DISCRETE PDF ESTIMATION IN THE PRESENCE OF NOISE

Byung-Jun Yoon and P. P. Vaidyanathan

Dept. of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125, USA
E-mail: bjyoon@caltech.edu, ppvnath@systems.caltech.edu

ABSTRACT

The problem of estimating a pdf from measurements has been widely studied by many researchers. However, most of the work was focused on estimating a probability density function of continuous random variables, especially in the absence of noise. In this paper, we consider a model for representing discrete probability density functions based on multirate dsp models. Using this model, we propose an efficient and stable scheme for pdf estimation when the measurements are corrupted by independent additive noise. This approach makes use of well-known results from multirate dsp theory, especially that of biorthogonal partners. Simulation results are given, which clearly show the advantage of the proposed method.

1. INTRODUCTION

The problem of estimating a probability density function from measurements has been widely studied by many researchers for decades. In this problem, we are given a number of observations $X_1, X_2, \ldots, X_N$, and the probability density function $f(x)$ that gave rise to these samples has to be estimated from them. The simplest approach is the histogram, and many other methods have been proposed, each with its own advantages. Despite the simplicity, the histogram approach yields a reasonable estimate of the original pdf when there are enough number of samples. However, it is discontinuous in nature, which makes it less attractive for estimating continuous density functions. It has been observed that model based methods such as the kernel estimators [1] and the wavelet estimators [2, 3] have certain advantages, especially when the number of samples are limited.

For example, let us consider the "kernel" based method, which assumes that the pdf $f(x)$ can be represented as

$$f(x) = \sum_k c_k \phi(x - s_k, \sigma_k) \tag{1}$$

where $\phi(x)$ is called the kernel function. It disperses the mass c_k around the center point s_k, where σ_k decides the extent to which it will disperse the mass. The kernel function $\phi(x)$ can be any appropriate positive function, such as a Gaussian, a spline, etc. The preceding model tries to represent the unknown pdf with a linear combination of shifted copies of the fixed function $\phi(x)$. With the shifts s_k and the dispersion factors σ_k typically fixed, the weighting factors c_k are adjusted based on the measurements of the random variable x, so that the resulting pdf estimate $\widehat{f}(x)$ approximates the original pdf $f(x)$ satisfactorily. One advantage of
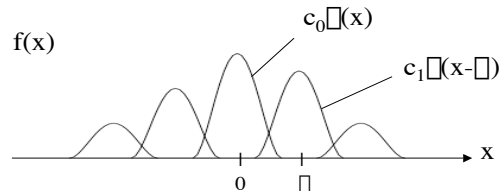
Fig. 1. The pdf representation as a linear combination of shifted versions of the kernel $\phi(x)$.

this method is the fact that the resulting pdf estimate $\widehat{f}(x)$ retains some of the properties of the kernel function. For example, if we choose a $\phi(x)$ with certain smoothness, the estimate $\widehat{f}(x)$ will also enjoy the same property. An example of such a $f(x)$ with uniform shifts and fixed σ_k is shown in Fig. 1. Further discussions on the model based methods can be found in many references, e.g. [1], [2], [3].

Now, let us consider the case when the observations are of the form

$$Y_i = X_i + Z_i, \qquad i = 0, 1, \ldots, N - 1$$

where X_i are the original samples and Z_i are i.i.d. noise that are independent of X_i. If we denote the pdf of X_i and Z_i as $f(x)$ and $k(z)$ respectively, then the density function of Y_i is

$$g(y) = \int k(y - x) f(x) dx. \tag{2}$$

Since the observed samples are distributed according to the pdf $g(y)$, estimating the original pdf $f(x)$ involves deconvolution of the noise pdf $k(z)$. However the inversion of this integral equation (2) is an ill-posed problem, and in general, there is no "good" solution [4]. Many methods have been proposed for estimating a continuous pdf in the presence of noise, and some of the interesting results can be found in [4], [5].

Even though much work has been done in the area of density estimation, most of the work was focused on the continuous case. New methods for modeling and estimating probability mass functions of discrete random variables have been recently proposed in [6], [7], [8]. These models are based on multirate dsp concepts, and they take advantage of well-known results from multirate signal processing theory. In this paper, we consider the discrete pdf model proposed in [6], and develop a novel scheme for estimating a discrete probability density function when the observations are corrupted by additive noise.

All notations are as in [9]. Thus $\downarrow M$ and $\uparrow M$ represent the M-fold decimator and expander respectively. Therefore $[X(z)]_{\downarrow M}$ denotes the z-transform of the decimated version $x(Mn)$, and

similarly $[X(z)]_{\uparrow M} = X(z^M)$ denotes the z-transform of the expanded version.

