B-SPLINE PDF: A GENERALIZATION OF HISTOGRAMS TO CONTINUOUS DENSITY MODELS FOR GENERATIVE AUDIO NETWORKS.

Yannis Agiomyrgiannakis

Speech Group, Google Inc.

agios@google.com

ABSTRACT

Many modern neural networks use histograms to efficiently model continuous random variables. This implies that the parametric space of the multinomial distribution is easier for training large neural networks. In applications like generative audio networks, this approach introduces audible quantization noise to the generated signal. This work presents a novel probability density function (PDF), referred to as B-Spline PDF, that is a direct generalization of histograms to continuous densities while retaining the multinomial parameter space. The latter uses k-th order B-Splines to ensure continuity up to the $(k-1)$-th order derivative. B-Spline PDF is amenable for neural network training via closed-form gradients that are easy and fast to compute. For other applications, one may use a novel algorithm, referred to as the *Expectation* algorithm, to efficiently estimate the model parameters. Further, a novel sample generation algorithm is derived that is fast and simple. The theoretical results, coupled with illustrative examples, suggest that B-Spline PDF may directly replace histograms in many related applications.

Index Terms: histograms, B-splines, non-parametric statistics, deep neural networks, generative models, waveform synthesis

1. INTRODUCTION

Recent advances in generative models allow speech and audio signal synthesis with high quality. Networks such as SampleRNN [1] are conditional probability models of the current sample given a number of previous samples conditioned on some application-specific side-information derived from text. A key characteristic of these probability models is that they are discrete rather than continuous as one would expect for audio signals. They model the conditional probability mass function of the indices of the quantized audio signal. The indices are constructed using companding (μ-law or A-law) followed by uniform quantization.

The advantage of treating a continuous signal as discrete via quantization is that the model has no assumptions regarding the global structure of the underlying probability density function (PDF). However, the overall process of quantization followed by discrete modelling corresponds to a histogram [2], which is a rather inefficient model for continuous variables. As a result, audio generated via SampleRNN using this model contains audible quantization noise.

The alternative is to construct a mixture of PDF kernels such as Gaussian, Radial Basis Functions (RBFs), or Logistics. Theoretically, these mixtures are universal approximators of the underlying PDF, but practically they do not scale well as the number of model parameters increases. Training them is much harder than training histograms, probably due to the more complicated structure of their parameter space. As a result, especially in neural network training that involves tuning multiple hyper-parameters, researchers tend to use either mixture models with few components or histograms [1].

Since histograms are mixture models of rectangular functions, both paradigms can be seeing as edge-cases of mixture models. Histograms are parameterized *locally* in the sense that each sample affects a portion of the overall PDF support, while kernel mixtures are parameterized *globally* in the sense that each sample affects the whole PDF.

The rigidity of the latter models leads to performance degradations when the underlying statistics deviate from the model triggering an increased interest in non-parametric statistics such as kernel estimators [3, 4]. Unfortunately, the computational cost for kernel estimators is too high for big-data and neural network training as they require the computation of a kernel function for each sample. Workarounds like Radial Basis Function (RBF) networks involve a two-step procedure that reverts them back to mixture models: first quantize the source and then center the kernels at the codepoints [4].

As a consequence, modern neural network design is constrained to use either discontinuous histograms or rigid parametric models because non-parametric models are not scalable to modern big-data sizes unless they are re-parameterized, which re-introduces the undesirable rigidity. From the perspective of generative audio networks, histograms based on softmax distributions tend to work better in practice than mixture densities [1].

Plausible, but less popular, alternatives blend between parametric and non-parametric models and are usually referred to as *functional* models; for example *log-spline* models and *orthogonal series* models. Log-spline models use polynomial splines or B-splines to model the logarithm of the PDF using variably-spaced knots [5, 6]. Orthogonal series models use an orthogonal transform (e.g. Fourier) to model the PDF [7, 3]. Both alternatives are quite flexible, more suitable than histograms to model continuous PDFs but not necessarily scalable to big-data as the lack of corresponding references implies.

