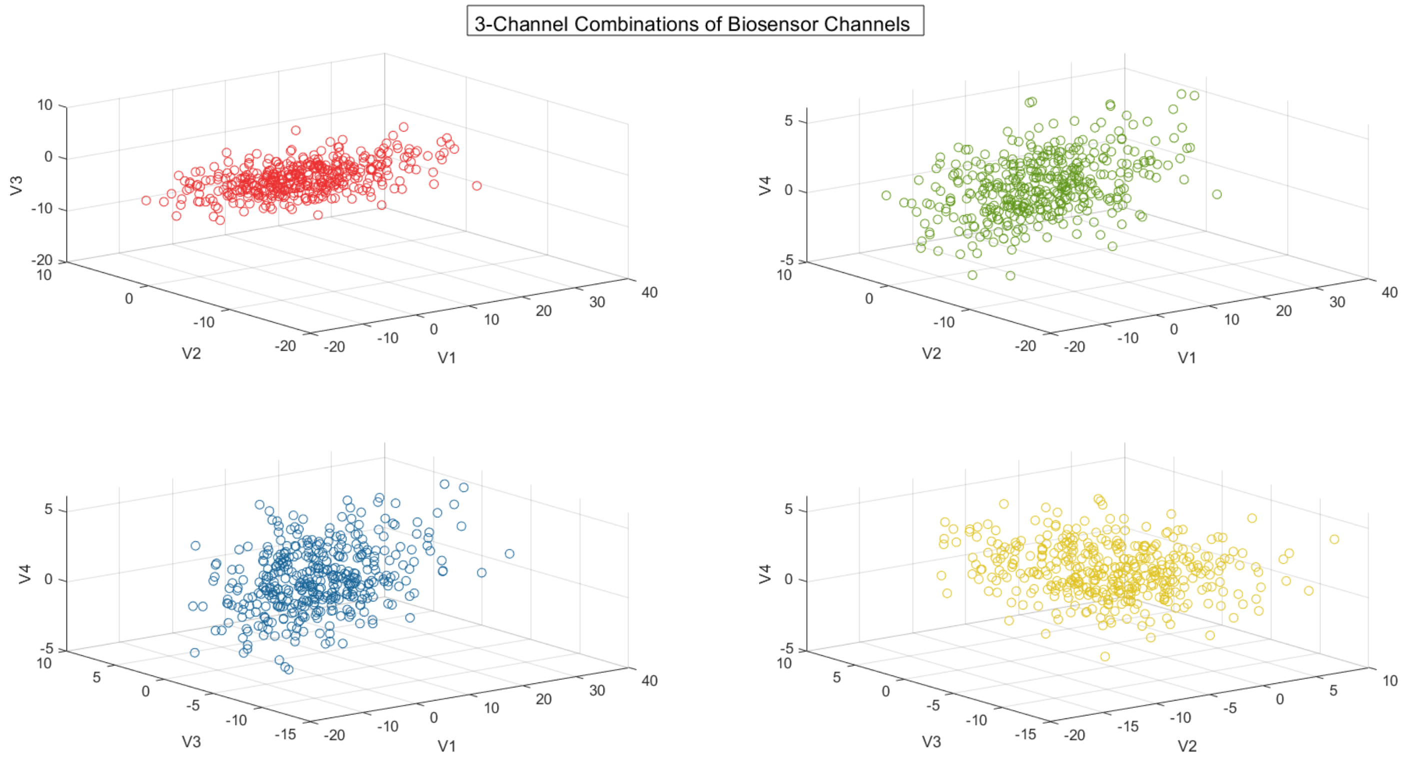


Figure 3.1.2 – scatter3() plot for *triplet combination of biosensor channels*

ii) In this task, we plotted data of the Q3stats as Histogram as follows:

