A MULTIRATE DSP MODEL FOR ESTIMATION OF DISCRETE PROBABILITY DENSITY FUNCTIONS

Byung-Jun Yoon, *Student Member*, *IEEE*, and P. P. Vaidyanathan, *Fellow*, *IEEE*

February 5, 2004

Contact Author: P. P. Vaidyanathan
Dept. of Electrical Engineering 136-93, California Institute of Technology, Pasadena, CA 91125, USA. Phone: (626) 395-4681 E-mail: ppvnath@systems.caltech.edu

EDICS: 2-MWAV, 2-FILB

ABSTRACT

The problem of estimating a probability density function from measurements has been widely studied by many researchers. Even though much work has been done in the area of PDF estimation, most of it was focused on the continuous case. In this paper, we propose a new model-based approach for modeling and estimating discrete probability density functions, or probability mass functions. This approach is based on multirate signal processing theory, and it has several advantages over the conventional histogram method. We illustrate the PDF estimation procedure and analyze the statistical properties of the PDF estimates. Based on this model, a novel scheme is introduced that can be used for estimating the PDF in the presence of noise. Furthermore, the proposed ideas are extended to the more general case of estimating multivariate probability density functions. Finally, we also consider practical issues such as optimizing the coefficients of a digital filter which is an integral part of the model. This allows us to apply the proposed model to solve real world problems. Simulation results are given where appropriate, in order to demonstrate the ideas. [1]

[1]Work supported in part by the ONR grant N00014-99-1-1002, USA.

1 Introduction

The problem of estimating a probability density function has been widely studied by many researchers in the mathematics as well as signal processing communities for many decades [1], [2], [3], [4], [5], [6], [10], [11], [14], [15], [16], [20]. The goal is to obtain a good estimate of a PDF $f(v)$ of a random variable v from the observations. The most common way to estimate density functions is the histogram method, and many other methods have been suggested, each with its own advantages and disadvantages. Even though histograms may give reasonable estimates of the true PDF when there are enough observations, it is discontinuous in nature, making it less preferable for estimating continuous random variables. It has been shown that a model-based approach has several advantages compared to the histogram method, especially when the number of observations is limited [3], [4]. For example, the kernel based method assumes that the PDF $f(v)$ can be represented as

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

where $\phi(v)$ 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 $\phi(v)$ can be any appropriate positive function, such as a Gaussian [5], a spline [6], etc. The preceding model tries to represent the unknown PDF with a linear combination of shifted copies of the fixed function $\phi(v)$. With the shifts s_k and the dispersion factors σ_k typically fixed[2], the weighting factors c_k are adjusted based on the measurements of the random variable v, so that the resulting PDF estimate $\widehat{f}(v)$ approximates the original PDF $f(v)$ satisfactorily. One advantage of this method is the fact that the resulting PDF estimate $\widehat{f}(v)$ retains most of the properties of the kernel function. For example, if we choose a $\phi(v)$ with certain smoothness, the estimate $\widehat{f}(v)$ will also enjoy the same property. Let us consider the histogram in Fig. 1(a). This can be considered to be a special case of (1) where $\phi(v)$ is chosen to be a rectangular pulse, σ_k are fixed so that the width of the pulse is Δ, and the shifts are uniform satisfying $s_k = k\Delta$. In this case, the mass c_k will be taken to be proportional to the number of observations that fall in the domain of the kth pulse $\phi(v - k\Delta)$. Generally, $\phi(v)$ will be chosen such that it is smooth so that we can obtain a smooth PDF estimate. Figure 1(b) shows an example of such a $f(v)$ with uniform shifts and fixed σ_k. Further discussions on model-based methods can be found in many references, e.g. [3], [4], [5], [6].

Even though much work has been done in the area of PDF estimation, most of it was focused on the continuous case. In this paper, we propose a PDF model for 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 argument n. The proposed

[2]The shifts s_k and the dispersion factors c_k may also be adjusted based on the measurements. For example, reference [7] addresses this in the context of the estimation of Gaussian mixture models.

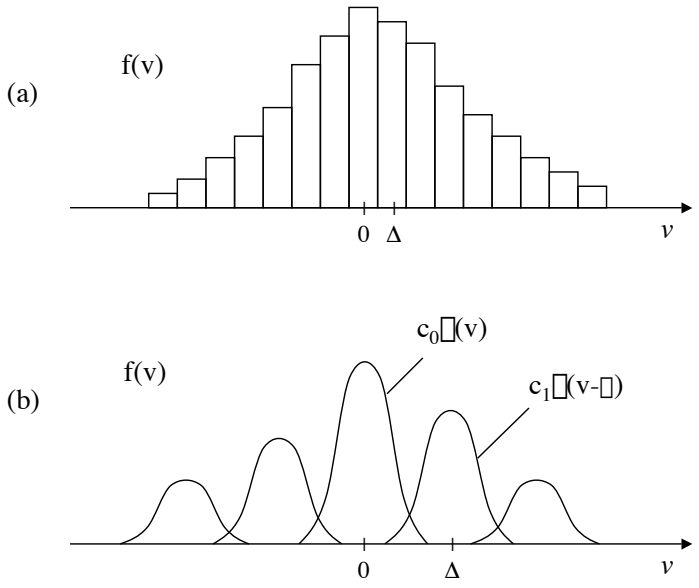

Figure 1: (a) Histogram as a special case of kernel based representation when $\phi(v)$ is rectangular. (b) The PDF representation as a linear combination of shifted versions of the kernel $\phi(v)$.

model is based on multirate signal processing concepts and it will be shown that it has several advantages over the traditional histogram based method.

1.1 Outline

In section 2 we introduce the basic multirate filter model for modeling a discrete probability density function. Analogy to the kernel method and the relation to the wavelet method will be pointed out. We illustrate how we can get a PDF estimate based on the observations. It will be seen that the concept of biorthogonal partners [8] plays an important role in the estimation procedure. In section 3 we illustrate the estimation procedure in more detail, considering practical issues such as how to obtain a positive PDF estimate. Simulation results will be given that clearly show the advantage of the proposed model. In section 4 we propose an efficient and stable scheme for estimating a PDF in the presence of noise. The statistical properties of the estimates such as bias and variance are analyzed in section 5, and compared to the bias and variance of the histogram estimate. The model-based approach can also be used for modeling and estimating a joint PDF of several random variables, which will be illustrated in section 6. Finally, in section 7 we consider optimizing the coefficients of a digital filter that is an integral part of the model, which opens up the way for this model to be used in practical applications.

1.2 Notations

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 Multirate Filter Model for Discrete PDFs

2.1 Basic Model

Let us consider a discrete probability density function $x(n)$ 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 shown in Fig. 2, where $F(z)$ is the z-transform of $f(n)$.

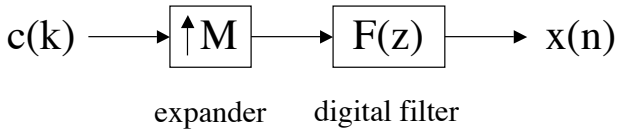

Figure 2: The basic PDF model.

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 $\{f(n - Mk)\}$ where k is any integer, then $x(n) \in \mathcal{V}_0$, and can be written as

$$x(n) = \sum_k c(k) f(n - Mk) \tag{2}$$

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). 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 subset of the ℓ_2 space[3]. Since this can be viewed as one channel of a M-channel synthesis filter bank, $\mathcal{V}_0$ is in fact a proper subspace of ℓ_2. If we choose $f(n)$ to be a lowpass filter, the resulting $\mathcal{V}_0$ will be a low frequency subspace.

We can add one or more channels to the model, thereby adding more fine scale components to the probability density function. Figure 3 shows one possible example of such a multi-channel PDF model. In this model $F_0(z)$ and $F_1(z)$ occupy different bands in the frequency domain. Suppose we have a multi-channel model with M channels with an m_k-fold expander in the k th channel, where

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

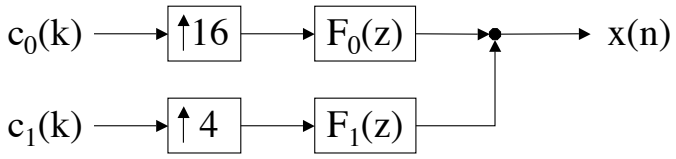

Figure 3: An example of a two channel PDF model.

m_k satisfies

$$\sum_{k=0}^{M-1} 1/m_k = 1. \tag{3}$$

In this case, if the corresponding filters $F_k(z)$ are from a perfect reconstruction filter bank, the subspace $\mathcal{V}_0$ can be the whole ℓ_2 space. This, in principle, can be related to the density estimation method based on wavelet thresholding, which was proposed by Donoho et. al [10]. Since $x(n)$ can be any function in ℓ_2, the model may look degenerate in this case. But the value of the wavelet based method lies in the fact that nonlinear operations (such as thresholding) can be done in the subbands to improve the quality of the estimate. In fact, it has been observed that this method suppresses the estimation error without compromising the sharpness of the underlying density function [10].

2.2 Estimating The PDF

Let us consider again the single channel model in Fig. 2. For a PDF $x(n) \in \mathcal{V}_0$, 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

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

This $G(z)$ is called a biorthogonal partner of $F(z)$ with respect to M [8]. One obvious example of such a filter is $G(z) = 1/F(z)$. In fact, any $G(z)$ that can be expressed in the form

$$G(z) = \frac{A(z)}{([A(z)F(z)]_{\downarrow M})_{\uparrow M}} \tag{5}$$

for some $A(z)$ is a biorthogonal partner of $F(z)$, hence the partner is not unique. It is also possible to have an FIR biorthogonal partner $G(z)$ under mild conditions on $F(z)$ [8]. A detailed study of biorthogonal partners can be found in [8].