The aim of this paper is to provide a continuous PDF model that is as versatile as the quantized ones, namely, the histogram. This work expresses histograms as 0-th degree B-splines in order to extend them to k-degree B-splines that are continuous up to the $(k-1)$-th derivative. The resulting model is *a generalization of histograms* that will hereafter be referred to as B-Spline PDF. It corresponds to successive smoothing of a uniform histogram using k convolutions with a rectangular function. Unlike log-spline models [5, 6], it models the PDF and not the logarithm of the PDF and uses uniformly distributed knots instead of non-uniformly distributed knots. Unlike the functional models in [3, 5, 6, 7], the parameters of the model can be multinomial distributions themselves, which renders the B-Spline PDF model suitable to be a direct replacement of histograms in many applications that had to quantize continuous variables in order to model them, such as the softmax-based histograms used in

generative neural networks [1]. This would eliminate the audible quantization noise in these models.

Section 2 presents published results on companded quantization using High-Rate theory [8] and some experimental validation of optimal companders. Section 3 demonstrates that companded quantization with multinomial modelling corresponds to a histogram. Section 4 generalizes histograms to B-Spline PDFs. Section 5 presents the gradient of the log-likelihood that can be used for neural-network training and proposes the *Expectation* algorithm, a simple iterative algorithm to efficiently estimate the model parameters that belongs to the family of Expectation-Maximization algorithms. Section 6 suggests an random generator algorithm for all B-Spline PDFs. Finally, section 7 concludes the paper.

2. COMPANDED QUANTIZATION

Generative models [1] quantize the 16-bit PCM input audio samples at 8-bits using companded quantization [8] following the A-law ITU standard [9]. The standard suggests two closely related companding functions, A-law and μ-law, that differ largely due to historical reasons.

The ITU-standard companders A-law and μ-law are effective by permitting fast implementations but they are not Rate-Distortion optimal [8], meaning that they cannot take advantage of knowledge regarding source statistics. Optimal companders for squared error distortion metrics can be derived using Bennet's High-Rate assumptions [8].

Let $s \sim f(s)$ be the audio input signal following the PDF $f(\cdot)$. Let the P-th power error $D_p(s, \hat{s}) = |s-\hat{s}|^P$ be the distortion metric that we want to minimize. Using High-Rate theory assumptions [8], we can derive the optimal codepoint density $\lambda(s)$ of a uniform quantizer to be:

$$\lambda(s) = \frac{f(s)^{1/(P+1)}}{\int f(s')^{1/(P+1)} ds'}, \tag{1}$$

which normalizes $f^{1/(P+1)}$ to integrate to one. Since $1/(P+1) \leq 1.0$, formula 1 has a smoothing effect on $f(s)$ that raises the probability in low-probability areas and decreases the probability in high-probability areas. This re-arrangement reduces the size of the biggest quantization cells as P increases because they yield higher distortions. Thus, the optimal compander $x = g(s)$ that converts $s \sim f(s)$ to $x \sim \lambda(s)$ is the cumulative codepoint density function of $x \sim \lambda(s)$ [8].

Figure 1 demonstrates the Rate-Distortion curves of several companders using squared-error distortion. The $P = 1, 2, 3$ companders are derived to be optimal for the statistics of s under distortion $D_p(\cdot, \cdot)$. As expected, the best compander overall is the $P = 2$ one because it corresponds to the evaluation distortion metric. The interesting part is that the best compander gains as much over the ITU companders as the later ones gain over no companding (linear compander). Further, the measured distortion is almost the same as the one predicted theoretically using High-Rate theory [8]. The experiment was made using the same EN-US single female speaker corpus as in [10].

Companding is a necessary pre-processing step when we need to model distributions with infinite support. We will henceforth assume that the input audio source is the companded audio signal $x = g(s)$ and not the original s.

Generalizing, companding allows us to handle infinite support distributions using finite-support ones and provides some justification on the use of uniformly distributed codepoints for being optimally distributed in the P-th power sense according to the global statistics of our source.