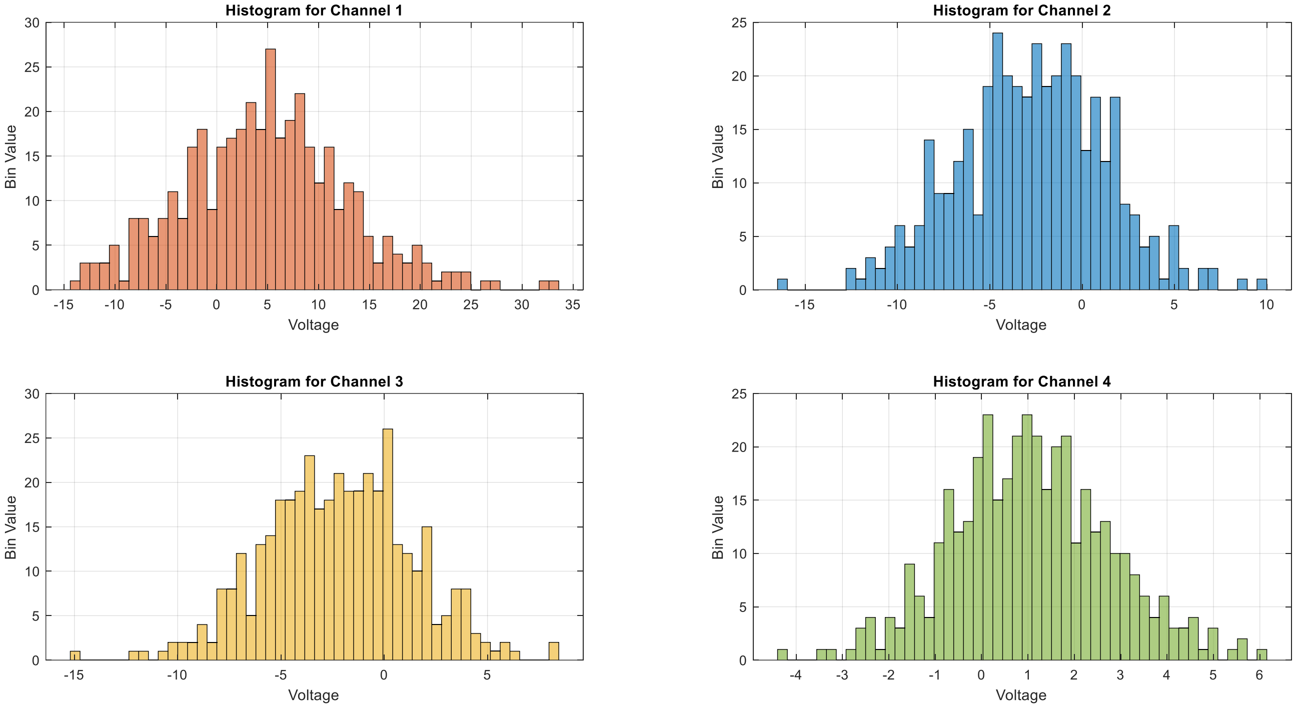


Figure 3.2 – Histogram plot for each biosensor channel

iii) In this task, we plotted pair-wise marginal distribution of unique biosensor pairs. The results are depicted as follows:

