Gaussian Approximation Using Integer Sequences

Arulalan Rajan, Ashok Rao, R. Vittal Rao, and H.S. Jamadagni

Abstract. The need for generating samples that approximate statistical distributions within reasonable error limits and with less computational cost, necessitates the search for alternatives. In this work, we focus on the approximation of Gaussian distribution using the convolution of integer sequences. The results show that we can approximate Gaussian profile within 1% error. Though Bessel function based discrete kernels have been proposed earlier, they involve computations on real numbers and hence increasing the computational complexity. However, the integer sequence based Gaussian approximation, discussed in this paper, offer a low cost alternative to the one using Bessel functions.

1 Introduction

Lindberg, in his work, [1], presents a family of kernels, which are the discrete analogue of the Gaussian family of kernels. The discrete Gaussian kernel in [1] uses modified Bessel function of integer order. It is well known that Bessel function evaluation is a computationally demanding process. This necessitates the need for

Arulalan Rajan
Dept. of Electronics & Communication Engg,
National Institute of Technology Karnataka, Surathkal
e-mail: perarulalan@gmail.com

Ashok Rao
Consultant
e-mail: ashokrao.mys@gmail.com

R. Vittal Rao · H.S. Jamadagni
Dept. of Electronics System Engg,
Indian Institue of Science, Bangalore
e-mail: {rvrao,hsjam}@cedt.iisc.ernet.in

S.M. Thampi, A. Gelbukh, and J. Mukhopadhyay (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 264,
DOI: 10.1007/978-3-319-04960-1_19,

looking at computationally less-demanding alternatives to approximate Gaussian profiles. Interestingly, while exploring the use of integer sequences [2] for generating window functions for digital signal processing applications [3], [4], we found that the convolution of symmetrised integer sequences resulted in a Gaussian like profile. This made it worth to explore the degree, to which, a single convolution of two symmetrised integer sequences, approximate a Gaussian profile. It is well known that the computations on integers and integer sequences are less demanding in terms of power and complexity. The aim of this paper is to throw light on the use of non-decreasing integer sequences to approximate discrete Gaussian profile. In this regard, we provide two techniques, based on convolution of integer sequences, to approximate Gaussian distribution, as listed below:

- Symmetrising the integer sequences, followed by their linear convolution.
- Linear convolution of non decreasing integer sequences and symmetrising the resulting sequence about its maximum.

These techniques result in mean squared error, between the estimated probability density function and the obtained density function, of the order of 10^{-8} or equivalently about 1% error. We use some of the sequences listed in the Online Encyclopedia of Integer Sequences [2]. The following notations are used throughout this paper:

- N : Sequence Length
- $x_i[n], x_j[n]$: Sequences used in linear convolution, of length L and M respectively
- $y[n]$: Sequence resulting from the linear convolution of two integer sequences, given by

$$y[n] = \sum_{k=0}^{L-1} x_i[k]x_j[n-k] \qquad n = 0,1,2,\ldots,L+M-1 \tag{1}$$

- ${X_{sc}}^1$, ${X_{cs}}^2$: Random variable that can assume values from the set $\{1,2,3,\ldots,N\}$, such that, the probability

$$P(X=n) = \frac{y[n]}{\sum_{k=1}^{N} y[k]} \qquad \forall n = 1,2,\ldots,N \tag{2}$$

In this paper, we restrict ourselves to integers, for the well known reason that, the computations on integers are much less complex than those on fixed point or floating point numbers. However, the comparison of complexity issues with regard to non-integer but fixed point methods with integers is beyond the scope of this paper.

[1] $(.)_{sc}$: Non decreasing input sequences are first symmetrized and then convolved
[2] $(.)_{cs}$: Non decreasing input sequences are first convolved and then symmetrized

2 Approximating Discrete Gaussian by Symmetrising the Integer Sequences, Followed by Convolution

2.1 Convolution of Symmetrized Arithmetic Progression Sequence

Consider, the arithmetic progression sequence $p[n] = a + nd$, where a, d, n are all integers. This sequence is labeled as A000027 in [2]. For a finite length, N,of the sequence, we symmetrize the sequence at $N/2$ or $(N+1)/2$, depending on whether the length is even or odd. Let us denote this symmetrized sequence by $x_s[n]$. This sequence is then convolved with itself to obtain $y[n]$. Let X_N be the random variable that can assume one of the values in the set $\{1,2,3,\ldots,2N-1\}$. The probability that X_N can take a specific value, n, is given by

$$P(X_{sc} = n) = \frac{y[n]}{\sum_{k=1}^{2N-1} y[k]}, \qquad \text{for all } n = 1, 2, 3, \ldots, 2N\text{-}1 \tag{3}$$

