

Fast Density Estimation from Histograms in Shift Invariant Spaces

Harald Schwab*

NUHAG, Department of Mathematics, University of Vienna,
Nordbergstraße 15, A-1090 Vienna, AUSTRIA
harald.schwab@univie.ac.at

Abstract

Histograms are mostly used as data presentation. In many applications we are interested in a good approximation of the density function, which creates the histograms. Standard techniques like the kernel density estimation are applied for approximating the density function. The problem of these techniques is that the data which define the histogram need to be known apriori. To avoid this problem we present an algorithm, which reconstructs the density function only from the given histogram (i.e., the width and the height of the bins are used as input) and without knowledge about the specific measurements. This becomes possible because we use techniques for *reconstruction from averages*. Using the fast efficient algorithm presented by Gröchenig and Schwab [7] it is shown in this paper, that this reconstruction scheme can be used for the case of averaging and provides good results for the approximation of the density function from a given histogram.

Key words: Density estimation, shift-invariant space, nonuniform sampling, banded matrix, localization, data segmentation.

2000 AMS Mathematics Subject Classification: 62G07, 41A15, 42C15, 46A35, 46E15, 46N99, 47B37

1 Introduction

Histograms can be viewed as a particular example of a density estimate and their appearance depends on both the choice of origin of the histogram and the width of the intervals used. Our goal is to find a good

*The author acknowledges partial support by the Austrian Science Fund project FWF P-14485.

approximation of the underlying density function. For some concrete applications of this problem see eg. [3], [2], [4], [8] or [9].

There are many different techniques that solve this problem. For example, the *kernel density estimation (KDE)* is a well-known technique in statistics. The simplest way of KDE can be thought as a smoothed version of histograms. The main problem in practice is to obtain a sufficiently smooth representation of the data while at the same time retaining its main features. In this context the choice of bin-width is critical.

In Section 2 we describe the standard technique of KDE and refer to Baxter and Beardah [3] who present a solution of the under/over smoothing problem.

These KDE techniques require an explicit knowledge of the observations that produce the histogram. Our goal is to avoid this assumption and to reconstruct the density function directly from the histogram. We can look at this problem as a kind of resolution enhancement, i.e., if multiple sensors take measurements over disjoint intervals we only get sums, respectively averages over these intervals and our intention is to get a resolution of arbitrarily small intervals.

The critical difference in this approach is that we do not use the histogram as a *presentation* of data but rather as *input* for the reconstruction of the density function.

2 Kernel Density Estimation

Let X_i , $i = 1, \dots, N$, be scalar measurements drawn from an arbitrary probability distribution f . One well-known approach (the kernel density estimation KDE, see also [10]) of finding an approximation \hat{f} for the unknown density function f is obtained based on a kernel function $K(u)$ and a bandwidth h as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right). \quad (1)$$

The kernel function K satisfies the following properties

- (i) $K(u) = K(-u) \geq 0$
- (ii) $K(u) = 0$ for $|u| > 1$
- (iii) $K(0) \geq K(u)$ for $u \neq 0$
- (iv) $\int_{-1}^1 K(u) = 1$

This method can be viewed as placing a ‘bump’ at each point and then summing the height of each bump at each point of the x-axis.

Compared to the histogram, the shape of \hat{f} does not depend upon the choice of origin but critical on the choice of bandwidth h . Large values of h over-smooth, while small values under-smooth the data.

In [3] Baxter and Beardah present a choice of h , which ‘optimizes’ the kernel density estimation in the sense that the calculated h makes the KDE as ‘close’ as possible. The measurement of ‘closeness’ is the asymptotic mean integrated square error (AMISE) which can be shown to have the form

$$AMISE(\hat{f}) = \frac{1}{nh}A + \frac{1}{4}h^4B. \quad (2)$$

As described in [3] the terms A and B in equation (2) are dependent on the known kernel, while B is also dependent on the integral of the squared second derivative of the unknown f . Then the optimal value of h which minimizes the AMISE is given by

$$h_{AMISE} = \left[\frac{A}{nB} \right]^{1/5}. \quad (3)$$

This expression, which through B depends upon the second derivative of the unknown density f , is the starting point for many methods for automatic selection of h .

Here in this paper we will not go into further details of selecting the optimal h , which can be found for example in [12] or [3].