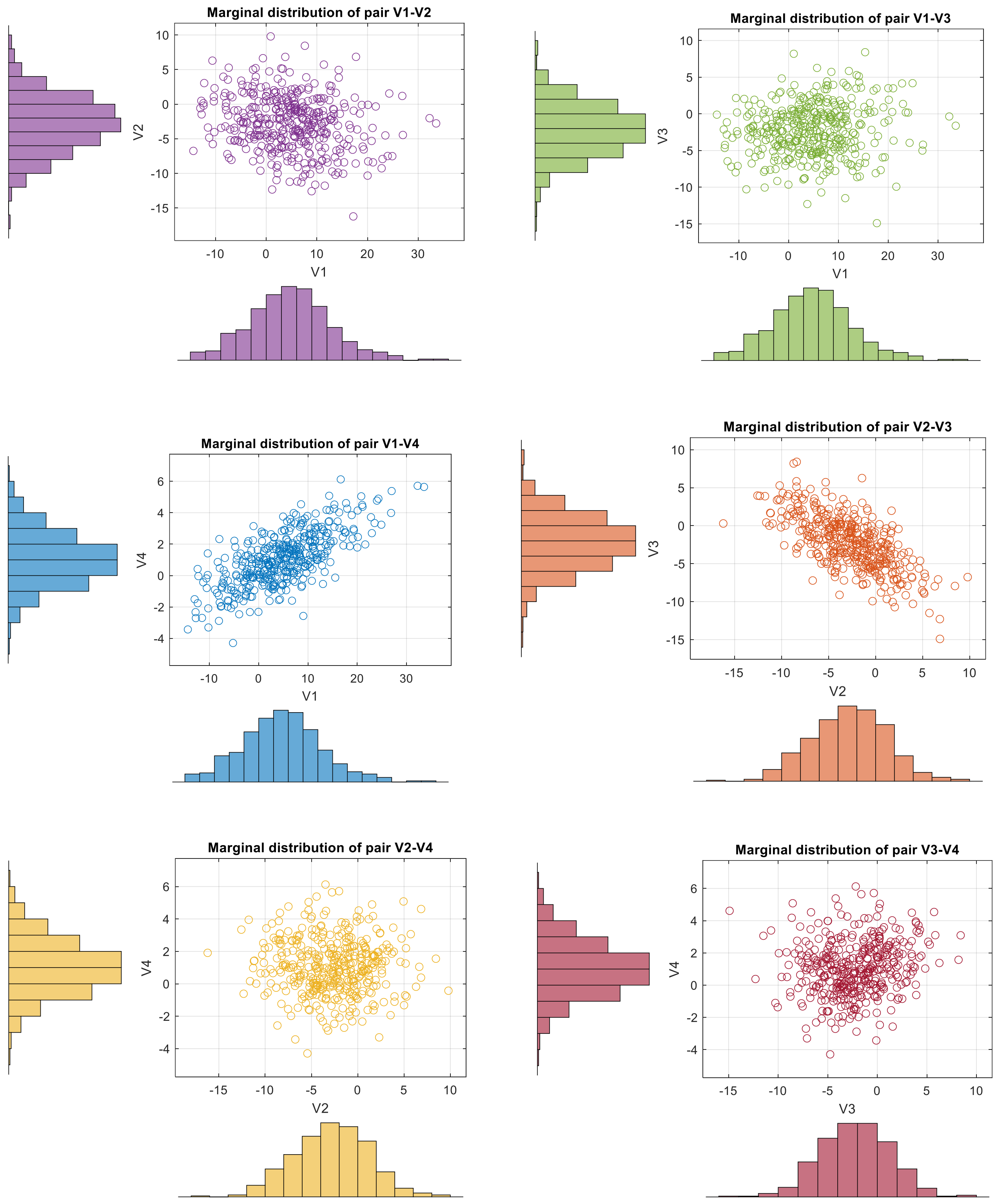


Figure 3.3.1 – *Marginal Distribution of each unique pair of bio sensor channel* as scatterhist plot