The plot given in fig.1 illustrates that one can indeed approximate Gaussian distribution by the sequence obtained by convolution of symmetrized arithmetic progression sequence, with a mean squared error of the order of 10^{-8}. The obtained density function, in fig.1 is same as the convolution profile. For the profile obtained, we compute the mean, μ, and variance, σ^2. For this mean and variance, we then fit a Gaussian distribution, using the MATLAB inbuilt function *normpdf*. This corresponds to the estimated density function mentioned in fig.1.

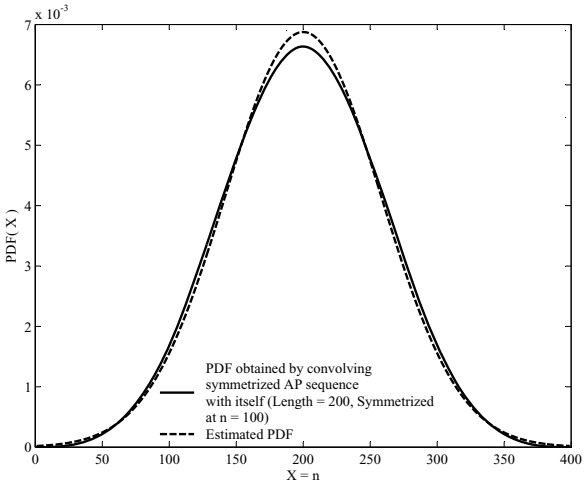

Fig. 1 Comparison between the obtained and estimated PDF

2.2 *Convolution of n^2 and Higher Powers of n*

We now look at integer sequences generated by higher powers of n, say n^2, n^3 etc. Fig. 2 gives the profiles obtained by convolving symmetrized integer sequences generated by various powers of n. The distributed square error, is given in fig. 3. We find that, the sequence n^2, after symmetrization and convolution, approximates discrete Gaussian with an error of the order of 10^{-9} or even less. The sequence, n^3, does a poor approximation of discrete Gaussian. Thus, it appears that there is an optimal value of the exponent, k, between 2 and 3. However, for $2 < k < 3$, the sequence n^k ceases to be an integer sequence.

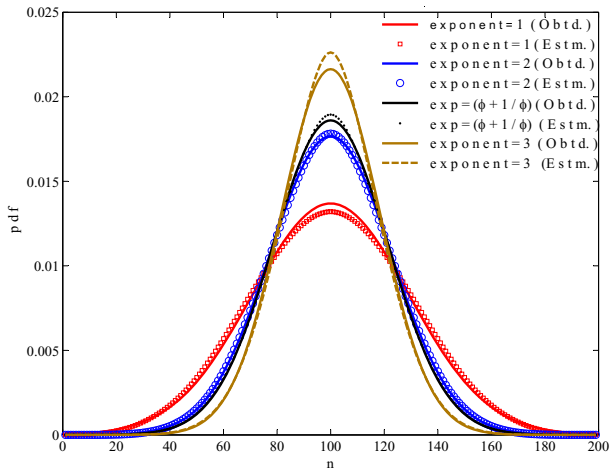

Fig. 2 Approximating Discrete Gaussian by sequences of the form n^k

Also, from fig. 4, we observe that, for higher powers of n, the variance saturates. Moreover, the lowest exponent with which we can obtain an integer sequence in the interval $(2,3)$ is 2. Thus it appears that, it may not be possible to obtain discrete Gaussian with different variance values, with integer sequences generated by n^k, for $k > 2$. We conjecture that the highest power k for which we can use n^k to approximate discrete Gaussian is related to the golden ratio, $\varphi = \frac{1+\sqrt{5}}{2}$. However, this sequence will not be an integer sequence. Therefore, it is necessary to look at other integer sequences which are slow growing and are referred to as the metafibonacci sequences [5]. In the subsections to follow, we look at the convolution of other symmetrized integer sequences. The sequences investigated include Golomb sequence [2], A005229 sequence [2], apart from those discussed in the paper. We present the simulation results for some of the integer sequences in the following subsections.