We want to focus on another problem:

Following the technique described above an apriori knowledge of the points X_i is assumed. In that case a histogram is only used as a method of data presentation. Now we want to concentrate on the problem that the histogram is the only input we have, i.e., the histogram does not appear as a data representation but as data itself. *Therefore our goal is*

to find a good approximation of the density function from the histogram itself without using the data/observations X_i . This leads us to the next section.

3 Reconstruction of the Density from Histograms

As mentioned above the input of our reconstruction algorithm is not the observations X_i because they may be unknown, but the histogram, i.e., the position, the width and the height of the bins.

The problem can be modeled as follows: Let be given n intervals of same length, i.e., for $i = 1, \dots, n$ and fixed $d > 0$:

$$\begin{aligned} I_i &= [a_i, a_{i+1}), \\ a_{i+1} &= a_i + d. \end{aligned}$$

Then we can describe a histogram by a step-function f_H

$$f_H(x) = \begin{cases} w_i, & \text{for } x \in [a_i, a_{i+1}) \\ 0, & \text{for } x < a_1 \text{ or } x \geq a_n \end{cases} \quad (4)$$

with $\int_{\mathbb{R}} f_H(x) dx = 1$ to get a normalized histogram.

The problem we want to handle is to get an approximation \hat{f} of the unknown density f from the given function f_H .

The first step is to define the samples for the reconstruction. Let the function u be defined as follows:

$$u(x) = \begin{cases} 1/d, & \text{for } x \in [-\frac{d}{2}, +\frac{d}{2}] \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Considering the convolution

$$\begin{aligned} (f * u)(x) &= \int_{\mathbb{R}} f(\xi) u(\xi - x) d\xi = \int_{x-d/2}^{x+d/2} f(\xi) u(\xi - x) d\xi \\ &= \int_{x-d/2}^{x+d/2} f(\xi) \frac{1}{d} d\xi \\ &= \frac{1}{d} \int_{x-d/2}^{x+d/2} f(\xi) d\xi. \end{aligned} \quad (6)$$

It is obvious that

$$(f * u)(x_0) = \frac{1}{d} \int_{x_0-d/2}^{x_0+d/2} f(\xi) d\xi \tag{7}$$

is the average of the function values of the interval $[x_0 - \frac{1}{d}, x_0 + \frac{1}{d}]$.

Since the bins of a histogram are normalized by their width d , the following equation holds for $i = 1, \dots, n$:

$$(f * u)(x_i) = f_H(x_i) \tag{8}$$

with

$$x_i = \frac{a_i + a_{i+1}}{2} \tag{9}$$

the midpoints of each bin.

3.1 Reconstruction with parabolic splines

In [5] Carl de Boor gives a spline interpolation for this problem:

Let be \hat{f} to be a parabolic spline, i.e., a piecewise polynomial function of order 3 with continuous first derivative,

$$\hat{f} \in \mathbb{P}_{3,\zeta} \cap C^1, \tag{10}$$

where $\mathbb{P}_{3,\zeta}$ is the linear space of piecewise polynomial functions of order 3 with breakpoint sequence ζ , which coincide with the sequence $a_i, \quad i = 1, \dots, n$.

If the underlying density function f is smooth and vanishes outside the interval $[a_1, a_{n+1}]$, then we have $f^{(j)}(a_1) = f^{(j)}(a_{n+1}) = 0$ for $j = 0, 1, \dots$ to the extent of the smoothness of f .

This gives altogether $n + 1$ interpolation conditions and $2n -$ homogeneous conditions, for a total of $3n$ conditions on the $3n$ polynomial coefficients. For the solution of the resulting linear system the reader is referred to [5].

3.2 Reconstruction with shift invariant functions

In this paper we want to present a more general solution for the density reconstruction from a histogram, i.e., we choose the approximation \hat{f}

for the unknown function f from the space of shift invariant functions $\hat{f} \in V(\varphi)$, which is defined by

$$V(\varphi) = \{g \in L^2(\mathbb{R}) : g(x) = \sum_{k \in \mathbb{Z}} c_k \varphi(x - k) \text{ for } (c_k) \in \ell^2(\mathbb{Z})\}, \quad (11)$$

where φ , the generator, is a continuous function with compact support of size S so that

$$\text{supp}(\varphi) \subseteq [-S, S]. \quad (12)$$

As mentioned above, the reconstruction of a density function from a histogram can be ascribed to the problem of *reconstruction from averages*. During the last years there were several publications on this topic. For example, Sun and Zhou presented some results in [11] for the bandlimited case. In this paper we concentrate on a fast local reconstruction method for sampling in shift invariant spaces formulated in [7] by K.Gröchenig and H.Schwab. Now we will show that this efficient algorithm can be applied to the case of reconstruction from averages.