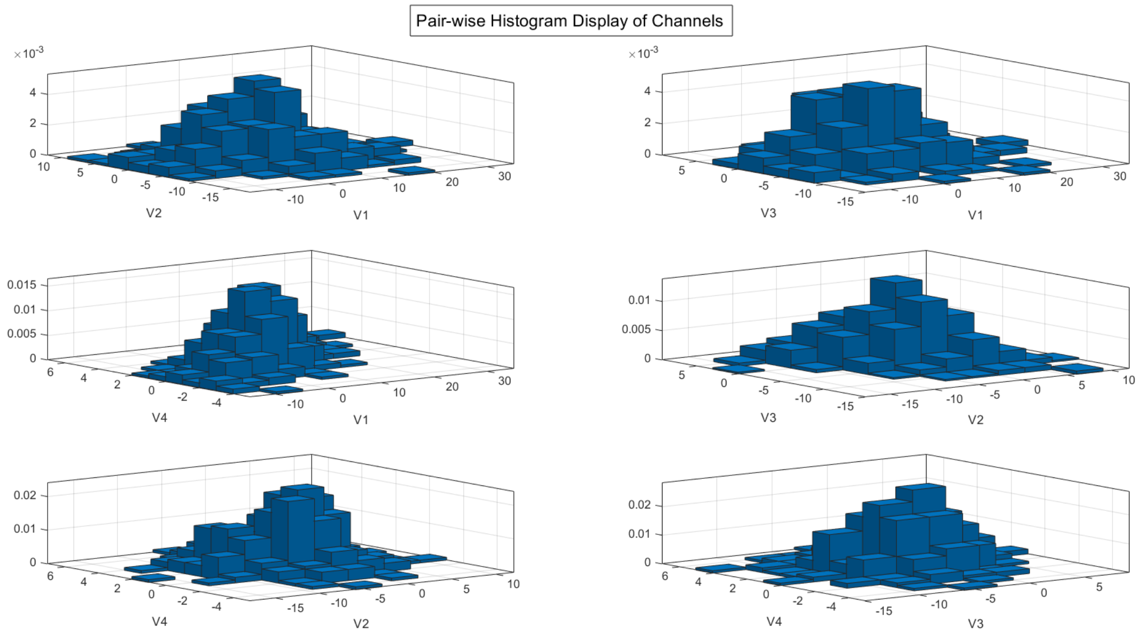
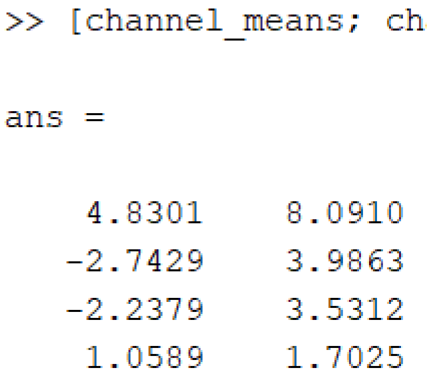


Figure 3.3.2 – *Marginal Distribution of each unique pair of biosensor channel* as histogram2 plot

Amongst the 4 channels, Channel V3 is the most probable predicted biosensor channel. As, V3 shows highest peak when paired with all the biosensor channels.

iv) In this task, we find out the empirical mean and standard deviation of each channel. Which is in the table generated as follows. Also, over further variation the mean deviation is not much affected.



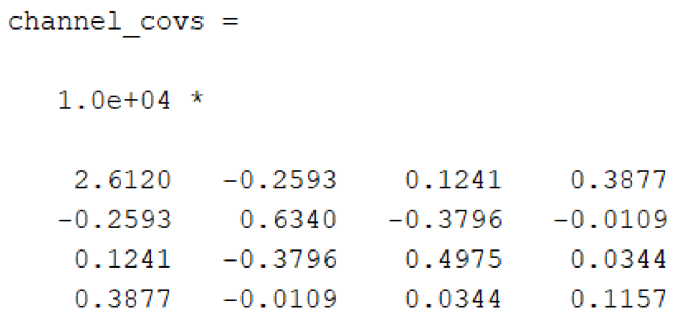
```
>> [channel_means; channel_stds]'

ans =

    4.8301    8.0910
   -2.7429    3.9863
   -2.2379    3.5312
    1.0589    1.7025
```

Figure 3.4 – *[Mean; Standard Deviations]* of each channel respectively

v) In this task, we can see covariance and associated correlation coefficient are maximum when the biosensor channel is paired with itself.

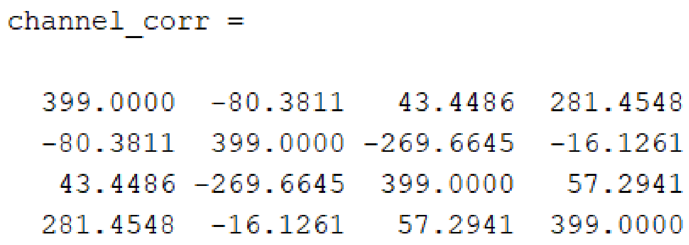


```
channel_covs =

   1.0e+04 *

    2.6120   -0.2593    0.1241    0.3877
   -0.2593    0.6340   -0.3796   -0.0109
    0.1241   -0.3796    0.4975    0.0344
    0.3877   -0.0109    0.0344    0.1157
```