The importance of biorthogonal partners in estimating the probability density function arises as follows. Let us consider $x(n)$ in Fig. 2. Its z-transform $X(z)$ can be written as

$$X(z) = C(z^M)F(z). \tag{6}$$

From this $x(n)$, we can recover the underlying driving signal $c(k)$ by using a biorthogonal partner $G(z)$ as in Fig. 4. To show this, observe that the output of Fig. 4 has the z-transform

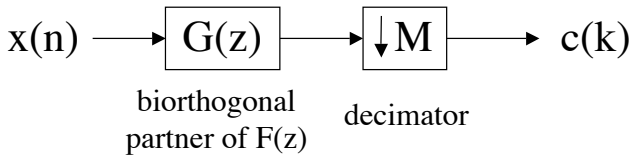

Figure 4: Reconstruction of the driving signal $c(k)$.

$$\begin{aligned}
[G(z)X(z)]_{\downarrow M} &= [G(z)C(z^M)F(z)]_{\downarrow M} \\
&= C(z)[G(z)F(z)]_{\downarrow M} \\
&= C(z) \qquad \text{(From Eq. (4))}
\end{aligned} \tag{7}$$

hence $c(k)$ is recovered. Figure 4 shows that $c(k)$ can be written as

$$c(k) = \sum_n x(n)g(Mk - n). \tag{8}$$

Notice that the signal $x(n)$ is a probability density function of an integer random variable n. Therefore the variable n in the equation above 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 (8) can be viewed as the expectation of the random variable $g(Mk - n)$ with respect to n. Therefore (8) can be rewritten as

$$c(k) = E_n[g(Mk - n)]. \tag{9}$$

This kind of interpretation of a signal as the expectation of a random variable naturally appears in almost any non-parametric density estimation scheme [4], [10], [11], [12], [13], the earliest being perhaps the work of Čencov [11] in 1962. In fact, this plays an important role in the PDF estimation method being proposed in this paper, since this allows us to relate the measurements to the PDF estimate. Assume that we have N measurements of the random variable n, and denote them as $n_i, 0 \leq i \leq N-1$. Given these measurements, the expectation in (9) can be approximated by its sample mean as follows

$$\hat{c}(k) = \frac{1}{N} \sum_{i=0}^{N-1} g(Mk - n_i). \tag{10}$$

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

$$\hat{c}(k) = \sum_n h(n)g(Mk - n). \tag{11}$$

Since $h(n)$ is nothing but the histogram obtained from the measurements $\{n_i\}$, this means that we can get an estimate of the driving signal $c(k)$ by feeding the histogram $h(n)$ to the decimation filter

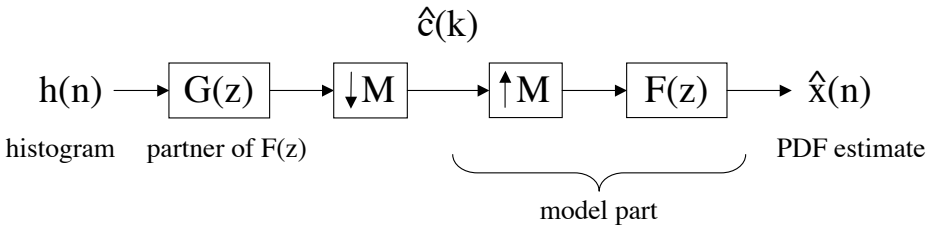

Figure 5: Estimation of the driving signal $c(k)$ from the histogram $h(n)$, and subsequent estimation of the PDF $x(n)$.

$g(n)$ and decimating the output by M. This is shown in Fig. 5. Now that we have the estimate $\hat{c}(k)$, this can be used in the original model shown in Fig. 2 to obtain the estimate $\hat{x}(n)$ of the original PDF. The entire picture is shown in Fig. 5. The following summarizes the procedure of PDF estimation:

1. We make measurements $\{n_i\}$ of the random variable n.
2. We construct the histogram $h(n)$ from the measurements. Notice that $h(n)$ is a coarse representation of the original PDF $x(n)$, and it need not belong to the subspace $\mathcal{V}_0$.
3. The histogram obtained from above is passed through the system that is shown in Fig. 5 to obtain the PDF estimate $\hat{x}(n)$. This PDF estimate belongs to $\mathcal{V}_0$ as the original PDF.

Note that if the input to the system in Fig. 5 is $x(n) \in \mathcal{V}_0$, since $G(z)$ and $F(z)$ are biorthogonal partners, the output is also $x(n)$. So, if we denote the input-output mapping of Fig. 5 by a linear operator $\mathcal{P}$, then $\mathcal{P}x(n) = x(n)$. Now consider an arbitrary input signal $x'(n) \in \ell_2$ and let us denote the corresponding output by $y(n) = \mathcal{P}x'(n)$. Since $y(n) \in \mathcal{V}_0$, we have $\mathcal{P}^2x'(n) = \mathcal{P}\{\mathcal{P}x'(n)\} = \mathcal{P}y(n) = y(n) = \mathcal{P}x'(n)$. This shows that $\mathcal{P}^2 = \mathcal{P}$, which means that the operator $\mathcal{P}$ is a projection operator. From this, we can see that the above estimation procedure is just a *projection of the histogram $h(n)$ onto the subspace $\mathcal{V}_0$*, where the original PDF $x(n)$ belongs.

2.3 The Choice of G(z)

Let us assume that $F(z)$ and M are fixed. Since the biorthogonal partner of a filter $F(z)$ is not unique, the quality of the estimate $\hat{x}(n)$ may vary depending on the choice of the partner $G(z)$. So, the natural question that may arise is how to choose $G(z)$ in order to obtain the best PDF estimate $\hat{x}(n)$, based on the limited number of measurements. To answer this question, let us consider the following. Given $h(n)$, suppose we wish to find the signal $\hat{x}(n) \in \mathcal{V}_0$ that is closest to $h(n)$ in least square sense, i.e. we want to minimize

$$\|h(n) - \hat{x}(n)\|^2 \stackrel{\triangle}{=} \sum_n |h(n) - \hat{x}(n)|^2. \tag{12}$$

The resulting $\hat{x}(n)$ is in fact the orthogonal projection of $h(n)$ onto $\mathcal{V}_0$. It can be shown [8] that if the filter $G(z)$ in Fig. 4 is chosen as

$$G(z) = \frac{\widetilde{F}(z)}{([\widetilde{F}(z)F(z)]_{\downarrow M})_{\uparrow M}} \tag{13}$$

where $\widetilde{F}(z) \stackrel{\triangle}{=} F^*(1/z^*)$, then $\hat{x}(n)$ is indeed the orthogonal projection of $h(n)$ onto $\mathcal{V}_0$. Since this $G(z)$ is also a biorthogonal partner of $F(z)$ [8], it is called the least squares biorthogonal partner (LSBP). If we choose any other biorthogonal partner $G(z)$, the projection $\mathcal{P}$ will be "oblique" rather than orthogonal. The advantage of the orthogonal projection is that the projected signal $\hat{x}(n)$ is guaranteed to be closer to the original PDF $x(n)$ than the histogram $h(n)$ is. In other words, we always have

$$\|h(n) - x(n)\| \geq \|\hat{x}(n) - x(n)\|. \tag{14}$$

In order to see this, note that since the $\hat{x}(n)$ is the orthogonal projection of $h(n)$ onto $\mathcal{V}_0$, we can write $h(n) = \hat{x}(n) + e(n)$ where $e(n) \in \mathcal{V}_0^c$. Therefore, we have $h(n) - x(n) = \hat{x}(n) - x(n) + e(n)$. As $\hat{x}(n) - x(n) \in \mathcal{V}_0$ and $e(n) \in \mathcal{V}_0^c$, it follows that

$$\begin{aligned} \|h(n) - x(n)\|^2 &= \|\hat{x}(n) - x(n)\|^2 + \|e(n)\|^2 \\ &\geq \|\hat{x}(n) - x(n)\|^2 \end{aligned} \tag{15}$$

hence proving (14). Now, suppose that we are going to choose the decimation filter $G(z)$ to be the LSBP of $F(z)$ with respect to M as in (13). If we look at the denominator $B(z) = ([\widetilde{F}(z)F(z)]_{\downarrow M})_{\uparrow M}$ of $G(z)$, we can see that it satisfies

$$B(z) = \widetilde{B}(z) = B^*(1/z^*). \tag{16}$$

Therefore if $B(z)$ has a zero at z_0, then there exists another zero at $1/z_0^*$. This can be a problem, since it means that $G(z)$ cannot have all the poles inside the unit circle, and therefore it cannot be a causal stable filter. One way to get around this problem is to choose $F(z)$ such that its magnitude square is Nyquist(M), i.e.

$$[\widetilde{F}(z)F(z)]_{\downarrow M} = 1. \tag{17}$$

In this case, the least squares partner becomes $G(z) = \widetilde{F}(z)$, which can be written as $g(n) = f^*(-n)$ in the time-domain. We can observe that (17) is equivalent to imposing the orthonormality constraint on the basis functions $\{f(n - kM)\}$ that span the subspace $\mathcal{V}_0$. An interpolation filter $F(z)$ that satisfies (17) can be designed using one of many known techniques [9].

2.4 The Square-Root Model

However, the use of an $F(z)$ that satisfies (17) suffers from one disadvantage, namely the fact that the positivity of the outcome $x(n)$ cannot be guaranteed. This is an important point when using

the estimation process shown in Fig. 5, since the resulting estimate $\hat{x}(n)$ may not be positive. If we design the filter $F(z)$ as in (17), then $f(n)$ will necessarily have negative coefficients unless it has order $< M$. Since the projection $\hat{x}(n)$ consists of a linear combination of shifted copies of $f(n)$, it is highly probable that $\hat{x}(n)$ will have some negative coefficients as well.

A simple way to overcome this problem is to use the output of the model in Fig. 2 to model the square-root of the PDF instead of the PDF itself. This is similar to the idea proposed by Good and Gaskins many years ago for estimating continuous PDFs [14]. In this model, we assume that $x_s(n)$ which is the square-root of the PDF can be represented as the output of the filter $F(z)$ preceded by an M-fold expander. Therefore we have

$$x_s(n) = \sum_k c_s(k) f(n - kM) \tag{18}$$

where the probability density function $x(n)$ is

$$x(n) = \{x_s(n)\}^2. \tag{19}$$

We can still use the model similar to Fig. 5 with slight modifications in the estimation procedure. The square-root PDF model is elaborated in [15], and it has the advantage that the resulting PDF estimate is guaranteed to be positive.