As mentioned above, we assume that the unknown density f is element of the space of shift invariant functions, i.e.,

$$f(x) = \sum_{k \in \mathbb{Z}} c_k \varphi(x - k), \quad (13)$$

which means that the function f is completely determined by the coefficients c_k or, in other words, the problem of finding the exact reconstruction f is equivalent to find the coefficients c_k .

Applying the convolution with u , we have

$$(f * u)(x) = \sum_{k \in \mathbb{Z}} c_k (\varphi * u)(x - k), \quad (14)$$

where obviously $(f * u)$ is determined by the same coefficients as f and is generated by $\varphi * u$, i.e.,

$$(f * u) \in V(\varphi * u). \quad (15)$$

From equation (8) we see that the midpoints of the bins $x_i = (a_i + a_{i+1})/2$ can be used as samples of the function $(f * u)$.

As we will see in the next chapter, we calculate the coefficients c_k from these samples of $(f * u)$ and then reconstruct the density function f with these coefficients.

This technique involves an important interpretation because we do not reconstruct the desired density f from samples but from the *averages* $(f * u)(x_i)$, $i = 1, \dots, n$.

3.2.1 Calculating the Coefficients of density function f

Here we follow the description of reconstruction methods in shift invariant spaces presented in [7]. To keep the notation clear we define

$$\psi = \varphi * u. \tag{16}$$

Then the averages, which we use as samples for the function $(f * u)$, can be written as

$$(f * u)(x_i) = \sum_{k=\lceil a_1 - S - d/2 \rceil}^{\lfloor a_n + S + d/2 \rfloor} c_k \psi(x_i - k) \tag{17}$$

for

$$x_i = \frac{a_i + a_{i+1}}{2} \quad i = 1, \dots, n. \tag{18}$$

Let U be the matrix with entries

$$U_{ik} = \psi(x_i - k) \tag{19}$$

with $i = 1, \dots, n$ and $k \in (a_1 - S - d/2, a_{n+1} + S + d/2) \cap \mathbb{Z}$. Then with $(f * u)|_X = ((f * u)(x_i))_{i=1, \dots, n}$ the equation (17) can be rewritten as

$$Uc = (f * u)|_X \tag{20}$$

or in the sense of normal equations

$$U^*Uc = U^*(f * u)|_X. \tag{21}$$

As described in [7] the structure of this problem must be exploited to receive a fast reconstruction algorithm:

Lemma 1 ([7]). *If $\text{supp}(\psi) \subseteq [-S - \frac{d}{2}, S + \frac{d}{2}]$, then $T = U^*U$ is a band matrix of (upper and lower) band-width $2S + d$.*

Let be $T = U^*U$, given by the entries

$$T_{kl} = \sum_{i=1}^n \overline{\psi(x_i - k)} \psi(x_i - l) \quad (22)$$

for $i = 1, \dots, n$ and $k, l = \lceil a_1 - S - \frac{d}{2} \rceil, \dots, \lfloor a_n + S + \frac{d}{2} \rfloor$. The next step is to compute $b = U^*(f * u)|_X$, i.e.,

$$b_k = \sum_{i=1}^n \overline{\psi(x_i - k)} (f * u)(x_i) \quad \text{for } k = \lceil a_1 - S - \frac{d}{2} \rceil, \dots, \lfloor a_n + S + \frac{d}{2} \rfloor \quad (23)$$

so that we have to solve the linear system

$$Tc = b. \quad (24)$$

Since T is a positive definite band matrix, the *band Cholesky algorithm* is an effective method for solving (24) because the operation count is *linear* to the number of samples n (see also [6]).