Figure 3.5.1 – Pair-wise *Channel Covariance* matrix



```
channel_corr =

  399.0000  -80.3811   43.4486  281.4548
  -80.3811  399.0000 -269.6645  -16.1261
   43.4486 -269.6645  399.0000   57.2941
  281.4548  -16.1261   57.2941  399.0000
```

Figure 3.5.2 – Pair-wise *Channel Correlation Coefficient* matrix

vi) In this task, Normal marginal distribution of V1-V2 Pair is projected with scatter plot over contour of

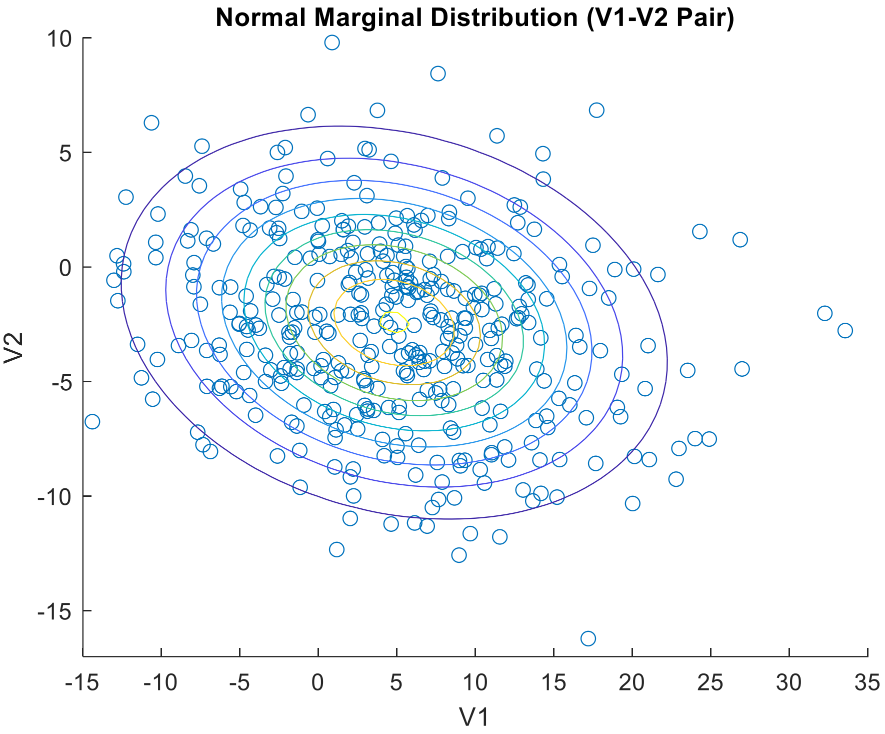


Figure 3.6.1 – Pair-wise *V1-V2 Channel* Normal Marginal Distribution **scatter plot** over **contour**