Despite this advantage, the square-root model has also several shortcomings. For example, in order to get a satisfactory estimate from the measurements, we have to adjust the sign of the square-root of the histogram, before it is used in the estimation procedure [15], [16]. The searching process for the optimal signature sequence can be computationally very expensive. Another disadvantage of this model is the fact that the estimation results are not easy to analyze analytically due to the nonlinearity of the model. In the next section, we consider a method that is free from all these problems.

3 The FIR Truncation of The LSBP

In this section, we consider a linear model for representing probability density functions, which ensures that the resulting estimate is positive, and uses only stable and realizable filters in the estimation procedure. Let us consider again the model in Fig. 2. In order to ensure that the PDF estimate is non-negative, all the coefficients of the filter $f(n)$ should be non-negative as well as the driving signal $c(k)$. Now take $G(z)$ to be the least squares partner of $F(z)$ as in (13). We know from section 2.4 that unless $F(z)$ has a filter order $< M$, $G(z)$ has poles both inside and outside the unit-circle, which 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 [17]. In order to illustrate the idea, let us consider an all-pole

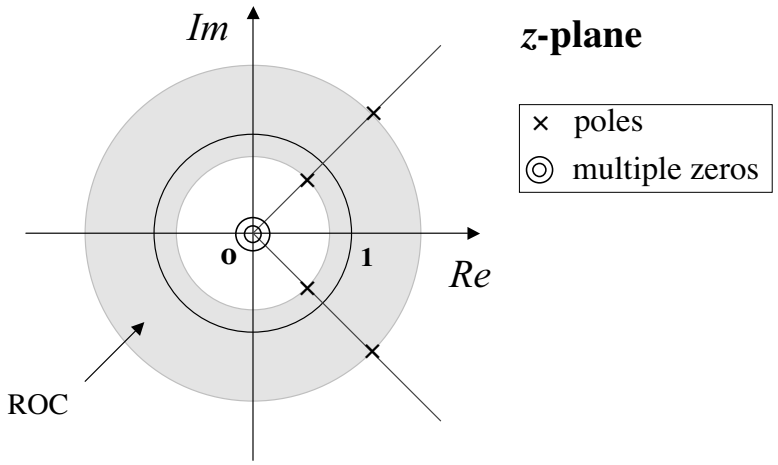

Figure 6: Pole-zero plot of $Q(z)$.

filter $Q(z)$ defined as

$$Q(z) = \prod_i \frac{1}{(1 - p_i z^{-1})}. \tag{20}$$

Its pole-zero plot may be as in Fig. 6. We assume that there are no poles lying on the unit circle. Let p_{in} be the pole with the largest modulus among all poles inside the unit circle. Similarly, let p_{out} be the pole with the smallest modulus among all poles outside the unit circle. We choose the region of convergence to be the annular region in the z-plane that satisfies $|p_{in}| < |z| < |p_{out}|$. Then $Q(z)$ can be expressed as follows using the partial fraction expansion

$$Q(z) = \sum_{|p_i|<1} \frac{c_i}{1 - p_i z^{-1}} + \sum_{|p_i|>1} \frac{c_i}{1 - p_i z^{-1}}. \tag{21}$$

The first term in the right hand side of (21) corresponds to a right-sided sequence, whereas the second term corresponds to a left-sided sequence. Using inverse z-transform, (21) will be

$$q(n) = \sum_{|p_i|<1} c_i p_i^n u(n) - \sum_{|p_i|>1} c_i p_i^n u(-n-1) \tag{22}$$

in the time domain. Now, we may truncate $q(n)$ to get an FIR filter $q_L(n)$ as follows.

$$q_L(n) = \begin{cases} q(n) & \text{if } |n| \le L \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

This corresponds to multiplying $q(n)$ with a rectangular window. Note that L should be large enough such that most of the energy of $q(n)$ is confined in $|n| \le L$. Unless there are poles very close to the unit circle, it is possible to approximate $q(n)$ with a reasonable length L. For example, it is shown in [17] that the cubic B-spline filter can be well approximated by a truncated FIR filter of length only five or seven.

Returning to our original interest, let us consider again $G(z)$, the least squares partner of $F(z)$. Let $B(z)$ be the denominator of $G(z)$. If we choose $f(n)$ such that $f(n) \geq 0$ and $B(z) = ([\widetilde{F}(z)F(z)]_{\downarrow M})_{\uparrow M}$ has zeros sufficiently away from the unit circle, it is possible to approximate $G(z)$ by an FIR filter $G_L(z)$, by truncating it using a window function. Using this FIR filter $G_L(z)$ in place of $G(z)$ in Fig. 5, we can use a similar estimation procedure as elaborated in section 2.2. However, one more remark remains to be made regarding the positivity of the PDF estimate. Since the interpolation filter $f(n)$ is non-negative, it is possible to make the output signal also non-negative by taking a non-negative driving signal $c(k)$. So, when modeling the original PDF $x(n) = \sum_k c(k)f(n - Mk)$, we can make it a valid PDF by choosing $c(k) \geq 0$ and normalizing $x(n)$ so that it adds up to 1. But when feeding the histogram $h(n)$ into the system shown in Fig. 5, there is no guarantee that the estimate $\hat{c}(k)$ will be non-negative for $\forall k$. Correspondingly, the orthogonal projection $\hat{x}(n)$ may not satisfy the non-negativity condition. In order to guarantee that the PDF estimate is non-negative, we simply drop the negative values of $\hat{x}(n)$ to obtain a positive estimate as follows.

$$\hat{x}_p(n) = \begin{cases} \hat{x}(n) & \text{if } \hat{x}(n) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Note that this $\hat{x}_p(n)$ may not necessarily belong to $\mathcal{V}_0$. However, this is not a problem since by taking $\hat{x}_p(n)$ instead of $\hat{x}(n)$, the PDF estimate gets even closer to the original PDF $x(n)$. To see this, notice that

$$\begin{aligned} \|\hat{x}(n) - x(n)\|^2 &= \sum_n |\hat{x}(n) - x(n)|^2 \\ &= \sum_{\hat{x}(n)\geq 0} |\hat{x}(n) - x(n)|^2 + \sum_{\hat{x}(n)<0} |\hat{x}(n) - x(n)|^2 \\ &\geq \sum_{\hat{x}(n)\geq 0} |\hat{x}(n) - x(n)|^2 + \sum_{\hat{x}(n)<0} |x(n)|^2 \\ &= \|\hat{x}_p(n) - x(n)\|^2. \end{aligned}$$

Combining this result with the inequality in (14), we get the following inequality

$$\|h(n) - x(n)\| \geq \|\hat{x}_p(n) - x(n)\| \tag{25}$$

which guarantees that the PDF estimate $\hat{x}_p(n)$ is always closer to the true PDF $x(n)$ than the histogram is.

In order to demonstrate the ideas, let us consider the following example. We assume $M = 2$ and use $F(z) = (1 + z)^6/2^6$. Notice that this filter leads to the 5th order spline function [18]. $G(z)$ is chosen to be the least squares partner of $F(z)$, truncated by a rectangular window of length 39. Since all the poles of $G(z)$ are located far away from the unit circle, the energy - or the square of the ℓ_2 norm - of the truncated filter $\|g_L(n)\|^2$ was almost identical to $\|g(n)\|^2$ in this case[4]. By

[4]In order to preserve a "near-biorthogonality", L has to be chosen such that $\|g_L(n)\|^2$ is almost identical to $\|g(n)\|^2$.

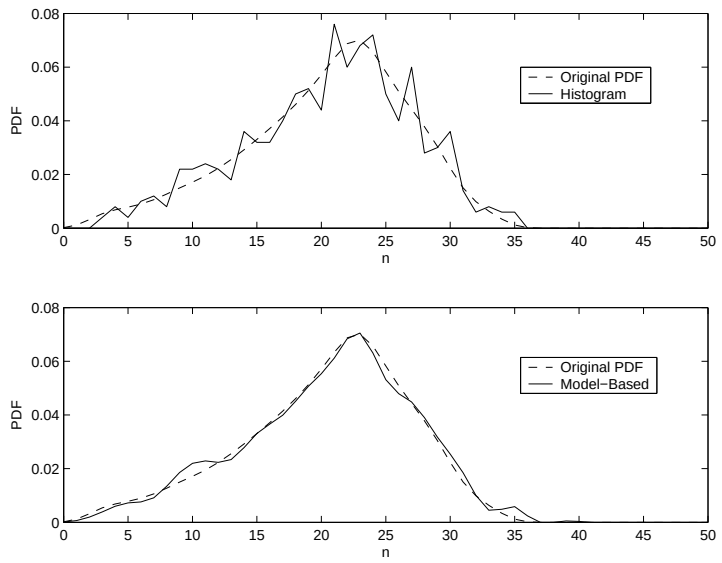

Figure 7: PDF estimation using FIR truncation of the LSBP. Top plot: the original PDF and the histogram. Bottom plot: the original PDF and the model-based PDF estimate.

choosing an appropriate driving signal $c(k)$ for the sake of generating an example, we obtained a sample PDF $x(n)$ of length 37. We made 500 measurements of the random variable n according to $x(n)$, and the histogram $h(n)$ was obtained from the observations. Then the histogram was passed through the system in Fig. 5 to get the orthogonal projection of the histogram. Finally, negative values in the output of Fig. 5 were dropped, and the result was normalized to get the PDF estimate $\hat{x}(n)$. Fig. 7 shows the simulation results. The histogram shown in Fig. 7(top) is quite different from the original PDF, whereas the model-based estimate is considerably close to the PDF, as can be seen in Fig. 7(bottom). The estimation error was

$$\sum_n |h(n) - x(n)|^2 = 0.00151128 \tag{26}$$

for the histogram, and

$$\sum_n |\hat{x}(n) - x(n)|^2 = 0.00017605 \tag{27}$$

for the model-based estimate, which is only about 12% of the error of the histogram. This clearly shows that the proposed approach is superior to the conventional histogram method.