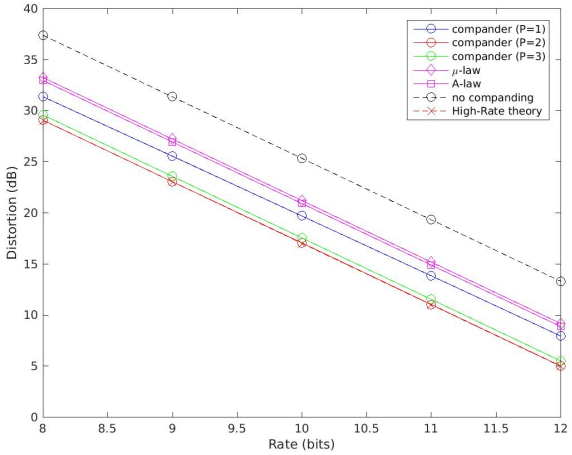

Fig. 1: Rate-Distortion for several companding functions using squared-error distortion metric.

3. THE SOFTMAX-BASED HISTOGRAM

Let $x_m^Q, m = 1, ..., M$ be the M uniformly distributed quantization points (codepoints) of x. Let $p(m)$ be the probability of having the m-th codepoint, following a multinomial distribution fed from a softmax-based layer of a neural network. The PDF that is used elsewere [1] to model the statistics of x can be described by the equation

$$p_0(x) = \sum_{m=1}^{M} p(m) K_0(\frac{x - x_m^Q}{\Delta}), \tag{2}$$

where Δ is the sampling interval of the uniformly spaced codepoints and $K_0(\cdot)$ is the rectangular function

$$K_0(x) = \begin{cases} 1 & |x| < 0.5 \\ 0.5 & |x| = 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Equation 2 corresponds to modelling a continuous source using a codebook and is essentially a histogram. From a statistical perspective, it would make sense to use the same PDF during synthesis, but all papers preferred to use the quantized samples x_m^Q according to $p(m)$ rather than generating x from $p_0(x)$ because equation 2 doubles the amount of additive noise in the generated signal. Thus, they used the following generative PDF:

$$p_{0,gen}(x) = \sum_{m=1}^{M} p(m) \delta(x - x_m^Q), \tag{4}$$

where $\delta(\cdot)$ is the Dirac function.

4. GENERALIZING HISTOGRAM AS B-SPLINE PDF

This paper proposes a generalization of the histogram PDF by successively smoothing it via convolutions with the rectangular function $K_0(\cdot)$. The smoothed histogram is of the form:

$$p_k(x) = \sum_{m=1}^{M} p(m) K'_{k,m}(x), \tag{5}$$

where $K'_{k,m}$ is the normalized kernel:

$$K'_{k,m}(x) = \frac{K_k(\frac{x-x_m^Q}{\Delta})}{\int_{x_{min}}^{x_{max}} K_k(\frac{x'-x_m^Q}{\Delta})dx'}. \quad (6)$$

The normalized kernel is based on the k-th order kernel $K_k(x)$, defined as k convolutions of the rectangular function with itself:

$$K_k(x) = \circledast_{k'=0}^{k} K_0(x), \quad (7)$$

where $\circledast$ is the convolution operator. By definition, the normalized kernel $K'_{k,m}(x)$ is a valid PDF that integrates to unity within the support of the random variable x, while the kernel $K_k(x)$ integrates to unity within $[(k-0.5)\Delta, (k+0.5)\Delta]$.

Let the support of x be $[x_{min}, x_{max}]$. According to section 2, for histograms ($k = 0$) a plausible placement of the quantization codepoints (spline knots) is the one that minimizes the mean-squared-error (MSE): $x_m^Q = x_{min} + \Delta(m - 0.5)$, where $\Delta = \frac{x_{max}-x_{min}}{M-1}$. For higher order B-Splines ($k > 0$) a plausible strategy when there is no prior knowledge of the PDF support would be to place the first and the last codepoint at x_{min} and x_{max}, respectively, in order to ensure that every point in the interval is described by exactly $k + 1$ kernels. This strategy suggests the same level of flexibility for the PDF across the whole support, but can be relaxed at the benefit of gaining 2 extra histogram bins. Thus, the codepoints can be set uniformly to: $x_m^Q = x_{min} + \Delta(m - 1)$, where $\Delta = \frac{x_{max}-x_{min}}{M-1}$. This places the first codepoint at $x_1^Q = x_{min}$ and the last at $x_M^Q = x_{max}$.