2. MODEL FOR DISCRETE DENSITY FUNCTIONS

2.1. The PDF Model

Let us consider a discrete probability density function of discrete random variables, which are restricted to have uniformly spaced values (assumed to be integers, without loss of generality). Thus, if we denote the pdf as $x(n)$, it will be a function of an integer random variable n. We assume that this $x(n)$ can be represented as the output of an interpolation filter $f(n)$ preceded by an M-fold expander as proposed in [6]. This can be seen in Fig. 2. The input signal $c(k)$ is the free parameter that is to be adjusted based on the measurements, while M and $f(n)$ are fixed. If we let the subspace $\mathcal{V}_0 = $ span of $\{f(n-kM)\}$ where k is any integer, then $x(n) \in \mathcal{V}_0$, and can be written as

$$x(n) = \sum_k c(k)f(n-kM) \qquad (3)$$

which is a linear combination of $f(n), f(n \pm M), f(n \pm 2M)$, and so on. Notice the analogy to the continuous case in (1).

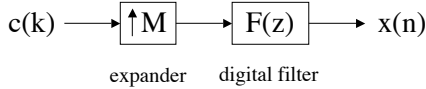**Fig. 2**. The basic pdf model.

If both the driving signal $c(k)$ and the impulse response $f(n)$ of the interpolation filter are in ℓ_2, the resulting pdf $x(n)$ also belongs to ℓ_2, hence $\mathcal{V}_0$ is a subspace of the ℓ_2 space[1]. Since this can be viewed as one channel of a M-channel synthesis filter bank, $\mathcal{V}_0$ is a proper subspace of ℓ_2. For example, if we choose $f(n)$ to be a lowpass filter, the resulting $\mathcal{V}_0$ will be a low frequency subspace. We may choose $f(n)$ such that $\mathcal{V}_0$ includes the pdfs that are of our interest. In fact, we can optimize the filter $f(n)$ for a given class of density functions, and some of the related issues are addressed in [8]. We may also add one or more channels to the model, thereby adding more fine structure to the probability density function $x(n)$. In this paper, we focus on the single channel model in Fig. 2, and explain how the estimation procedure works starting from the measurements.

2.2. Estimation of the PDF

Let us consider again the pdf model in Fig. 2. Assuming that the probability density function $x(n)$ can be represented as the output of this model, how can we get the best estimate based on the measurements? In order to answer this question, let us consider a filter $G(z)$ that satisfies

$$[G(z)F(z)]_{\downarrow M} = 1 \qquad (4)$$

This $G(z)$ is called a biorthogonal partner of $F(z)$ with respect to M [10]. The importance of biorthogonal partners in estimating the probability density function arises as follows. Let us consider

---

[1]Strictly speaking, $F(e^{j\omega})$ should be bounded for this.
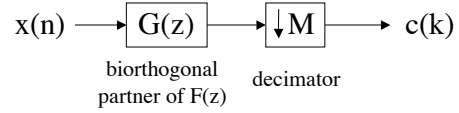

biorthogonal
partner of F(z)   decimator

Fig. 3. Reconstruction of the driving signal $c(k)$.

a signal $x(n)$ that can be represented as in Fig. 2. Therefore, we have

$$X(z) = C(z^M)F(z)$$

From this $x(n)$, we can recover the underlying driving signal $c(k)$ by using a biorthogonal partner $G(z)$ as in Fig. 3. This is not hard to see, since the output of Fig. 3 has the z-transform

$$
\begin{aligned}
[G(z)X(z)]_{\downarrow M} &= [G(z)C(z^M)F(z)]_{\downarrow M} \\
&= C(z)[G(z)F(z)]_{\downarrow M} \\
&= C(z) \qquad \text{(From Eq. (4))}
\end{aligned}
$$

hence $c(k)$ is recovered. Figure 3 shows that $c(k)$ can be written as

$$c(k) = \sum_n x(n)g(Mk-n) \qquad (5)$$

Notice that the signal $x(n)$ is a pdf of an integer random variable n. Therefore the variable n in (5) should be interpreted as a random variable that is distributed according to $x(n)$ (instead of as the traditional "time index"). From this point of view, $g(Mk-n)$ is also a random variable because n is random, and the right hand side of (5) can be viewed as the expectation of the random variable $g(Mk-n)$ with respect to n. Therefore (5) can be rewritten as

$$c(k) = E_n[g(Mk-n)] \qquad (6)$$

Now, assume that we have N measurements of the random variable n, and denote them as $n_i, 0 \le i \le N-1$. Given these measurements, the expectation in (6) can be approximated by its sample mean as follows