4 Estimating The PDF in The Presence of Noise

Another advantage of the PDF model in Fig. 2 is the fact that it gives rise to a simple and efficient way of removing noise present in the measurements as we shall show next. Suppose that the original

samples are corrupted by additive noise. Our measurements $\{m_i\}$ can be expressed as

$$m_i = n_i + e_i \qquad (i = 0, 1, ..., N-1) \tag{28}$$

where $\{n_i\}$ are the original samples and $\{e_i\}$ are i.i.d. noise that are independent of $\{n_i\}$. Let the probability density function of n_i be $x(n)$ and let $e(n)$ be the PDF of the noise random variable e_i. Then the density $y(n)$ of the measurement m_i is

$$y(n) = (x * e)(n). \tag{29}$$

Since the PDF $x(n)$ comes from the model in Fig. 2, $y(n)$ can be represented as the output of the model shown in Fig. 8. Therefore if we let $D(z) = F(z)E(z)$, we can write $Y(z) = C(z^M)D(z)$.

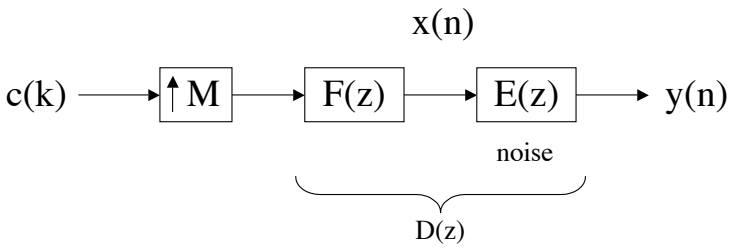

Figure 8: The original PDF convolved with the noise PDF.

Now let us define $S(z)$ as the least squares partner of the filter $D(z)$, so that

$$S(z) = \frac{\widetilde{D}(z)}{([\widetilde{D}(z)D(z)]_{\downarrow M})_{\uparrow M}} \tag{30}$$

We can recover the driving signal $c(k)$ by passing $y(n)$ through $S(z)$ and decimating it by M as shown in Fig. 9. This can be proved in a similar manner as (7). Knowing 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 in section 3. The whole estimation procedure is illustrated in Fig. 10. The system in Fig. 10 takes $h(n)$ as the input, which is a coarse representation of the PDF $y(n)$ of the noisy observations. It eliminates the effect of the noise and

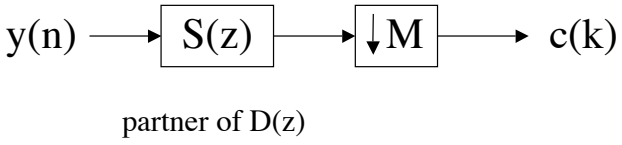

Figure 9: Reconstruction of the driving signal from the PDF in the presence of noise.

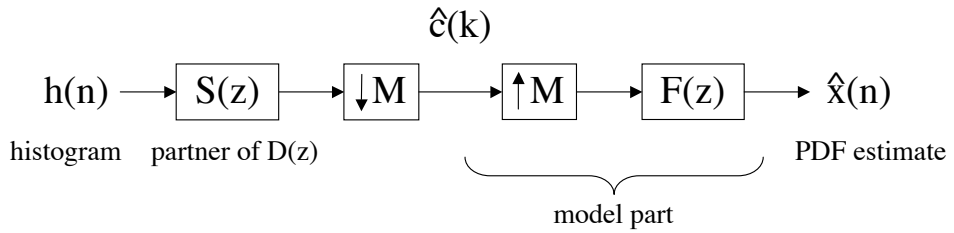

Figure 10: Estimation of the PDF in the presence of noise.

yields an estimate $\hat{x}(n)$ of the original PDF $x(n)$ as expected. In order to ensure that $\hat{x}(n)$ is a valid PDF, we should drop the negative coefficients and then normalize the PDF estimate as described in section 3.

Experiment shows that this approach has a considerable advantage over the inverse filtering method. Knowing the noise PDF $e(n)$, we may use the inverse filter $1/E(z)$ to filter out the noise as shown in Fig. 11. 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 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 section 3. But experiment shows that this tends to amplify the estimation error that is present in the histogram, which makes the PDF estimate very unreliable. The reason why this did not happen when we were using the least squares partner $S(z)$ in the previous discussion, was because the whole system worked as an orthogonal projection operator, which tends to suppress the error instead of amplifying it.

To demonstrate, consider the following example. We assume $M = 2$, and use the same filter as in section 3. In addition to this, a sample noise PDF $e(n)$ of length 5 is chosen[5]. From this, we compute the filter $D(z) = F(z)E(z)$, and let $S(z)$ be the least squares partner of $D(z)$. We made 500 measurements $n_i \sim x(n)$ and generated the same number of observations of the random noise $e_i \sim e(n)$. Adding these measurements 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.

[5]The noise PDF $e(n)$ had the following coefficients : $[e(0) \cdots e(4)] = [0.0532\ 0.2339\ 0.3780\ 0.2660\ 0.0689]$.

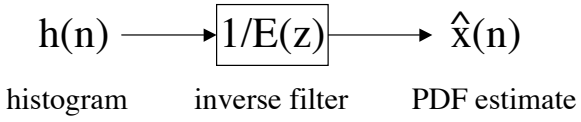

Figure 11: Traditional way to estimate the PDF in the presence of noise using the inverse filter $1/E(z)$.

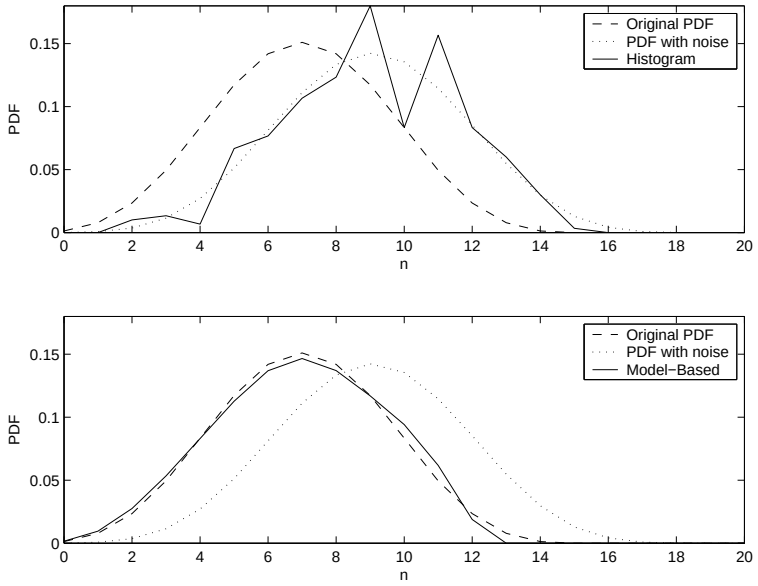

Figure 12: PDF estimation when noise is present. Top plot: the original PDF, the PDF with noise and the histogram. Bottom plot: the original PDF, the PDF with noise and the model-based PDF estimate.

12(top). The figure clearly shows that there is a huge difference between the original PDF and the histogram due to the noise. In addition to this, there exists also a considerable estimation error between the histogram and the PDF with noise, resulting in an error of

$$\sum_n |h(n) - y(n)|^2 = 0.0068871 \tag{31}$$

where $y(n) = (x * e)(n)$ is the PDF of the noisy samples. Now, consider the model-based PDF estimate. The histogram is put into the model in Fig. 10, 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. 12. It can be noticed that the noise is effectively removed, resulting in an excellent estimate of the original PDF. The final estimation error between the estimate $\hat{x}(n)$ and the original PDF $x(n)$ was

$$\sum_n |\hat{x}(n) - x(n)|^2 = 0.0004767. \tag{32}$$

Note that this error is much smaller than the initial estimation error between the histogram $h(n)$ and the PDF with noise $y(n)$. 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 Bias and Variance of The Estimates

Since the PDF estimates are based on random observations, the estimates themselves are random as well. Therefore, it is important to understand their statistical properties such as the bias and the variance. Let θ be the true value (or function) that is to be estimated, and $\hat{\theta}_N$ be an estimate based on N observations. An estimate is unbiased if $\mathcal{E}\{\hat{\theta}_N\} = \theta$ [19]. It is desirable to have an estimate that is unbiased and has a small variance at the same time. In the following sections, we focus on the model in Fig. 2, analyze its bias and variance, and finally compare them with those of the histogram method. It will be demonstrated that both methods yield unbiased estimates. However the model-based method results in a smaller variance, giving us a more reliable estimate than the histogram method.

5.1 Histogram Method

Let us first consider the histogram method. Assume that we have N observations of a random variable n, where the underlying PDF is $x(n)$. The probability that the i th observation will be $n_i = n$ is

$$P\{n_i = n\} = x(n) \tag{33}$$

for all $i = 0, 1, 2, ..., N-1$. Therefore if we let $t(n)$ be the number of observations in $\{n_i\}$ that have the value n, the probability $P\{t(n) = k\}$ is simply

$$P\{t(n) = k\} = \left(\begin{array}{c} N \\ k \end{array} \right) x^k(n)\Big(1 - x(n)\Big)^{N-k} \tag{34}$$

and therefore $t(n)$ is a binomial random variable that follows $B(N, x(n))$. From this, we have $\mathcal{E}[t(n)] = Nx(n)$ and $\mathrm{Var}[t(n)] = Nx(n)\Big(1 - x(n)\Big)$. Notice that the histogram can be represented as $h(n) = t(n)/N$. Therefore the expectation of $h(n)$ is

$$\mathcal{E}[h(n)] = \frac{1}{N}\mathcal{E}[t(n)] = x(n) \tag{35}$$

which shows that the histogram estimate is unbiased. Also from $h(n) = t(n)/N$, we get the following variance for $h(n)$.

$$\mathrm{Var}[h(n)] = \frac{1}{N^2}\mathrm{Var}[t(n)] = \frac{1}{N}x(n)\Big(1 - x(n)\Big). \tag{36}$$

The variance of the histogram estimate, which is defined as $\sum_n \mathrm{Var}[h(n)]$ is therefore

$$\mathcal{E}\Big[\|h(n) - x(n)\|^2\Big] = \sum_n \mathrm{Var}[h(n)] = \frac{1}{N}\Big(1 - \sum_n x^2(n)\Big). \tag{37}$$

From (37) we can see that the variance of the estimate decreases as the number of observations N increases, as expected.