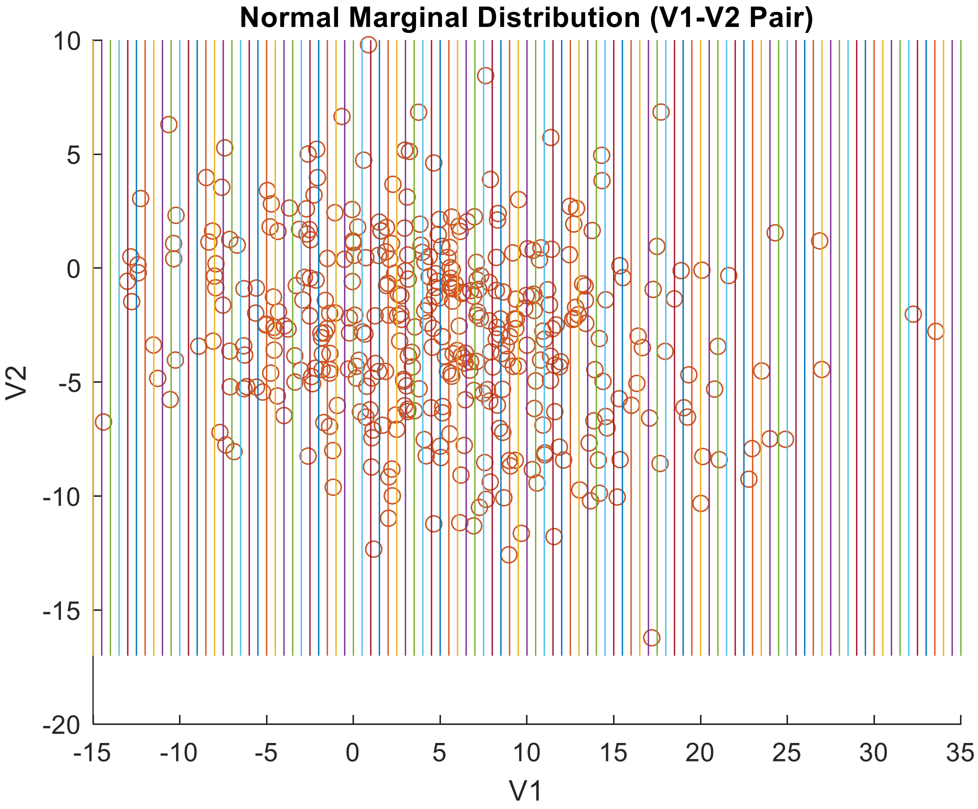


Figure 3.6.2 – Pair-wise *V1-V2 Channel* Normal Marginal Distribution **scatter plot** over **plot**

vii) In this task, for pair of V2-V3 biosensors, scatter plot over regression line in between of lower and upper standard deviation