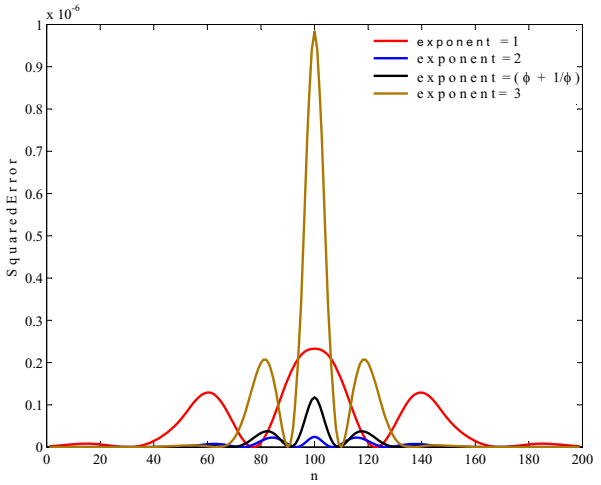

Fig. 3 Error Distribution in Approximating Discrete Gaussian by sequences of the form n^k

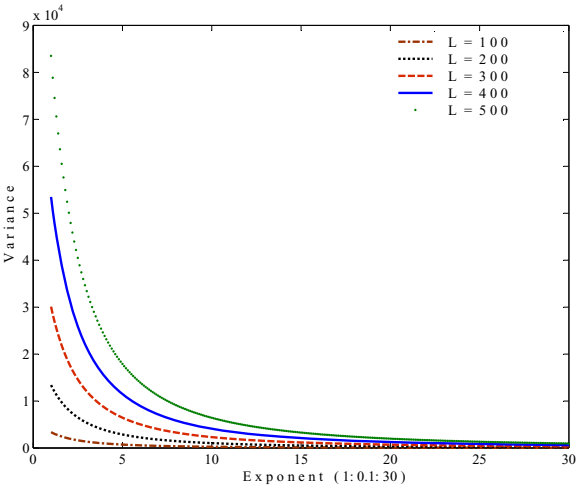

Fig. 4 Variance Saturates for higher values of k

2.3 *Convolution of Symmetrized Hofstadter-Conway (HC) Sequence*

The next sequence that we look at is the Hofstadter-Conway sequence [2], generated by the recurrence relation,

$$a[n] = a[a[n-1]] + a[n - a[n-1]] \qquad a[1] = a[2] = 1; \tag{4}$$

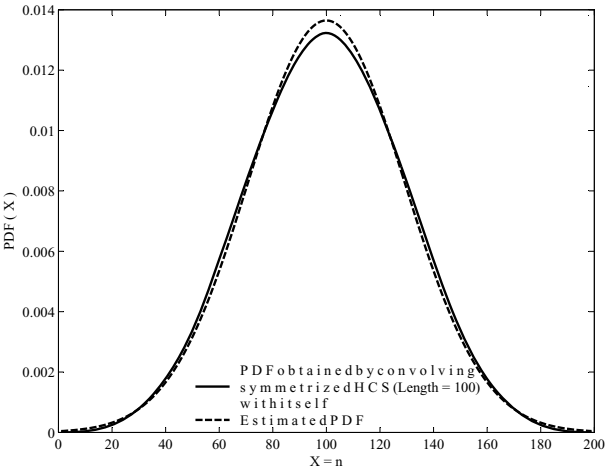

Fig. 5 Obtained PDF vs Estimated PDF

Fig.5 gives the profile obtained by the convolution of symmetrized Hofstadter Conway sequence with itself. A plot of the estimated density function is given in fig.5. This shows that a Gaussian density function can be approximated very closely by convolving symmetrized HC sequence with itself.

2.4 Convolving Hofstadter-Conway Sequence and Sequence A006158

We also found that convolving two different symmetric integer sequences also resulted in a profile similar to Gaussian.

Now, we look at the convolution of symmetrized HC sequence and another sequence labelled as A006158, in the OEIS [2], generated by the recurrence relation,

$$a[n] = a[a[n-3]] + a[n-a[n-3]] \qquad a[1] = a[2] = 1; \tag{5}$$

Both the sequences were considered to be equal in length and the length was taken to be 100. From fig.5, we find that Gaussian distribution with a desired variance can be obtained by varying the length of the sequences to be convolved. Interestingly, from fig.6, we can infer that, to approximate a Gaussian distribution with a desired variance, using integer sequences, and with minimal mean squared error, there are two options, namely,

- varying the lengths, N_1 and N_2, of the two symmetric sequences
- choice of the sequences.