5.2 Model-Based Method

Let us consider the model in Fig. 5 again. We can write the output as

$$\hat{x}(n) = \sum_l \sum_k h(k) g(Ml - k) f(n - Ml). \tag{38}$$

Therefore, the expectation $\mathcal{E}[\hat{x}(n)]$ can be written as

$$\begin{aligned} \mathcal{E}[\hat{x}(n)] &= \sum_l \sum_k \mathcal{E}[h(k)] g(Ml - k) f(n - Ml) \\ &= \sum_l \sum_k x(k) g(Ml - k) f(n - Ml). \end{aligned} \tag{39}$$

We can see that the last expression in (39) is the output of Fig. 5 when the input signal is $x(n)$. Since $x(n) \in \mathcal{V}_0$ and as passing an input signal through Fig. 5 is just a projection onto $\mathcal{V}_0$, the right hand side of (39) reduces to $x(n)$. This proves that the model-based method results in an unbiased PDF estimate.

In order to compute the variance of $\hat{x}(n)$, let us first consider $\mathcal{E}[\hat{x}^2(n)]$

$$\mathcal{E}[\hat{x}^2(n)] = \sum_{k_1,k_2} \sum_{l_1,l_2} \mathcal{E}[h(k_1)h(k_2)] g(Ml_1 - k_1) g(Ml_2 - k_2) f(n - Ml_1) f(n - Ml_2). \tag{40}$$

It can be shown that $\mathcal{E}[h(k_1)h(k_2)]$ is

$$\mathcal{E}[h(k_1)h(k_2)] = \begin{cases} \frac{N-1}{N} x(k_1)x(k_2) & \text{if } k_1 \neq k_2 \\ \frac{1}{N} x(k_1)\Big(1 + (N-1)x(k_1)\Big) & \text{if } k_1 = k_2. \end{cases} \tag{41}$$

Detailed derivation of (41) can be found in the appendix. From this, we get

$$\mathcal{E}[\hat{x}^2(n)] = \frac{1}{N} \sum_k \sum_{l_1,l_2} x(k) g(Ml_1 - k) g(Ml_2 - k) f(n - Ml_1) f(n - Ml_2) + \frac{N-1}{N} x^2(n). \tag{42}$$

Since $\mathcal{E}[\hat{x}(n)] = x(n)$, we get the following variance for $\hat{x}(n)$

$$\begin{aligned} \mathrm{Var}[\hat{x}(n)] &= \mathcal{E}[\hat{x}^2(n)] - \Big(\mathcal{E}[\hat{x}(n)]\Big)^2 \\ &= \frac{1}{N} \sum_k \sum_{l_1,l_2} x(k) g(Ml_1 - k) g(Ml_2 - k) f(n - Ml_1) f(n - Ml_2) - \frac{1}{N} x^2(n). \end{aligned} \tag{43}$$

Using this, it can be shown that the variance of the PDF estimate is

$$\begin{aligned} \mathcal{E}\Big[\|\hat{x}(n) - x(n)\|^2\Big] &= \sum_n \mathrm{Var}[\hat{x}(n)] \\ &= \frac{1}{N} \left(\sum_k \sum_l x(k + Ml) g(-k) f(k) - \sum_n x^2(n) \right). \end{aligned} \tag{44}$$

Now, let us compare the variance of the two estimates. It can be shown that the variance of the model-based estimate is always smaller than that of the histogram. In fact, this is an immediate

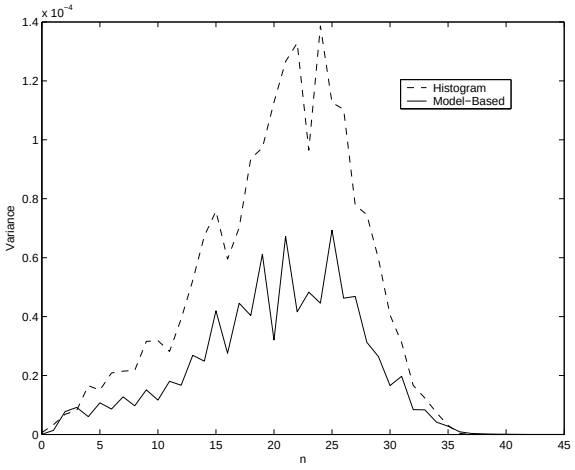

Figure 13: The variance of the histogram and the model-based estimate.

result of (15). If we subtract the variance of the model-based estimate from that of the histogram, we get

$$\begin{aligned} \mathcal{E}\Big[\|h(n)-x(n)\|^2\Big] - \mathcal{E}\Big[\|\hat{x}(n)-x(n)\|^2\Big] &= \mathcal{E}\Big[\|h(n)-x(n)\|^2 - \|\hat{x}(n)-x(n)\|^2\Big] \\ &= \mathcal{E}\Big[\|e(n)\|^2\Big] \geq 0. \end{aligned} \tag{45}$$

This shows that the model-based PDF estimate in Fig. 5 has a smaller variance than the histogram-based estimate. The reduced variance is due to the fact that the PDF estimate $\hat{x}(n)$ is restricted to $\mathcal{V}_0$, which is a proper subspace of ℓ_2.

Consider again the example given in section 3. We now compute the bias and the variance of the histogram and the model-based estimate. We made 500 measurements of the random variable n and constructed the histogram and the model-based estimate. This experiment was repeated 100 times. The mean value of the 100 estimates was almost identical to the original PDF $x(n)$ for both methods, which verifies that both the histogram estimate and the model-based estimate are unbiased. Figure 13 shows the variance $\mathrm{Var}[h(n)]$ and $\mathrm{Var}[\hat{x}(n)]$ at each n. The figure shows that the variance of the histogram is larger than that of the model-based PDF estimate, at nearly all points. The variance, or equivalently the mean squared error (MSE), of the histogram estimate was

$$\sum_n \mathrm{Var}[h(n)] = 0.00191544 \tag{46}$$

and the variance of the model-based estimate was

$$\sum_n \mathrm{Var}[\hat{x}(n)] = 0.00091091. \tag{47}$$

which is less than half of the variance of the histogram estimate. These values are indeed very close to the theoritical values computed from (37) and (44). The variances predicted by equations (37)

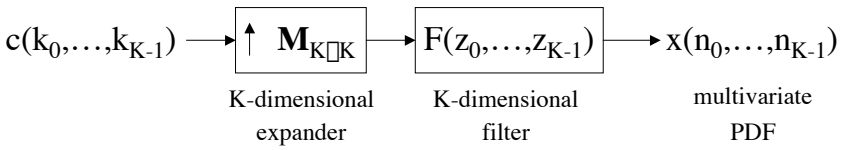

Figure 14: The multivariate PDF model.

and (44) are

$$\mathcal{E}\Big[\|h(n) - x(n)\|^2\Big] = 0.00191044, \quad \mathcal{E}\Big[\|\hat{x}(n) - x(n)\|^2\Big] = 0.00091044 \tag{48}$$

which are very close to the values in (46) and (47) obtained from the simulation.

6 The Multivariate PDF Estimation

Up to this point, we have considered only probability density functions of a single random variable n. However, it is not hard to extend the basic ideas in the univariate PDF estimation to the multivariate case. Suppose we have K random variables $(n_0, ..., n_{K-1})$, where the joint PDF is $x(n_0, ..., n_{K-1})$. We may model such a K-dimensional density function, using a K-dimensional interpolation filter $F(z_0, ..., z_{K-1})$ and a $K \times K$ sampling matrix $\mathbf{M}$. This is shown in Fig. 14. Suppose we are going to use a separable filter

$$F(z_0, ..., z_{K-1}) = F_0(z_0)F_1(z_1)\cdots F_{K-1}(z_{K-1}) \tag{49}$$

and a diagonal sampling matrix

$$\mathbf{M} = \begin{bmatrix} M_0 & 0 & \cdots & 0 \\ 0 & M_1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & M_{K-1} \end{bmatrix} \tag{50}$$

for modeling the PDF. In this case, the estimation procedure becomes very similar to that in section 2.2. First of all, we compute the least squares partner of $F_i(z)$ as in (13) for all $i = 0, \ldots, K-1$. Therefore, we have

$$G_i(z) = \frac{\widetilde{F}_i(z)}{([\widetilde{F}_i(z)F_i(z)]_{\downarrow M_i})_{\uparrow M_i}} \qquad (i = 0, 1, \ldots, K-1). \tag{51}$$

As we have seen in section 3, we may approximate this least squares partner by a truncated FIR filter, if necessary. Now, these filters can be used to construct the following filter

$$G(z_0, ..., z_{K-1}) = G_0(z_0)G_1(z_1)\cdots G_{K-1}(z_{K-1}). \tag{52}$$

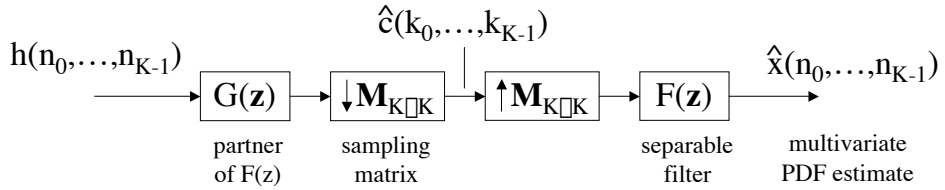

Figure 15: Estimation of a multivariate PDF.