$$\hat{c}(k) = \frac{1}{N} \sum_{i=0}^{N-1} g(Mk-n_i)$$

which allows us to relate the measurements to the pdf estimate. If we define the signal $h(n)$ as the relative occurrence of the integer value n in the measurements $\{n_i\}$, we can write $\hat{c}(k)$ as

$$\hat{c}(k) = \sum_n h(n)g(Mk-n) \qquad (7)$$

This shows that we can get an estimate of the driving signal $c(k)$ by feeding the histogram $h(n)$ to the decimation filter $g(n)$ and decimating the output by M, as shown in Fig. 4. Now that we have the estimate $\hat{c}(k)$, this can be used in the original model Fig. 2 to obtain the estimate $\hat{x}(n)$ of the original pdf. The entire picture is shown in Fig. 4. Note that $\hat{x}(n) \in \mathcal{V}_0$, and the above estimation procedure can be viewed as a projection of the histogram $h(n)$ onto the subspace $\mathcal{V}_0$, where the original pdf $x(n)$ belongs.

Since the biorthogonal partner of $F(z)$ is not unique [10], the quality of the estimate $\hat{x}(n)$ may vary depending on the choice of the partner $G(z)$. If we choose $G(z)$ to be the least squares partner defined as

$$G(z) = \frac{\widetilde{F}(z)}{([\widetilde{F}(z)F(z)]_{\downarrow M})_{\uparrow M}} \qquad (8)$$
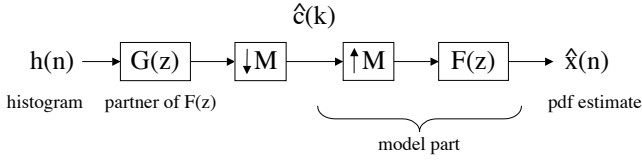
Fig. 4. Estimation of the driving signal $c(k)$ from the histogram $h(n)$, and subsequent estimation of the pdf $x(n)$.

the entire estimation procedure in Fig. 4 becomes simply the orthogonal projection of $h(n)$ onto $\mathcal{V}_0$ [10]. It is shown in [6] that in this case the estimate $\hat{x}(n)$ is guaranteed to be closer to the true pdf $x(n)$, than the histogram $h(n)$ is, in ℓ_2 sense. In other words, we always have

$$\|h(n) - x(n)\| \geq \|\hat{x}(n) - x(n)\|. \tag{9}$$

One problem in taking $G(z)$ to be the least squares partner of $F(z)$ is that $G(z)$ may be an unstable filter. If we consider the denominator of $G(z)$, which is

$$B(z) = ([\widetilde{F}(z)F(z)]_{\downarrow M})_{\uparrow M}$$

we can see that it satisfies $B(z) = \widetilde{B}(z) = B^*(1/z^*)$. Therefore if $B(z)$ has a zero at z_0, then there exists another zero at $1/z_0^*$, hence it has zeros both inside and outside the unit circle. This can be a problem since it means that $G(z)$ cannot be a causal stable filter. However, it is possible to approximate such a filter by an FIR filter by choosing the region of convergence properly, as long as there are no poles on the unit circle. It is shown in [6] that we may use an FIR truncation of the least squares partner $G(z)$, and it was observed that this approach yields impressive estimation results when compared to the traditional histogram approach.

3. ESTIMATION OF PDFS IN THE PRESENCE OF NOISE

Now, let us suppose that the original samples are corrupted by noise. Our measurement $\{m_i\}$ can be expressed as

$$m_i = n_i + e_i, \qquad i = 0, 1, \ldots, N-1$$

where $\{n_i\}$ is the original sample and $\{e_i\}$ is i.i.d. noise that is independent of $\{n_i\}$. Let $x(n)$ be the probability density function of n_i and let $e(n)$ be the pdf of the noise e_i. Then the density $y(n)$ of the observation m_i can be written as

$$y(n) = x(n) * e(n)$$

Since the pdf $x(n)$ comes from the model in Fig. 2, $y(n)$ can be represented as the output of the following model in Fig. 5.
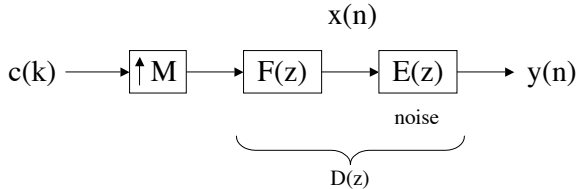

Fig. 5. The original pdf convolved with the noise pdf.