The plots indicate that one can closely approximate a Gaussian distribution of a specific variance by convolving symmetrized integer sequences. Fig.7 compares the

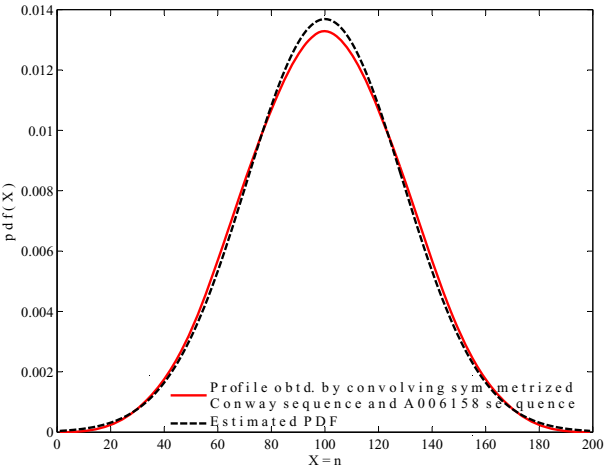

Fig. 6 Obtained PDF vs Estimated PDF

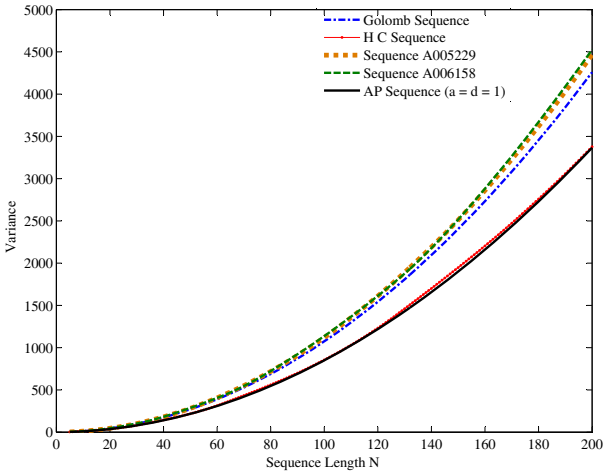

Fig. 7 Variance versus Sequence Length for Different Sequences

manner in which the variance, of distributions generated by convolving symmetrized integer sequences, depends on the length of the sequence.

From fig. 7 it is clear that we can obtain the length of the sequence required for any particular variance of the Gaussian PDF. However, this automatically fixes the mean. At present, it appears that only one of the two parameters - variance or mean, can be realised by the choice of the lengths of the sequences.

3 Approximating Gaussian Distribution by Convolving Integer Sequences and Symmetrizing

In this approach, we convolve two finite length, non-decreasing sequences. Once the convolution is done, we symmetrize the result at that N^*, where the convolution result has its absolute maximum. The resulting sequence is used as the probability density function of the discrete random variable, X.

3.1 Integer Sequence in Arithmetic Progression (AP)

Consider an arithmetic progression $p[n] = a + nd$, where a, d, n are all integers. We then define a truncated sequence $x[n]$ with N terms as

$$x[n] = p[n](u[n] - u[n-N]) \tag{6}$$

where $u[n]$ is the unit step function. Consider the convolution of this truncated sequence,(of length N) with itself. This results in a sequence, $y[n]$, of length $2N-1$ and is defined by

$$y[n] = \sum_{k=0}^{N-1} x[k]x[n-k] \tag{7}$$

$$y[n] = \sum_{k=0}^{N-1} p[k]p[n-k](u[n-k] - u[n-k-N]) \tag{8}$$

Clearly $y[n] = 0$ for $n < 0$ and is defined differently for different regions namely $0 \le n \le N-1$ and $N \le n \le 2N-1$. Thus

$$y[n] = \begin{cases} 0 & n < 0 \\ \sum_{k=0}^{n} p[k]p[n-k] & 0 \le n \le N-1 \\ \sum_{k=n-N+1}^{N-1} p[k]p[n-k] & N \le n \le 2N-1 \end{cases} \tag{9}$$

Evaluating the summations we get

$$y[n] = \begin{cases} 0 & n < 0 \\ a^2(n+1) + ad(n(n+1)) + d^2\left[\dfrac{n(n^2-1)}{6}\right] & 0 \le n \le N-1 \\ a^2(2N-n-1) + ad[n(2N-n-1)] + d^2 t_1 & N \le n \le 2N-1 \end{cases} \tag{10}$$

where

$$t_1 = \left[\frac{N(N-1)(3n-2N+1) - (n-N)(n-N+1)(n+2N-1)}{6} \right] \tag{11}$$