It can be seen that $G(\mathbf{z})$ is a biorthogonal partner of $F(\mathbf{z})$ with respect to the sampling matrix $\mathbf{M}$, where $\mathbf{z} = [z_0 \cdots z_{K-1}]$. This is shown in the following.

$$\begin{aligned}
[G(\mathbf{z})F(\mathbf{z})]_{\downarrow\mathbf{M}} &= [G_0(z_0)\cdots G_{K-1}(z_{K-1})F_0(z_0)\cdots F_{K-1}(z_{K-1})]_{\downarrow\mathbf{M}} \\
&= [G_0(z)F_0(z)]_{\downarrow M_0} \ \cdots \ [G_{K-1}(z)F_{K-1}(z)]_{\downarrow M_{K-1}} \\
&= 1
\end{aligned} \tag{53}$$

Note that we can reconstruct $C(\mathbf{z})$, if we pass $x(\mathbf{n})$ through $G(\mathbf{z})$ and decimate it by the sampling matrix $\mathbf{M}$. To prove this, notice that when using the filter in (49) and the sampling matrix in (50), the output $X(\mathbf{z})$ of Fig. 14 can be written as

$$X(\mathbf{z}) = C(z_0^{M_0} \ldots z_{K-1}^{M_{K-1}})F(\mathbf{z}). \tag{54}$$

If we pass it through the filter $G(\mathbf{z})$ and decimate by $\mathbf{M}$, we get

$$\begin{aligned}
[X(\mathbf{z})G(\mathbf{z})]_{\downarrow\mathbf{M}} &= [C(z_0^{M_0} \ldots z_{K-1}^{M_{K-1}})F(\mathbf{z})G(\mathbf{z})]_{\downarrow\mathbf{M}} \\
&= C(\mathbf{z})[F(\mathbf{z})G(\mathbf{z})]_{\downarrow\mathbf{M}} \\
&= C(\mathbf{z}).
\end{aligned}$$

Now the estimation procedure is as follows. Firstly, we construct the histogram based on the observations. Next, we pass it through the K-dimensional filter $G(\mathbf{z})$ and decimate the output by $\mathbf{M}$. This yields the estimate $\widehat{C}(\mathbf{z})$ of the original driving signal $C(\mathbf{z})$. Then we feed this estimate to the PDF model in Fig. 14, to get the PDF estimate $\hat{x}(n_0, ..., n_{K-1})$. The entire estimation procedure can be found in Fig. 15. It has to be noted that, although the filter $F(z_0, ..., z_{K-1})$ is separable and the matrix $\mathbf{M}$ is diagonal, the resulting PDF may not be a separable PDF. For example, let us choose $C(z_0, z_1), F(z_0, z_1)$ as follows

$$\begin{aligned}
C(z_0, z_1) &= 1 + z_0^{-1} z_1^{-2} \\
F(z_0, z_1) &= (1 + z_0^{-1})(1 + z_1^{-1})
\end{aligned}$$

and use the following sampling matrix

$$\mathbf{M} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}. \tag{55}$$

Then the output $X(z_0, z_1)$ will be

$$X(z_0, z_1) = (1 + z_0^{-2} z_1^{-6})(1 + z_0^{-1})(1 + z_1^{-1}) \tag{56}$$

which is clearly not separable, since it cannot be separated into $X(z_0, z_1) = X_0(z_0)X_1(z_1)$.

In order to demonstrate the idea, consider the following example. We take $F_0(z) = F_1(z) = (1 + z^{-1})^6/2^6$, and define $F(z_0, z_1) = F_0(z_0)F_1(z_1)$. The sampling matrix $\mathbf{M}$ is assumed to be the following diagonal matrix

$$\mathbf{M} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}. \tag{57}$$

By feeding a proper driving signal $c(k_0, k_1)$ into the model in Fig. 14, we obtained the bivariate PDF $x(n_0, n_1)$ in Fig. 16(top) as the output. Based on this joint PDF $x(n_0, n_1)$, we made 5000 measurements of (n_0, n_1), and computed the PDF $h(n_0, n_1)$ from these observations. The resulting two-dimensional histogram is shown in Fig. 16(center). We can see that there are many peaks in the histogram, which were not present in the original PDF, indicating that the estimation error has degraded the estimate considerably. Now let us define $G_0(z)$ and $G_1(z)$ as the least squares partners of $F_0(z)$ and $F_1(z)$ respectively, as in (51). The analysis filter in Fig. 15 is chosen to be $G(z_0, z_1) = G_0(z_0)G_1(z_1)$. The histogram $h(n_0, n_1)$ is passed through the single channel filter bank in Fig. 15, which results in the PDF estimate $\hat{x}(n_0, n_1)$. This is shown in the bottom plot in Fig. 16. It clearly shows that the model-based approach yields a much better estimate of the original PDF, by removing most of the peaks in the histogram to give us a smoother output. The estimation error of each PDF estimate was

$$\sum_{n_0} \sum_{n_1} |h(n_0, n_1) - x(n_0, n_1)|^2 = 0.00019012 \tag{58}$$

and

$$\sum_{n_0} \sum_{n_1} |\hat{x}(n_0, n_1) - x(n_0, n_1)|^2 = 0.00003194. \tag{59}$$

We can see that the proposed method results in a considerably smaller estimation error, which is only 16.80% compared to that of the histogram method.

7 Optimization of F(z)

7.1 Optimizing F(z) for PDF Estimation

We have seen in the previous discussion, that the model-based approach has many advantages over the traditional histogram method. Knowing that the original PDF $x(n)$ belongs to $\mathcal{V}_0$, we could reduce the estimation error dramatically. Moreover, the model-based approach yielded an unbiased estimate with a smaller variance. However, in general we do not know in advance what the subspace $\mathcal{V}_0$ will look like. Unless there is a way to choose the interpolation filter $f(n)$ such that it includes

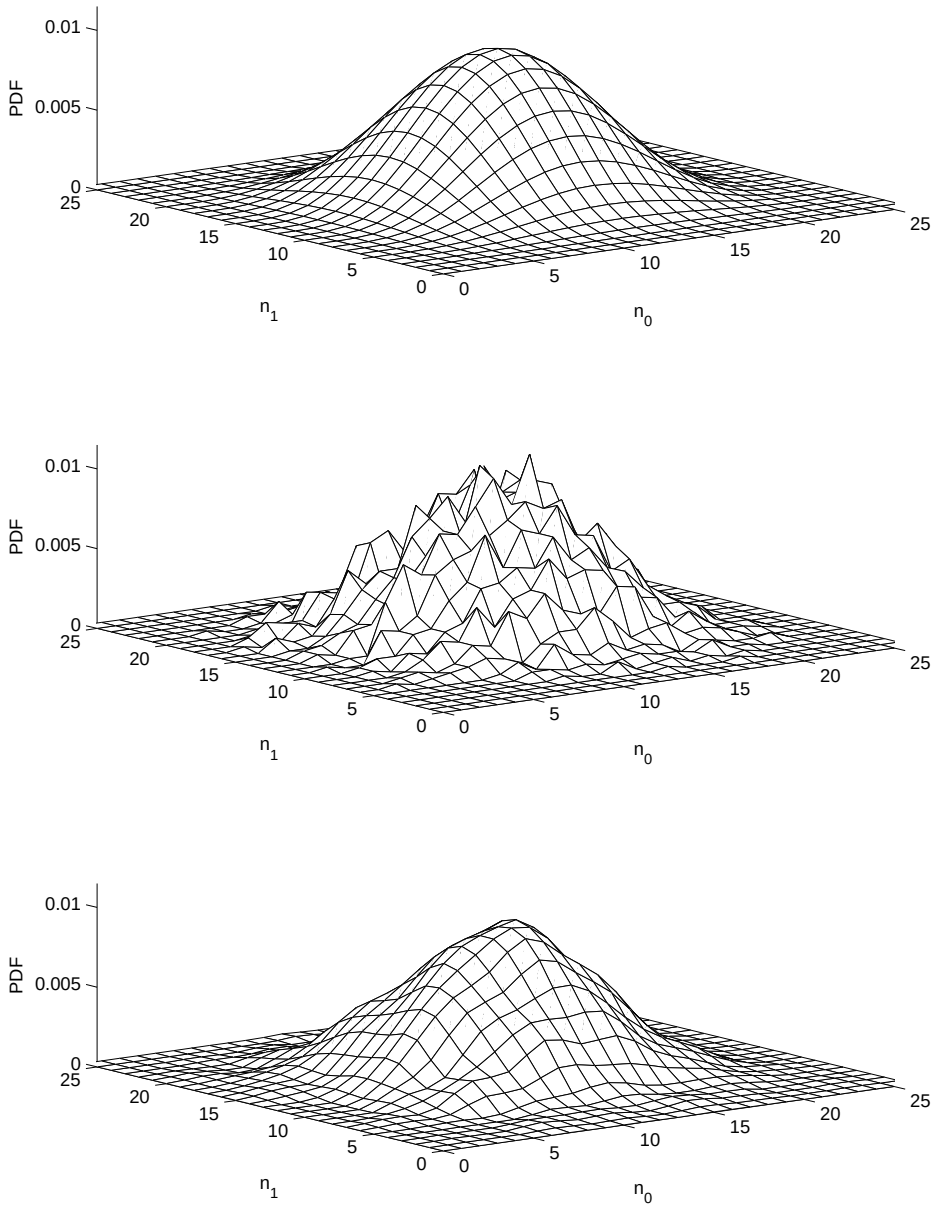

Figure 16: Multivariate PDF estimation. Top plot: the original PDF. Center plot: histogram estimate. Bottom plot: model-based estimate.

all the PDFs of interest, the estimation procedure elaborated in the previous sections may not result in a satisfactory PDF estimate. Therefore the need for optimizing the filter coefficients of the interpolation filter $f(n)$ arises naturally.

Let $\mathcal{W}_0$ be a set of I PDFs $\{x_i(n)|i = 0, 1, \ldots, I-1\}$ belonging to a certain class (e.g., samples from Laplacian PDFs, samples from Poisson PDFs, etc.). We would like our PDF subspace $\mathcal{V}_0$ to be such that the orthogonal projections $\hat{x}_i(n)$ of $x_i(n)$ onto $\mathcal{V}_0$ are as close as possible to the original PDFs $x_i(n)$ when averaged over all elements in the given set $\mathcal{W}_0$. That is, we would like to minimize the sum of the ℓ_2 norms

$$\sum_i \|x_i(n) - \hat{x}_i(n)\|^2 \tag{60}$$

by choosing $\mathcal{V}_0$, or equivalently by designing $f(n)$. Since $\hat{x}_i(n) \in \mathcal{V}_0$ and $x_i(n) - \hat{x}_i(n) \in \mathcal{V}_0^c$, we have

$$\|x_i(n)\|^2 = \|\hat{x}_i(n)\|^2 + \|x_i(n) - \hat{x}_i(n)\|^2. \tag{61}$$

Therefore, minimizing the sum of $\|x_i(n) - \hat{x}_i(n)\|^2$ is equivalent to maximizing the sum of the energies of the projection $\|\hat{x}_i(n)\|^2$. This is in fact an energy compaction problem, which is similar to the problems addressed in [21], [22], [23].