After we evaluate the coefficients c from the function $(f * u)$ we can use them to construct the restriction of f to $[a_1, a_n]$ by

$$f(x) = \sum_{k=\lceil a_1 - S \rceil}^{\lfloor a_n + S \rfloor} c_k \varphi(x - k) \quad \text{for } x \in [a_1, a_n]. \quad (25)$$

Once again: *The coefficients of the function $(f * u)$, which reconstruct the averages (=midpoints of the bins), are the same as the coefficients of the density function f .*

As described in [7] the operator count of this algorithm is

$$\mathcal{O}(n \cdot C) \quad (26)$$

where C depends quadratic on the length of the generator ψ . In other words, the costs of the algorithm are linear to the number of data (=bins of the histogram).

For a detailed description see [7].

3.2.2 Initialization of the Histogram

For exact reconstruction the set of samples, which are the midpoints of the bins, needs to fulfill certain conditions. If φ is a B-spline of order N , i.e., $\varphi = \chi_{[0,1]} * \dots * \chi_{[0,1]}$ ($N + 1$ convolutions), then the main result of [1] implies that the maximum gap condition

$$\sup_{i \in \mathbb{Z}} (x_{i+1} - x_i) = \delta < 1 \tag{27}$$

is sufficient for exact reconstruction. To hold this condition we have to factorize the histogram, i.e., we have to choose a $\lambda > 0$ such that

$$\begin{aligned} \hat{x}_i &= \frac{a_i + a_{i+1}}{2\lambda}, \\ \sup_i (\hat{x}_{i+1} - \hat{x}_i) &= \delta < 1, \\ \text{for } i &= 1, \dots, n - 1 \end{aligned} \tag{28}$$

and the step function in (4) becomes

$$\hat{f}_H(x) = \begin{cases} w_i, & \text{for } x \in [\frac{a_i}{\lambda}, \frac{a_{i+1}}{\lambda}) \\ 0, & \text{for } x < \frac{a_1}{\lambda} \text{ or } x \geq \frac{a_n}{\lambda}. \end{cases} \tag{29}$$

Consequently the length of the modified histogram is

$$\hat{d} = \frac{a_{i+1} - a_i}{\lambda} \quad \text{for } i = 1 \dots n. \tag{30}$$

With these modified data we can reconstruct from the samples

$$(\hat{f} * u)(\hat{x}_i) = \hat{f}_H(\hat{x}_i) \tag{31}$$

the desired density function by

$$f(x) = \hat{f}(x \cdot \lambda). \tag{32}$$

3.2.3 Averages created by different average-functions u_j

As described above we have examined the situation, that the averages are always taken with respect to the same window u .

Now let us study the more general situation that we are taking averages over areas of different width, i.e.,

$$u_j(x) = \begin{cases} 1/\mu_j, & \text{for } x \in [-\frac{\mu_j}{2}, \frac{\mu_j}{2}] \\ 0, & \text{otherwise} \end{cases} \quad (33)$$

Figure 6 shows this situation. We are using the same density function, but the averages are taken over areas of different length.

Consequently the sampling values w_j are given by

$$w_j = (f * u_j)(x_j) = \frac{1}{\mu_j} \int_{x_j - \mu_j/2}^{x_j + \mu_j/2} f(\xi) d\xi. \quad (34)$$

Therefore we need to modify the algorithm described before in the following way:

1. For each sample-value w_j , we must calculate the appropriate average-function u_j , $j = 1, \dots, J$. Let be $f \in V(\varphi)$, then

$$\psi_j = \varphi * u_j. \quad (35)$$

2. We must store these ψ_j to compute $b = U^* w$

$$b_k = \sum_{j=1}^J \overline{\psi_j(x_j - k)} w_j. \quad (36)$$

3. Also the calculation of the matrix T must be modified:

$$T_{kl} = \sum_{j=1}^J \overline{\psi_j(x_j - k)} \psi_j(x_j - l). \quad (37)$$

A closer look at the system matrix T verifies that the matrix T does not have the same structure as before.

Since the support of ψ_j depends on the width of the bins, i.e.,

$$\text{supp}(\psi_j) \subseteq [-S - \frac{\mu_j}{2}, S + \frac{\mu_j}{2}] \quad (38)$$

the width of the band in the matrix T varies, too (see figure 7).