The closed form of convolution of arithmetic progression is given in eq.10. The convolution plot is given in fig.8

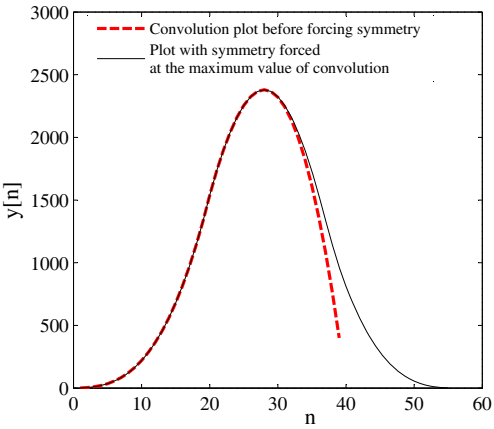

Fig. 8 Convolution of AP Sequences and Reflection about the index of maximum value

From fig.8, we find that, to get a Gaussian-like profile, we need to truncate the convolved sequence at the point of its absolute maximum and employ symmetry. To do so, it is necessary to obtain the point at which the absolute maximum occurs. Let us denote this point as N^*, the point at which, $y[n]$, defined by eq.10, attains its absolute maximum. It can be shown that the maxima is obtained at

$$N^* \approx \frac{(2N-1)}{\sqrt{2}} - \frac{(2-\sqrt{2})a}{d} \tag{12}$$

Thus we get, from eq.12, the point at which the result of convolution of an AP sequence with itself has its maximum value. At this N^* we symmetrize the result of convolution. This results in a sequence, $y_1[n]$, which closely approximates the Gaussian curve. The length of $y_1[n]$ is $2N^*$. In this case, where we convolve first and then employ symmetry about the point of absolute maximum, the values that the random number X can assume is from the set $\{1,2,3,\ldots,N^*,\ldots,2N^*\}$. Hence, the probability that the random variable X can take is defined by

$$P(X_{cs} = n) = \frac{y_1[n]}{\sum_{k=1}^{2N^*} y_1[k]} \tag{13}$$

where $y_1[n]$ is the sequence obtained by the convolution of the AP sequence and employing symmetry at N^*. The following were the values taken for the AP sequence: $a = 1$; $d = 1$; $N = 100$.

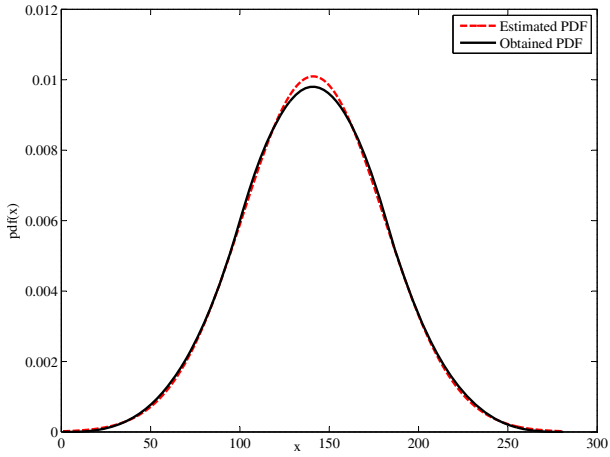

Fig. 9 Comparison between the obtained and estimated PDF

Fig.9 compares the obtained density function with the estimated one. Further investigations show that convolution of various other metafibonacci sequences indeed result in very close approximation of a Gaussian profile.

3.2 Validation Using Fourier Fit

To validate that the two techniques, presented in the previous sections, a four term Fourier fit was done. This involved the following steps:

- ***Step 1:*** The integer sequences were symmetrized and then convolved or vice-versa.
- ***Step 2:*** The mean and variance of the resulting sequence were obtained.
- ***Step 3:*** For that mean and variance, a discrete Gaussian distribution was obtained. The mean squared error was found out.
- ***Step 4:*** Assuming that it is a continuous distribution, a four term Fourier fit was obtained at random points. The number of points were the same as the length of the sequence resulting from step 1. However, the sample points were randomly chosen.
- ***Step 5:*** For this distribution, the mean and variance were obtained.
- ***Step 6:*** With this as the mean and variance, a Gaussian PDF was estimated.
- ***Step 7:*** The mean square error was obtained as the absolute difference between the profile obtained in step 4 and step 6.