In order to make this a well-posed optimization problem, let us consider the ith sample PDF $x_i(n)$. Let $\hat{x}_i(n)$ be the orthogonal projection of $x_i(n)$ onto $\mathcal{V}_0$ as before. This projection $\hat{x}_i(n)$ can be obtained by passing $x_i(n)$ through the model in Fig. 5 where $G(z)$ is the least squares partner of $F(z)$. Suppose we let the error ξ_i be

$$\begin{aligned} \xi_i &= \|x_i(n) - \hat{x}_i(n)\|^2 \\ &= \sum_n |x_i(n) - \hat{x}_i(n)|^2. \end{aligned} \tag{62}$$

Then small ξ_i implies that the sample PDF $x_i(n)$ can be represented by the model in Fig. 2 satisfactorily. Now, the optimization problem can be stated as follows. We want to find the filter coefficients $f(n)$ that minimize the following objective function

$$J = \sum_{i=0}^{I-1} \alpha_i \xi_i \tag{63}$$

where the error ξ_i is defined as in (62), and α_i is a weighting factor that satisfies $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$.

When minimizing J, there are several constraints that have to be taken into account. We may first limit the length of the filter to be $\leq R$, so that $f(n)$ is zero outside $0 \leq n \leq R-1$. Secondly, we restrict the filter coefficients to be non-negative, i.e. $f(n) \geq 0, \forall n$. This is necessary in order to make the PDF estimate non-negative. Thirdly, we impose the constraint on $f(n)$ such that the

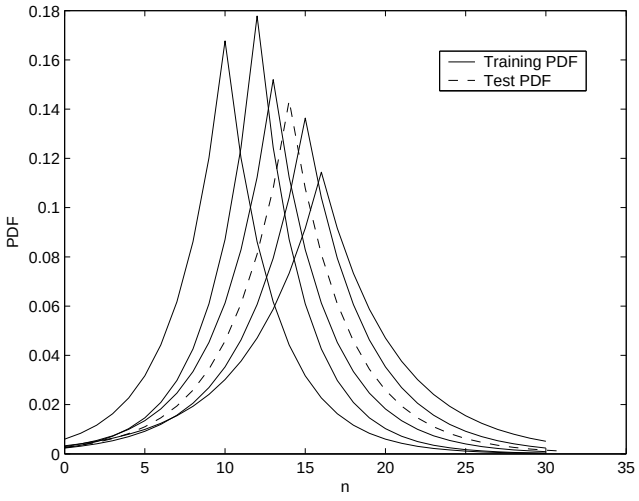

Figure 17: Discrete Laplacian PDFs that are used for optimizing $f(n)$ and testing the performance.

zeros of $[\widetilde{F}(z)F(z)]_{\downarrow M}$ are not too close to the unit circle. If this condition is not satisfied, the least squares partner $G(z)$ will have poles that are close to the unit circle (or even on the unit circle), in which case the FIR truncation approach is no more practical. We may also impose other constraints such as $f(0) \neq 0$ and $\sum_n f(n) = 1$ in order to make the optimal solution unique.

7.2 Optimization Results for Commonly Used PDFs

In order to demonstrate the idea, we now optimize the coefficients of $f(n)$ for commonly used probability density functions such as the Laplacian and the Poisson PDF. We assume that a set of density functions that belong to the same class - based on some criterion, which is not necessarily known - are given. This set of density functions will serve as the training set for optimizing $f(n)$. Next, another PDF that belongs to the same group but was not included in the training set will be chosen in order to test the performance of the model-based estimator.

Let us first consider a set of Laplacian PDFs. The continuous Laplacian density function is defined as

$$p(x) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}|x-\mu|/\sigma}. \tag{64}$$

We may obtain a discrete PDF $x(n)$ by sampling and truncating the above function, and then normalizing it in order to make the coefficients add up to unity. We obtained several discrete PDFs from continuous Laplacian density functions with different mean and variance. These are shown in Fig. 17. Five density functions were used as the training set for optimizing $f(n)$, and the remaining one was used for evaluating the performance of the model-based PDF estimation.

***Choice of* M** When computing the optimal $f(n)$ based on this training set, we assumed $M = 2$ and the length of $f(n)$ was restricted to be ≤ 5 for simplicity. In general, a larger M results

in increasing the minimum value of the weighted projection error J. However, if the PDFs of our interest can be well represented by the model in Fig. 2 with the given M and the optimized $f(n)$, a larger M can result in a smaller estimation error. Therefore, there is a trade-off in choosing the value of M, and it should be chosen according to the given training set. Once M is decided, choosing the length of the filter $f(n)$ to be around $2M+1$ usually yields a satisfactory estimation result. After the optimization process, if the error (63) is too large compared to $\sum_{i=0}^{I-1} \alpha_i \|x_i(n)\|^2$, we may have to decrease M as well as allow the filter $f(n)$ to have a longer length.

In our simulations, we used a *sequential quadratic programming (SQP)* method [24] for optimizing the filter coefficients. Once we obtained the optimal $f(n)$, the corresponding least squares partner $g(n)$ was computed as in (13). Next, we took the test PDF in Fig. 17 to make 300 measurements of the r.v. n and constructed the histogram based on these observations. Finally, this histogram was passed through the system in Fig. 5 to get the PDF estimate $\hat{x}(n)$. The result is shown in Fig. 18. This figure shows that the model-based approach results in an excellent PDF estimate, whereas the histogram estimate is not very satisfactory. The estimation errors were

$$\sum_n |h(n) - x(n)|^2 = 0.00260485 \quad \text{and} \quad \sum_n |\hat{x}(n) - x(n)|^2 = 0.00048105 \tag{65}$$

where the error of the model-based estimate was only 18.47% of the error of the histogram based estimate. There is another important point that has to be noted. Although the model-based estimation procedure reduces the estimation error by a considerable amount, it does not compromise the sharpness of the original PDF, as can be seen in Fig. 18. This is indeed another advantage of

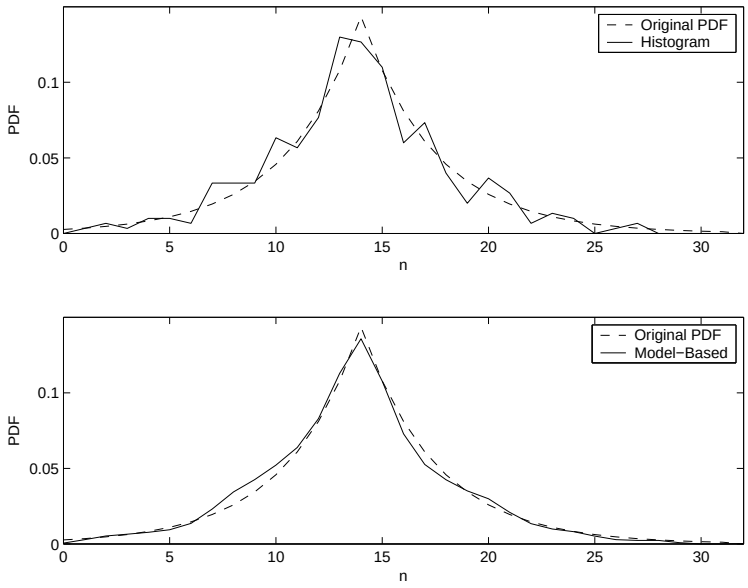

Figure 18: Top plot: the original Laplacian PDF and the histogram. Bottom plot: the original PDF and the model-based estimate.

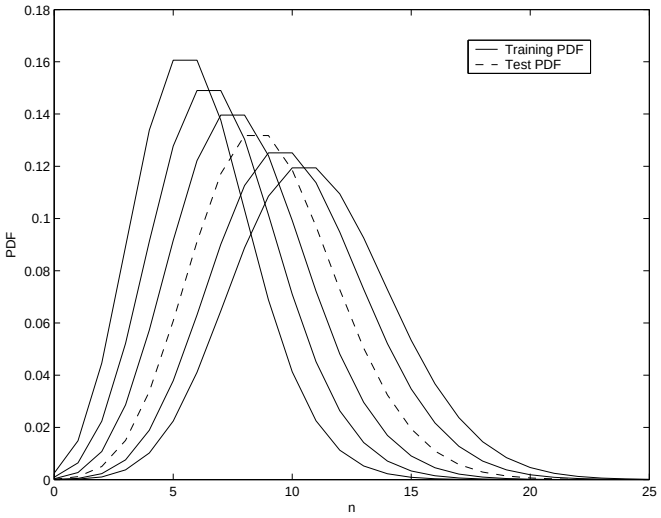

Figure 19: Poisson PDFs used for optimizing $f(n)$ and testing the performance.

	Average Estimation Error	
PDF	Histogram	Model-Based
Laplacian	0.00289717	0.00115754
Gaussian	0.00304934	0.00103529
Rayleigh	0.00298871	0.00119290
Poisson	0.00292003	0.00094862

Table 1: Estimation results using optimized $f(n)$ for various class of PDFs.

the model-based approach, making it more attractive compared to the traditional histogram based density estimation.

A similar experiment has been performed based on a set of Poisson PDFs. The training PDFs and the test PDF are shown in Fig. 19. Figure 20 shows the histogram and the model-based estimate. The estimation errors were

$$\sum_n |h(n) - x(n)|^2 = 0.00170609 \quad \text{and} \quad \sum_n |\hat{x}(n) - x(n)|^2 = 0.00024454 \tag{66}$$

hence the estimation error of the model-based estimate was only 14.33% when compared to that of the histogram estimate.

Table 1 summarizes the estimation results for several class of density functions. The filter coefficients $f(n)$ were optimized for each class of PDFs, and we estimated the test PDFs using this $f(n)$. The estimation errors have been averaged for 10 simulations. This result shows that the optimization of $f(n)$ yields satisfactory PDF estimates that are significantly better than the histograms.