Therefore if we let $D(z) = F(z)E(z)$, we can write $Y(z) = C(z^M)D(z)$. Now, let us define $S(z)$ as the least squares partner of the filter $D(z)$, so that

$$S(z) = \frac{\widetilde{D}(z)}{([\widetilde{D}(z)D(z)]_{\downarrow M})_{\uparrow M}}$$

We can recover the driving signal $c(k)$ by passing $y(n)$ through $S(z)$ and decimating it by M as shown in Fig. 6.
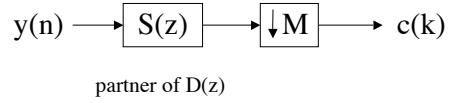

Fig. 6. Reconstruction of the driving signal from the pdf in the presence of noise.

To prove this, note that the output of Fig. 6 has the z-transform

$$\begin{aligned}
[S(z)Y(z)]_{\downarrow M} &= [S(z)C(z^M)D(z)]_{\downarrow M} \\
&= C(z)[S(z)D(z)]_{\downarrow M}
\end{aligned}$$

Since $S(z)$ is the LSBP of $D(z)$, they satisfy $[S(z)D(z)]_{\downarrow M} = 1$, and the above expression reduces to $C(z)$. Now that we know the driving signal $c(k)$, if we pass it through the system in Fig. 2, we can get the original pdf $x(n)$ back. As $S(z)$ has poles both inside and outside the unit circle (unless $D(z)$ has order $< M$), it cannot be directly used. But we can use the FIR truncation $S_L(z)$ instead, as explained in [6]. The whole estimation procedure is illustrated in Fig. 7.
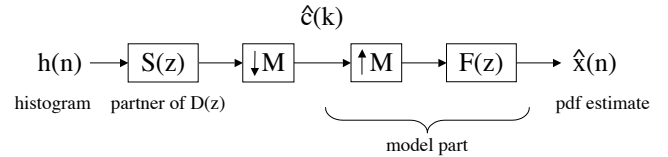

Fig. 7. Estimation of the pdf in the presence of noise.

The system in Fig. 7 takes $h(n)$ as the input, which is a coarse representation of the corrupted pdf $y(n)$. It eliminates the effect of the noise and yields an estimate $\hat{x}(n)$ of the original pdf $x(n)$ as the output. In order to ensure that $\hat{x}(n)$ is a valid pdf, we may drop the negative coefficients and then normalize the pdf estimate.

Experiment shows that this approach has a considerable advantage over the traditional inverse filtering method, which is as follows. Since we know that $h(n)$ is a representation of $y(n) = x(n) * e(n)$, we may pass $h(n)$ through the inverse filter $1/E(z)$ in order to get rid of the measurement noise, as shown in Fig. 8. Generally, the noise pdf $e(n)$ will be symmetric around $n = 0$, resulting in a zero-mean random noise. Due to the symmetry, $E(z)$ will have zeros both inside and outside the unit circle. Therefore
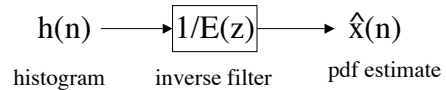

Fig. 8. Traditional way to estimate the pdf in the presence of noise using the inverse filter $1/E(z)$.
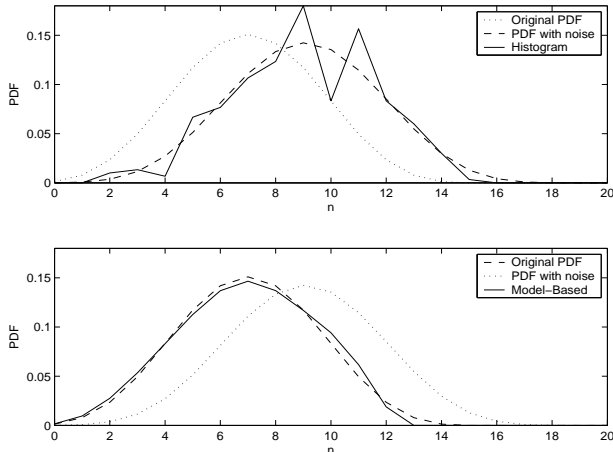
Fig. 9. PDF estimation when noise is present. Top plot: the original pdf, the pdf with noise and the histogram. Bottom plot: the original pdf, the pdf with noise and the model based pdf estimate.

the corresponding inverse filter $1/E(z)$ will have poles both inside and outside the unit circle resulting in an unstable filter. In such cases, we may again use the FIR truncation method, as in [6]. But experiment shows that this tends to amplify the estimation error present in the histogram, making the pdf estimate very unreliable. The reason why this does not happen when we use the least squares partner $S(z)$ as in the previous discussion, is because the whole system works as an orthogonal projection operator, which tends to suppress the error instead of amplifying it.