Applying the modifications described above, we get the following approximation of the density function as plotted in figure 9.

4 Numerical Simulations

We tested this algorithm on several different data sets. We used also data of large dimension, i.e. a histogram with 10^4 bins. By handling this quantity of data two features of the proposed algorithm are of great importance:

1. The number of operations is linear to the number of bins.

As mentioned above the calculation of the coefficients of the unknown density function requires $\mathcal{O}(n \cdot C)$ multiplications, where $n = 10^4$ and C depends quadratic on the length of $\text{supp}(\psi)$.

2. The algorithm uses data segmentation for the reconstruction.

Since we have a local reconstruction method we can handle non-adjacent intervals separately. Therefore we can separate the reconstruction in several intervals depending on numerical limitations e.g. memory capacity.

Using this data segmentation we approximate the density function from 10^4 given average-values. After calculating the coefficients we evaluate the density function at 10^7 values, which can be viewed as an improvement of the resolution of the factor 10^3 . The error of the approximation \hat{f} measured by

$$\text{err}(\hat{f}) = \frac{\|f - \hat{f}\|_2}{\|f\|_2}$$

is 1.12%.

Figure 1 shows a part of the reconstruction and the given histogram.

To see more details we also want to present an example of small size: Figure 2 shows the histogram with the underlying density function. We want to find a good approximation for the density function from the given histogram by using approximations from the space of shift invariant functions.

In our simulation we use MATLAB. We use the shift-invariant spline spaces with the B -spline of order 3

$$\varphi = \underbrace{\chi_{[-1/2, 1/2]} * \dots * \chi_{[-1/2, 1/2]}}_{4 \text{ times}} \quad (39)$$

$$4 \text{ times} \quad (40)$$

as the generator of $V(\varphi)$. Thus $\text{supp}(\varphi) \subseteq [-2, 2]$ and $S = 2$ (see Figure 3).

As we can see in Figure 4, the algorithm yields to a good approximation of the original density function.

In Figure 5 the difference between approximation and original density function is plotted.

In Figure 6 we see the exact reconstruction of the midpoints of the bins x_i by the function $(\hat{f} * u)$, i.e., for $i = 1, \dots, n$

$$(\hat{f} * u)(x_i) = w_i \tag{41}$$

and therefore

$$\int_{a_i}^{a_{i+1}} (\hat{f} * u)(x) dx = w_i \cdot d, \tag{42}$$

which means that the area under the function $(\hat{f} * u)$ is the same as the area of the bins. The function $(\hat{f} * u)$ can be viewed as moving average.

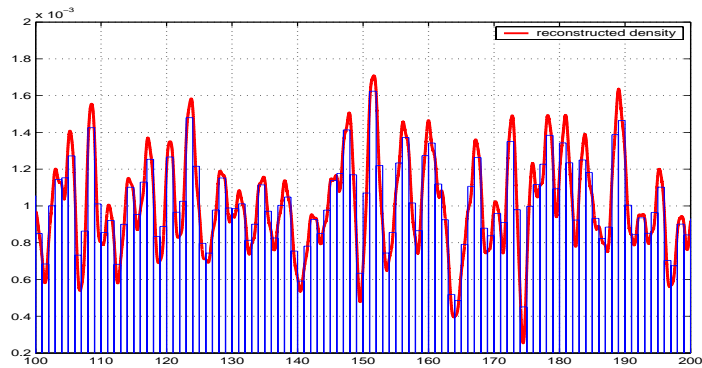


Figure 1: Histogram with reconstruction.

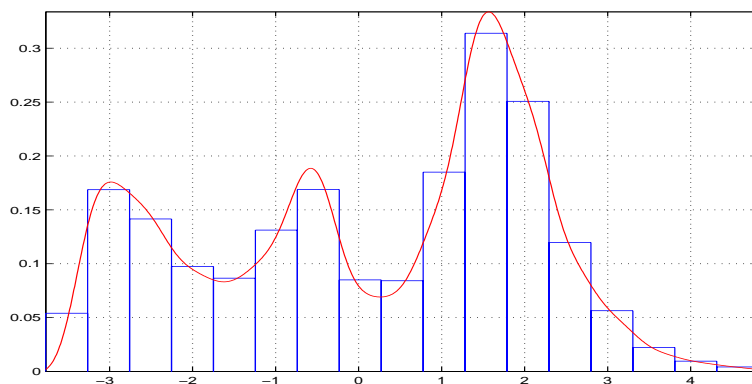


Figure 2: Histogram with underlying density function.