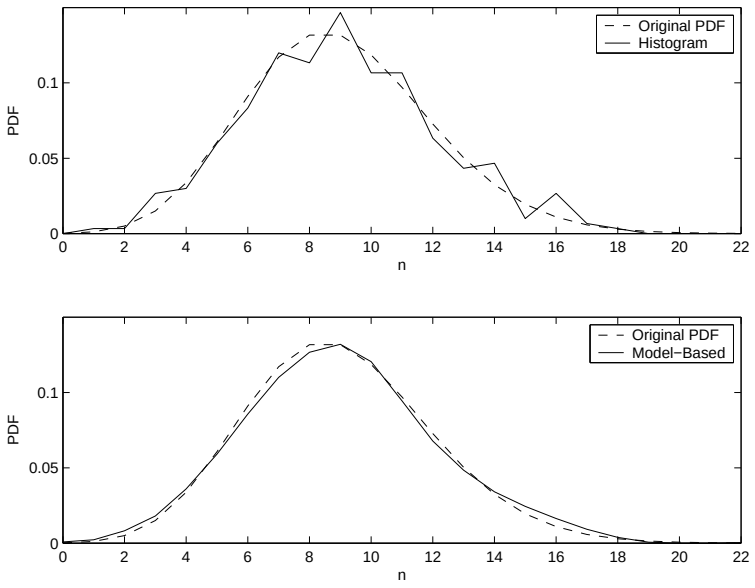

Figure 20: Top plot: the original Poisson PDF and the histogram. Bottom plot: the original PDF and the model-based estimate.

8 Concluding Remarks

Although many researchers in the mathematics as well as the signal processing communities have been interested in estimating the probability density function of a random variable v, most of the results were focused on the continuous case. In this paper, we proposed a new method for the estimation of discrete probability density functions. The proposed method is based on multirate signal processing theory, and it makes extensive use of the notion of biorthogonal partners. This model-based approach has several advantages over the traditional histogram approach. It guarantees a smaller estimation error and it provides an efficient denoising scheme when the observations are corrupted by additive noise. Furthermore, the analysis of the model-based estimates shows that they are unbiased and have a smaller variance than the histogram estimates. It was also shown that the proposed model can be used for modeling and estimating multivariate probability density functions. Finally, it was demonstrated that the interpolation filter $f(n)$ could be optimized for a class of density functions. Simulation results show that the use of an optimized filter can decrease the estimation error dramatically. There are many interesting extensions of the proposed ideas, which remain to be considered. For example, we may consider modeling and estimating multivariate probability density functions with a non-separable filter $F(\mathbf{z})$ and a non-diagonal matrix $\mathbf{M}$. This will require the concept of multi-dimensional biorthogonal partners. Another interesting problem is the estimation of a PDF when the noise is dependent on the original samples. These are topics for future research.

9 Appendix

In this appendix, we provide the detailed derivation of (41). Let us first compute $\mathcal{E}[t(k_1)t(k_2)]$. If $k_1 = k_2$, then

$$\begin{aligned}
\mathcal{E}[t(k_1)t(k_2)] &= \mathcal{E}[t^2(k_1)] \\
&= \mathrm{Var}[t(k_1)] + \Big(\mathcal{E}[t(k_1)]\Big)^2 \\
&= Nx(k_1)\Big(1 - x(k_1)\Big) + N^2x^2(k_1) \\
&= Nx_1\Big(1 + (N-1)x(k_1)\Big) \qquad (67)
\end{aligned}$$

Now, if $k_1 \neq k_2$, then

$$\begin{aligned}
\mathcal{E}[t(k_1)t(k_2)] &= \sum_{m=0}^{N} P\{t(k_1) = m\} \cdot m \cdot \mathcal{E}[t(k_2)|t(k_1) = m] \\
&= \sum_{m=0}^{N} P\{t(k_1) = m\} \cdot m \cdot (N-m)\frac{x(k_2)}{1 - x(k_1)} \\
&= \frac{x(k_2)}{1 - x(k_1)}\left(N \cdot \mathcal{E}[t(k_1)] - \Big[\mathrm{Var}[t(k_1)] + \Big(\mathcal{E}[t(k_1)]\Big)^2\Big]\right) \\
&= \frac{x(k_2)}{1 - x(k_1)}\left(N^2x(k_1) - \Big[Nx(k_1)\Big(1 - x(k_1)\Big) + N^2x^2(k_1)\Big]\right) \\
&= N(N-1)x(k_1)x(k_2) \qquad (68)
\end{aligned}$$

Since $\mathcal{E}[h(k_1)h(k_2)] = \mathcal{E}[t(k_1)t(k_2)]/N^2$, we obtain (41) by dividing (67) and (68) by N^2.

References

[1] C. N. Morris, "Natural exponential families with quadratic variance functions: statistical theory", *Ann. Stat.*, vol. 11, pp. 515-529, 1983.

[2] G. G. Walter and G. G. Hamedani, "Bayes empirical bayes estimation for natural exponential families with quadratic variance functions", *Ann. Stat.*, vol 19, pp. 1191-1224, 1991.

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

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

[5] B. W. Silverman, "Density estimation for statistics and data analysis", *Monographs on Statistics and Applied Probability*, vol. 26, Chapman and Hall, London, 1986.

[6] C. Chen, W. A. Fuller and F. J. Breidt, "Spline estimators of the density function of a variable measure with error", *Statistics - Simulation and Computation*, vol. 32, no. 3, pp. 73-86, Jan. 2003.

[7] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parametric estimation for Gaussian mixture and hidden Markov models", *Technical Report ICSI-TR-97-021*, University of Berkeley, 1997.

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

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

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

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

[12] B. Vidakovic, *Statistical modeling by wavelets*, John Wiley & Sons, Inc., NY, 1999.

[13] L. Couvreur and C. Couvreur, "Wavelet-based method for nonparametric estimation of HMMs", *IEEE Signal Processing Letters*, vol. 7, pp. 25-27, Feb. 2000.

[14] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities", *Biometrika*, 58:255-277, 1971.

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

[16] B.-J. 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.

[17] B. Vrcelj and P. P. Vaidyanathan, "Efficient implementation of all-digital interpolation", *IEEE Trans. Image Processing*, vol. 10(11), pp. 1639-46, Nov. 2001.

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

[19] C. W. Therrien, *Discrete random signals and statistical signal processing*, Prentice Hall, Inc., NJ, 1992.

[20] B.-J. 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.

[21] A. Tkacenko and P. P. Vaidyanathan, "On the least squares signal approximation model for overdecimated rational nonuniform filter banks and applications", *Proc. 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, Apr. 2003.

[22] A. Kiraç and P. P. Vaidyanathan, "Theory and design of optimum FIR compaction filters", *IEEE Trans. Signal Processing*, vol. 46(4), pp. 903-919, Apr. 1998.

[23] A. N. Akansu, and Y. Liu, "On signal decomposition techniques", *Opt. Eng.*, vol. 30, pp. 912-920, July 1991.

[24] P. E. Gill, W. Murray, and M. H. Wright, *Practical optimization*, Academic Press, London, 1981.

Byung-Jun Yoon (S'02) was born in Seoul, Korea in 1975. He received the B.S.E. (summa cum laude) degree from Seoul National University (SNU), Seoul, Korea in 1998 and the M.S. degree from California Institute of Technology (Caltech), Pasadena, in 2002, both in electrical engineering. He is currently pursuing the Ph.D. degree in electrical engineering at Caltech. His research interests include filter bank theory, multirate signal processing and applications, wavelets, signal regularization, Bioinformatics and Genomic signal processing. He received the Killgore fellowship in 2001 from Caltech, and he was selected as a Microsoft Research fellow for the year of 2004-2005. In 2003, he was awarded a prize in the student paper contest in the 37th Asilomar conference on

signals, systems, and computers.

P. P. Vaidyanathan (S'80–M'83–SM'88–F'91) was born in Calcutta, India on Oct. 16, 1954. He received the B.Sc. (Hons.) degree in physics and the B.Tech. and M.Tech. degrees in radiophysics and electronics, all from the University of Calcutta, India, in 1974, 1977 and 1979, respectively, and the Ph.D degree in electrical and computer engineering from the University of California at Santa Barbara in 1982. He was a post doctoral fellow at the University of California, Santa Barbara from Sept. 1982 to March 1983. In March 1983 he joined the electrical engineering department of the Calfornia Institute of Technology as an Assistant Professor, and since 1993 has been Professor of electrical engineering there. His main research interests are in digital signal processing, multirate systems, wavelet transforms and signal processing for digital communications.

Dr. Vaidyanathan served as Vice-Chairman of the Technical Program committee for the 1983 IEEE International symposium on Circuits and Systems, and as the Technical Program Chairman for the 1992 IEEE International symposium on Circuits and Systems. He was an Associate editor for the IEEE Transactions on Circuits and Systems for the period 1985-1987, and is currently an associate editor for the journal IEEE Signal Processing letters, and a consulting editor for the journal Applied and computational harmonic analysis. He has been a guest editor in 1998 for special issues of the IEEE Trans. on Signal Processing and the IEEE Trans. on Circuits and Systems II, on the topics of filter banks, wavelets and subband coders. Dr. Vaidyanathan has authored a number of papers in IEEE journals, and is the author of the book Multirate systems and filter banks. He has written several chapters for various signal processing handbooks. He was a recepient of the Award for excellence in teaching at the California Institute of Technology for the years 1983-1984, 1992-93 and 1993-94. He also received the NSF's Presidential Young Investigator award in 1986. In 1989 he received the IEEE ASSP Senior Award for his paper on multirate perfect-reconstruction filter banks. In 1990 he was recepient of the S. K. Mitra Memorial Award from the Institute of Electronics and Telecommuncations Engineers, India, for his joint paper in the IETE journal. He was also the coauthor of a paper on linear-phase perfect reconstruction filter banks in the IEEE SP Transactions, for which the first author (Truong Nguyen) received the Young outstanding author award in 1993. Dr. Vaidyanathan was elected Fellow of the IEEE in 1991. He received the 1995 F. E. Terman Award of the American Society for Engineering Education, sponsored by Hewlett Packard Co., for his contributions to engineering education, especially the book Multirate systems and filter banks published by Prentice Hall in 1993. He has given several plenary talks including at the Sampta'01, Eusipco'98, SPCOM'95, and Asilomar'88 conferences on signal processing. He has been chosen a distinguished lecturer for the IEEE Signal Processing Society for the year 1996-97. In 1999 he was chosen to receive the IEEE CAS Society's Golden Jubilee Medal. He is a recepient of the IEEE Signal Processing Society's Technical Achievement Award for the year 2002.

List of Figures

List of Tables