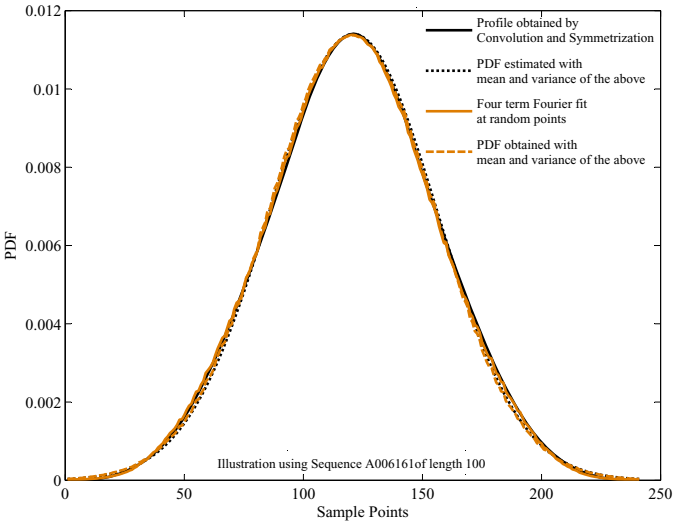

Fig. 10 Fourier Fit

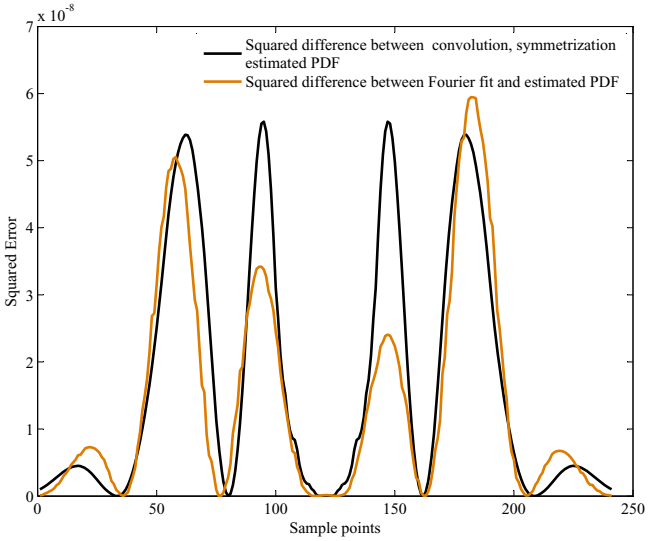

Fig. 11 Squared Error Plot

The mean square error was found to be of the order of 10^{-8}. This clearly shows that the sequence obtained by convolution of integer sequences or the sequence obtained by convolution of integer sequences and symmetrization, indeed, approximated discrete Gaussian distribution with 1% error. Fig.10 illustrates the same,

using sequence A006161, of length 100, convolved and symmetrized, while fig.11 compares the error distribution. The Fourier series coefficients, in this case, are:

$$a_0 = 0.004277; a_1 = -0.005424; a_2 = 0.001288; a_3 = -0.0002101; a_4 = 9.588 \times 10^{-5};$$
$$b_1 = -0.0006025; b_2 = 0.000289; b_3 = -7.239 \times 10^{-5}; b_4 = 4.543 \times 10^{-5}; w = 0.02688;$$

It can be seen from fig.10 that the sampling points are different from the ones used in step 2 and step 3.

4 Conclusion

In this paper, we proposed two techniques to approximate discrete Gaussian distribution with integer sequences. We found that convolution of slow-growing sequences can approximate a 4 term cosine fit within 1% error, which in turn approximates a Gaussian distribution with very low approximation error of about 1%. We also found that discrete Gaussian distribution with a specific variance and mean can be controlled by varying the length of the integer sequences or by choosing the sequences or both. Also, as operations on integers are known to be less demanding in terms of computations, we find that the integer sequences can indeed be used as alternatives to approximate probability distributions.

References

1. Lindberg, T.: Scale-Space for Discrete Signals. IEEE Trans. on Pattern Analysis and Machine Intelligence 12(3), 234–254 (1990)
2. Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences, Ed.2008 published electronically at, `http://oeis.org/SequenceNumber`
3. Antoniou, A.: Digital Filters: Analysis, Design and Applications, 2nd edn. Tata McGraw-Hill (1999)
4. Rajan, A.: Some applications of Integer sequences in signal processing and their implications on performance and architecture, PhD thesis, Indian Institute of Science (September 2011)
5. Vajda, S.: Fibonacci and Lucas Numbers, and the Golden Section: Theory and Applications, Dover (2008)