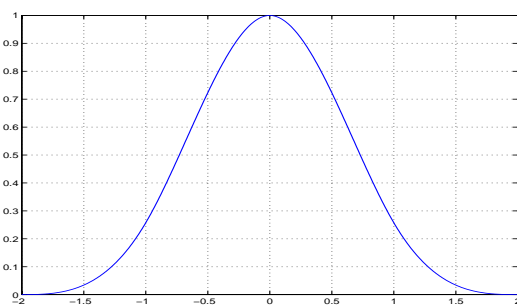


Figure 3: Generator φ : B-spline of order 3 with $\text{supp}(\varphi) \subseteq [-2, 2]$.

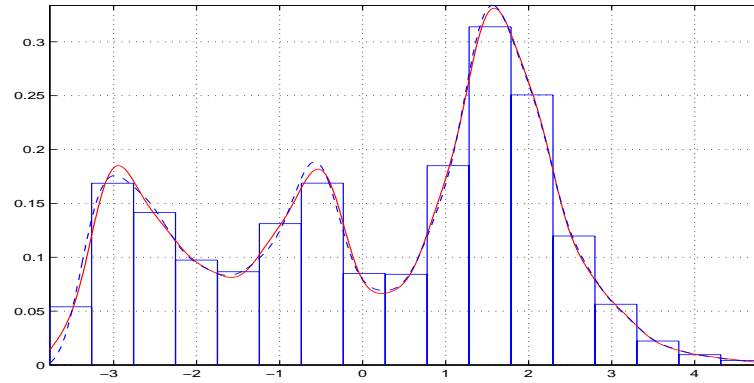


Figure 4: Reconstruction (solid line) and the original density function (dotted line).

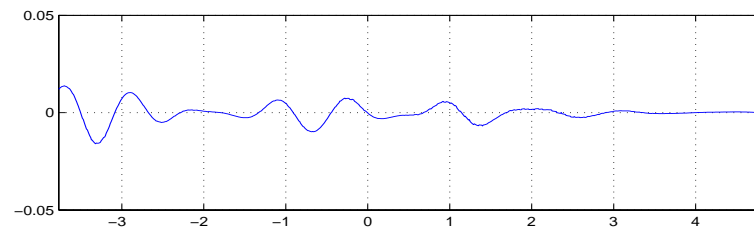


Figure 5: Error: Reconstructed versus original density function.

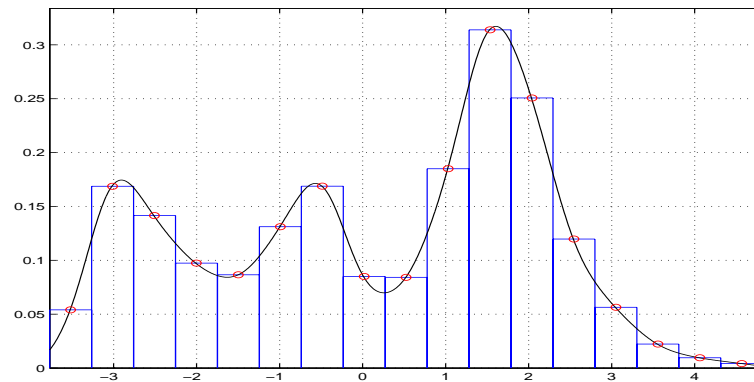


Figure 6: The histogram with marked midpoints, which are used as samples for the function $(f * u)$. Solid line: Reconstructed function $(\hat{f} * u)$.

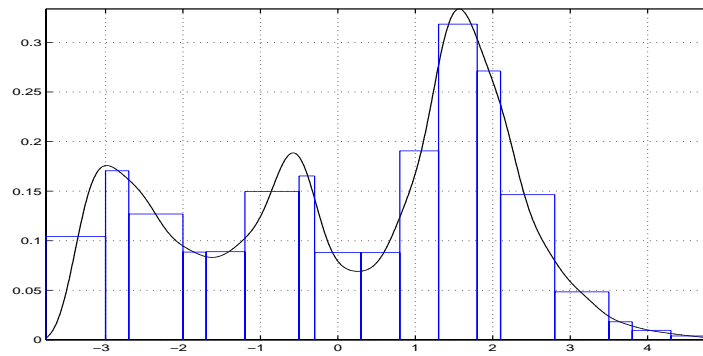


Figure 7: The density function and the histogram with different bin-width.