A closer look at equation 5 reveals that it corresponds to the basic splines (B-splines) first introduced by Schoenberg [11, 12] in 1946. However, they are not completely unconstrained since their coefficients are non-negative probabilities $p(m)$, altogether forming a convex combination.

The 0-th order B-Spline PDF is the well-known histogram, a discontinuous function. The 1-st order B-Spline PDF is the continuous piecewise linear function, with discontinuous derivatives. The 2-nd order B-Spline PDF is a piecewise quadratic function with continuity up to the 1-st derivative. Extending, the k-th order B-Spline PDF is a piecewise quadratic function with continuity up to the $(k-1)$-th derivative.

5. ESTIMATION

Estimation of the B-Spline PDF parameters can be made by optimizing the average log-likelihood:

$$L = \frac{1}{N}\sum_{n=1}^{N} \log p_k(x_n) \quad (8)$$

where $x_n, n = 1, ..., N$ are the N data samples. In the context of a neural network training, we need the partial derivatives of equation 8 with respect to the parameters $p(m)$:

$$\frac{\partial L}{\partial p(m)} = \frac{1}{N}\sum_{n=1}^{N} \frac{K'_{k,m}(x_n)}{\sum_{m=1}^{M} p(m)K'_{k,m}(x_n)}. \quad (9)$$

Figure 2 demonstrates the result of fitting B-Spline PDFs of several orders to 10,000 samples drawn from a Gaussian $N(0, 1)$ distribution using gradient-based maximization of the average log-likelihood according at equation 9.

It can be clearly seen that improving the order improves the quality of the fit, while higher order B-splines closely follow $N(0, 1)$.

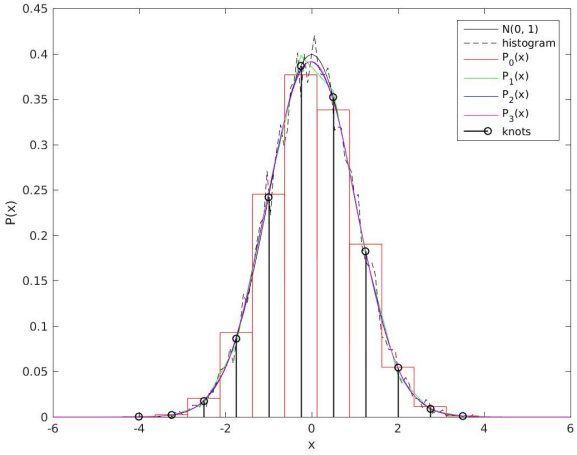

Fig. 2: B-Spline PDFs of order 0,1,2,3 fitting 10,000 samples drawn randomly from $N(0, 1)$.

Further, it is interesting to observe that the higher-order B-splines can capture the peak of the Gaussian even when the peak is located in-between the knots. This is an important advantage over histograms, which fail to capture PDF peaks and it is particularly useful to algorithms that depend on maximum-a-posteriori inference.

In the context of a typical PDF estimator, we can employ the mixture model interpretation of the B-Spline PDF to construct a very simple iterative algorithm, the *Expectation* algorithm, using conditional expectation:

$$\hat{p}(m) = E_x\{p(m|x)\} = E_x\left\{\frac{p(m)p(x|m)}{p(x)}\right\} \approx \frac{1}{N}\sum_{n=1}^{N} \frac{p(m)K'_{k,m}(x_n)}{\sum_{m'} p(m')K'_{k,m'}(x_n)}, \quad (10)$$

where $E_x\{\cdot\}$ is the expectation operator over x. The Expectation algorithm is as follows:

1. Initialize $\hat{p}(m) = \frac{1}{M}$.
2. Iterate $\hat{p}(m) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \frac{\hat{p}(m)K'_{k,m}(x_n)}{\sum_{m'} \hat{p}(m')K'_{k,m'}(x_n)}$ until convergence.

The latter algorithm is the expectation step of the well-known Expectation-Maximization algorithm for mixtures with fixed components and variable component weights. As such, it has the same convergence properties.