4. SIMULATION RESULTS

In order to demonstrate the ideas, we present the following simulation results. We assume $M = 2$ and use $F(z) = (1+z)^L/2^L$ with $L = 6$. Notice that this filter leads to the 5th order spline function [12]. By choosing an appropriate driving signal $c(k)$, we obtained a sample pdf $x(n)$ of length 37. In addition to this, a sample noise pdf $e(n)$ of length 5 is chosen[2]. Now we let $D(z) = F(z)E(z)$, and compute $S(z)$, which is the least squares partner of $D(z)$. We generated 500 random variables $n_i \sim x(n)$ and the same number of noise random variables $e_i \sim e(n)$. Adding these random variables respectively, we obtained 500 noisy observations $m_i = n_i + e_i$. The histogram was constructed from these observations, as can be seen in Fig. 9 (top). The figure clearly shows that there is a large difference between the original pdf and the histogram due to the noise. In addition to this, there exists also a considerable amount of estimation error between the histogram and the pdf with noise. In fact

$$\sum_n |h(n) - y(n)|^2 = 0.0068871$$

where $y(n) = x(n) * e(n)$ is the pdf with noise. Now, let us consider the model based pdf estimate. The histogram is fed into the model in Fig. 7, and the output is normalized after removing the negative coefficients, to get the final estimate $\hat{x}(n)$. The result is shown in the bottom plot of Fig. 9. It can be noticed that the

noise is effectively removed, resulting in an excellent estimate of the original pdf. The estimation error between the estimate $\hat{x}(n)$ and the original pdf $x(n)$ was

$$\sum_n |\hat{x}(n) - x(n)|^2 = 0.0004767$$

Note that this error is much smaller than the initial estimation error between the histogram $h(n)$ and the corrupted pdf $y(n)$. Conventional inverse filtering of the histogram yielded a very oscillatory output, a considerable portion of which was negative, and therefore we have not shown the result here.

5. CONCLUDING REMARKS

In this paper, we proposed a novel scheme for estimating a pdf of a discrete random variable in the presence of measurement noise. It was observed that the proposed scheme can practically get rid of the effect of the noise, yielding a satisfactory estimate of the original pdf $x(n)$. An interesting extension of the proposed ideas is the estimation of a probability density function when the noise is dependent on the original samples. This is a topic for future work.

6. REFERENCES

[1] Chiu, S. T. "Bandwidth selection for kernel density estimation", Ann. Stat., vol. 19, pp. 1883-1905, 1991.

[2] Walter, G. G. and Ghorai, J. K., "Advantages and disadvantages of density estimation with wavelets", Comp. Sci., Stat., vol. 24, pp/ 234-243, 1993.

[3] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. "Density estimation by wavelet thresholding", Ann. Stat., vol. 24, pp. 508-539, 1996.

[4] Walter, G. G., "Density estimation in the presence of noise", Stat. and Prob. letters, 41, pp. 237-246, 1999.

[5] Walter, G. G. and Shen, X., "Continuous Non-Negative Wavelets and Their Use In Density Estimation", Communication in Statistics 28, 1-18, 1999

[6] Byung-Jun Yoon and P. P. Vaidyanathan , "Improved estimation of discrete probability density functions using multirate models", Proc. 37th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2003.

[7] P. P. Vaidyanathan and Byung-Jun Yoon, "Discrete probability density estimation using multirate DSP models", Proc. 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, Apr. 2003.

[8] Byung-Jun Yoon and P. P. Vaidyanathan, "Non-parametric estimation of discrete probability density functions using multirate DSP models", Internal Report, Dept. of Electrical Engineering, California Institute of Technology, May 2003.

[9] Vaidyanathan, P. P., *Multirate systems and filter banks*, Prentice Hall, Inc., 1993.

[10] P. P. Vaidyanathan and B. Vrcelj, "Biorthogonal partners and applications", IEEE Trans. Signal Processing, vol. 49(5), pp. 1013-1027, May 2001.

[11] Čencov, N. N., "Evaluation of an unknown distribution density from observations", Doklady, (3):1559-1562, 1962.

[12] Strang, G. and Nguyen, T. *Wavelets and filter banks*, Wellesley-Cambridge Press, MA, 1997.

---

[2]The noise pdf $e(n)$ had the following coefficients : {0.0532, 0.2339, 0.3780, 0.2660, 0.0689}