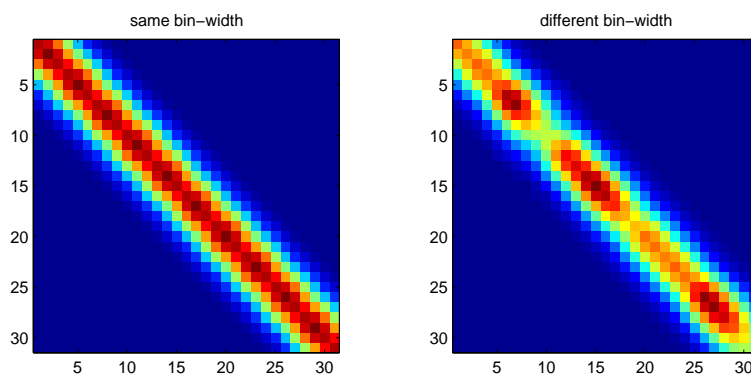


Figure 8: The system matrix T in two different situations.

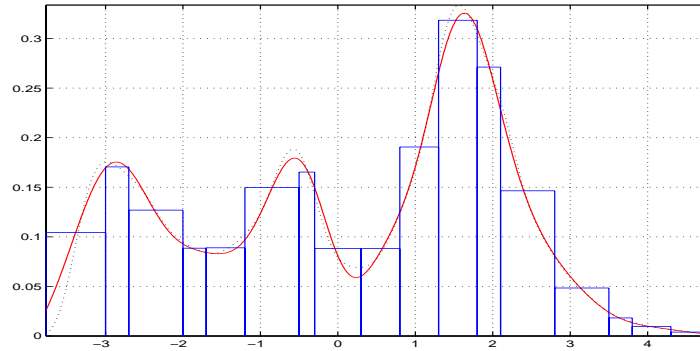


Figure 9: Reconstruction (solid line) and the original density function (dotted line).

References

- [1] A. Aldroubi and K. Gröchenig. Beurling-Landau-type theorems for non-uniform sampling in shift invariant spline spaces. *J. Fourier Anal. Appl.*, 6(1):93–103, 2000.
- [2] G.J. Babu, A. J. Canty, and Y.P. Chaubey. Application of bernstein polynomials for smooth estimation of a distribution and density function. *J. Stat. Plann. Inference* 105, 2:377–392, 2002.
- [3] M.J. Baxter and C.C. Beardah. Beyond the histogram - improved approaches to simple data display in archaeology using kernel density estimates. *Archeologia e Calcolatori 7, Proceedings of the 3rd International Symposium on Computing and Archaeology*, pages 397–408, 1996.
- [4] H. Chen and P. Meer. Robust computer vision through kernel density estimation. *7th European Conference on Computer Vision*, I:236–250, 2002.
- [5] C. De Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. 27. New York, NY: Springer., New Yourk, 1978.
- [6] G.H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

- [7] K. Gröchenig and H. Schwab. Fast local reconstruction methods for nonuniform sampling in shift-invariant spaces. *SIAM Journal on Matrix Analysis and Applications*, 24(4):899–913, 2003.
- [8] P. Hall and M.P. Wand. On the accuracy of binned kernel density estimators. *Journal Multivariate Analysis*, 56:165–184, 1996.
- [9] M. Pawlak and Stadtmüller U. Kernel density estimation with generalized binning. *Scandinavian Journal of Statistics*, 26:1–23, 1999.
- [10] J.A. Rice. *Mathematical Statistics and Data Analysis. 2nd ed.* Belmont, CA: Duxbury Press. xx, 602 p., 1995.
- [11] W. Sun and X. Zhou. Average sampling in spline subspaces. *Appl. Math. Lett.*, 15(2):233–237, 2002.
- [12] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Monographs on Statistics and Applied Probability. 60. London: Chapman & Hall. xii, 212 p., 1995.