For the case of $k = 0$, where the B-Spline PDF corresponds to the histogram, the kernel $K'_{k,m'}(x)$ is either zero or one:

$$\frac{p(m)K'_{k,m}(x)}{\sum_{m'} p(m')K'_{k,m'}(x)} = 1\,(x \in Q_m)\,, \quad (11)$$

where $1(\cdot)$ is the indicator function and the interval Q_m is defined as:

$$Q_m \triangleq [x_m^Q - \frac{\Delta}{2}, x_m^Q + \frac{\Delta}{2}]. \quad (12)$$

Equation 11 allows us to obtain the expectation in a single step:

$$\hat{p}(m) = \frac{\sum_{n=1}^{N} 1\,(x_n \in Q_m)}{N}. \quad (13)$$

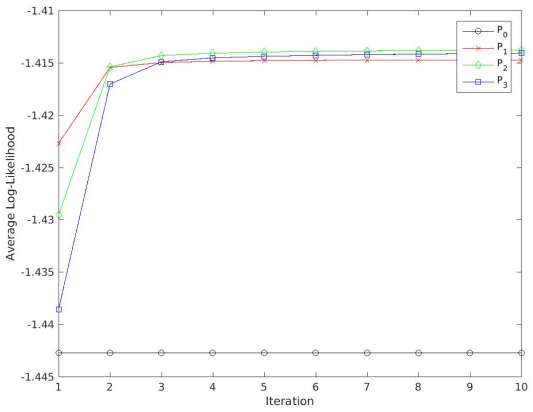

Fig. 3: Training B-Spline PDFs of order 0,1,2,3 with 1000 $N(0,1)$ samples using the Expectation algorithm.

This equation corresponds to the typical frequentistic estimation of histogram bin probabilities. Therefore, histogram bin counting can also be seen as the first step of the Expectation algorithm.

Figure 3 shows the convergence of the Expectation algorithm for B-Spline PDFs of order 0,1,2,3 with 1000 samples drawn from a $N(0,1)$ and $M = 11$ codepoints at locations $x_m^Q = -4.0 + (m-1) * 0.75, m = 1, ..., 11$. The initialization was made using the uniform distribution $p(m) = 1/M$, for all m. A total of 10 iterations are displayed, while the evaluation was made in terms of average log-likelihood. We can observe that the 0-th order B-Spline PDF converges from the first iteration, as indicated by the non-iterative nature of equation 13. Further, the higher the order of the kernel, the more iterations it takes to converge.

6. RANDOM SAMPLING

Drawing samples from a B-Spline PDF is straight-forward; first we randomly select the generating (normalized) kernel and then we generate a sample from that kernel. The kernel is selected by drawing a sample from a multinomial distribution with probabilities $p(m), m = 1, ..., M$. The sample from the selected kernel is constructed using the convolution-related property of moment-generating functions [13]. According to this property and equation 7, the random variable $u_k \sim K_k(\cdot)$ that follows the k-th kernel $K_k(\cdot)$ is equal to the sum of $K+1$ uniform variables $u_0 \sim K_0(\cdot)$: $u_{k'} = \sum_{k=0}^{K} u_0$. Therefore, the overall algorithm that generates a random sample x is:

1. $m' = \text{RandomMultinomial}\{p(1), p(2), ..., P(M)\}$.
2. $x = x_{m'}^Q + \Delta * \sum_{k'=0}^{k} \text{RandomUniform}\{-0.5, 0.5\}$.
3. if $x \notin [x_{min}, x_{max}]$ goto Step 2.

The last step ensures that the drawn sample is within the support of the PDF.

Figure 4 demonstrates the histogram of 500,000 samples generated from the 1-st order B-Spline PDF estimated in the experiment in section 5. Evidently, the histogram closely follows the polyline nature of $P_1(\cdot)$.