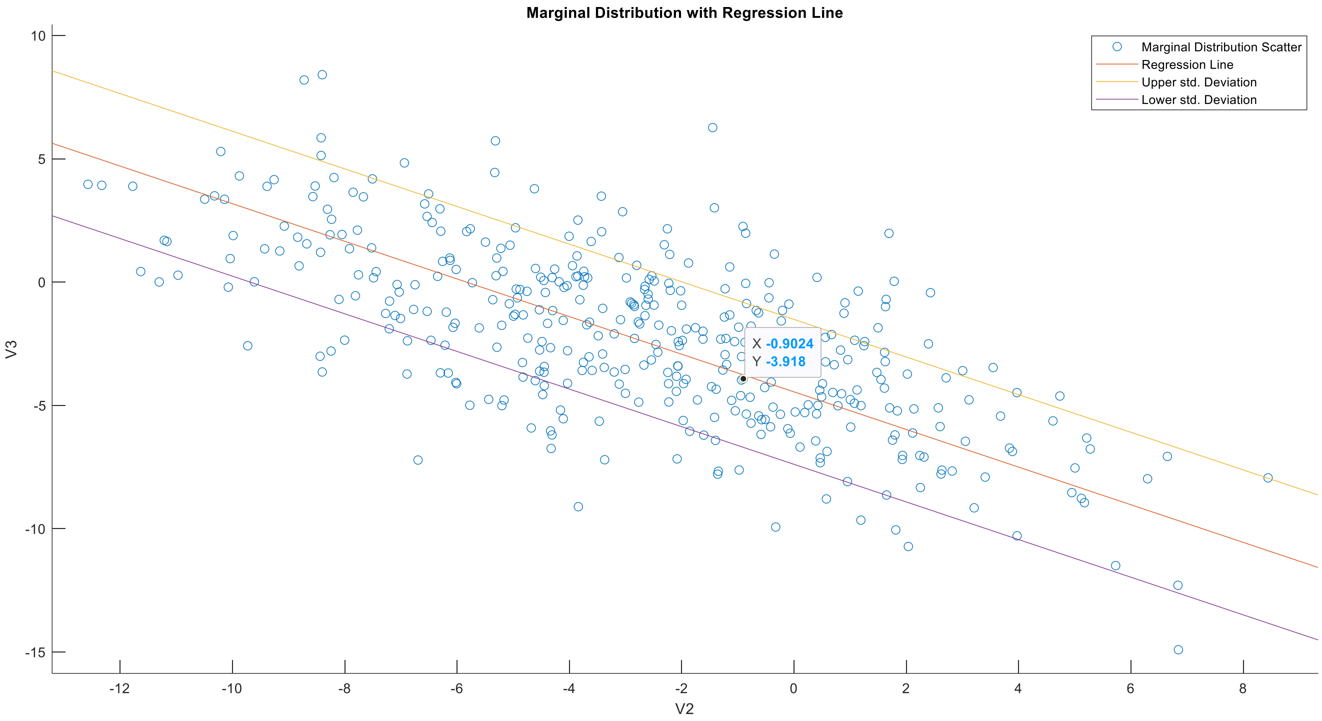


Figure 3.7 – Pair-wise *V2-V3 Channel* Marginal Distribution plot

The highlighted point **V2 ≈ -0.9mV**, we get V3 ≈ -3.92 mV i.e. **V3 < -3.8** for the given value of V2.

viii) In this task, we plotted Bivariate Marginal Distribution on the scatter plot of V2-V4 biosensor channel pair. **Y given X** and **X given Y** regression line differs in slope. Hence, for this pair the scatter values differs in multiple of the difference of slope.

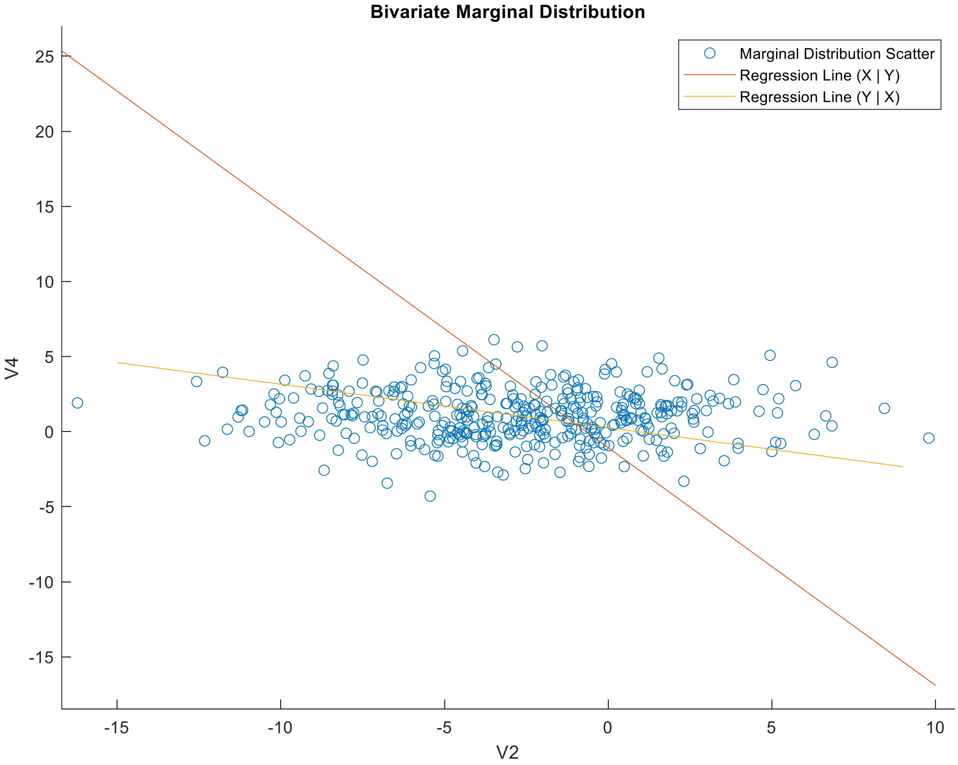


Figure 3.8 – Pair-wise *V2-V4 Channel* Bivariate Marginal Distribution without **contour**

References:

1. MATLAB Documentation: https://uk.mathworks.com/help/index.html
2. Probability Modelling Course: http://www.mee.tcd.ie/~aquinn/3e3/
3. https://en.wikipedia.org/wiki/Probability_density_function