7. DISCUSSION

This work demonstrates that the statistical modelling performed by modern generative audio networks corresponds to a uniformly-spaced histogram applied to the companded audio source. Results

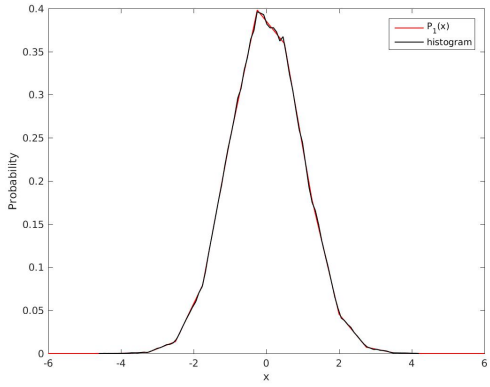

Fig. 4: 1-st order B-Spline PDF and histogram of 500,000 generated random samples.

from high-rate theory provide some justification to use uniformly-spaced histograms to model any source, even ones with infinite support. However, histograms are not suitable models for continuous random variables. This is tackled by generalizing the uniformly-spaced histogram to a B-Spline PDF with a predefined finite degree of continuity: a k-th order B-Spline is continuous up to the $k-1$-th derivative. In fact, histograms can be expressed as 0-th order B-Splines.

The paper provides a very simple formulation of B-Spline PDFs as convolutional mixture models; B-Spline PDF is a uniformly-spaced histogram that is successively smoothed by convolution with a single rectangular function. The simplicity of this formulation allows the derivation of an iterative algorithm, the *Expectation* algorithm, that can be used to estimate the model parameters. The convergence properties of the Expectation algorithm stem from the fact that it corresponds to the expectation step of the well-known Expectation-Maximization algorithm. The algorithm is computationally efficient and robust in the sense that it lacks the maximization step that is frequently hampered by degenerate solutions. Finally, the convolutional nature of the B-Spline PDF allows us to construct a simple random sample generator.

The advantage of the continuous B-Spline PDFs ($P_k(\cdot), k > 0$) over histograms ($P_0(\cdot)$) is clearly depicted in Figure 2: they can closely follow the underlying Gaussian despite having exactly the same number of parameters as histograms.

The fact that researchers choose to use multinomial distributions to model continuous random variables implies that the complexity of the parametric spaces of the alternative continuous models makes training harder. By construction, the B-Spline PDF maintains the simplicity of the parametric space of the multinomial distribution, allowing the development of continuous models that can be trained as efficiently as multinomials.

This paper focused on presenting the theoretical framework of B-Spline PDFs rather than providing application examples, mainly due to space limitations. Accordingly, the author chose to defer the presentation of real world applications to a later publication. Here, the properties of B-Spline PDFs and the suggested algorithms are experimentally demonstrated using illustrative examples. These results suggest that B-Spline PDFs can provide a direct replacement of histograms in any application.

8. REFERENCES

[1] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron C. Courville, and Yoshua Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *CoRR*, vol. abs/1612.07837, 2016.

[2] David W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[3] H. Luter, "Silverman, b. w.: Density estimation for statistics and data analysis. chapman & hall, london new york 1986," *Biometrical Journal*, vol. 30, no. 7, pp. 876–877, 1988.

[4] G. A. Wright and S. M. Zabin, "Nonparametric density estimation for classes of positive random variables," *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1513–1535, Sep 1994.

[5] Charles Kooperberg and Charles J. Stone, "A study of logspline density estimation," *Comput. Stat. Data Anal.*, vol. 12, no. 3, pp. 327–347, Nov. 1991.

[6] Charles J. Stone, "Large-sample inference for log-spline models," *Ann. Statist.*, vol. 18, no. 2, pp. 717–741, 06 1990.

[7] Sam Efromovich, "Orthogonal series density estimation," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 467–476, 2010.

[8] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

[9] ITU Recommendation G.711, "Pulse code modulation of voiced frequencies," *ITU Standards*, 1988.

[10] Yannis Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *ICASSP*, 2015.

[11] I. J. SCHOENBERG, "Contributions to the problem of approximation of equidistant data by analytic functions: Part a.on the problem of smoothing or graduation. a first class of analytic approximation formulae," *Quarterly of Applied Mathematics*, vol. 4, no. 1, pp. 45–99, 1946.

[12] M. Unser, "Splines: a perfect fit for signal and image processing," *IEEE Signal Processing Magazine*, vol. 16, no. 6, pp. 22–38, Nov 1999.

[13] Robert V Hogg and Allen T Craig, *Introduction to mathematical statistics.(5-th edition)*, Upper Saddle River, New Jersey: Prentice Hall, 1995.