Maximum Likelihood Wavelet Density Estimation with Applications to Image and Shape Matching

Adrian Peter[1] and Anand Rangarajan[2]

[1]Dept. of ECE, [2]Dept. of CISE, University of Florida, Gainesville, FL

December 21, 2007

Abstract

Density estimation for observational data plays an integral role in a broad spectrum of applications, e.g. statistical data analysis and information-theoretic image registration. Of late, wavelet based density estimators have gained in popularity due to their ability to approximate a large class of functions; adapting well to difficult situations such as when densities exhibit abrupt changes. The decision to work with wavelet density estimators (WDE) brings along with it theoretical considerations (e.g. non-negativity, integrability) and empirical issues (e.g. computation of basis coefficients) that must be addressed in order to obtain a bona fide density. In this paper, we present a new method to accurately estimate a non-negative density which directly addresses many of the problems in practical wavelet density estimation. We cast the estimation procedure in a maximum likelihood framework which estimates the square root of the density $\sqrt{p}$; allowing us to obtain the natural non-negative density representation $\left(\sqrt{p}\right)^2$. Analysis of this method will bring to light a remarkable theoretical connection with the Fisher information of the density and consequently lead to an efficient constrained optimization procedure to estimate the wavelet coefficients. We illustrate the effectiveness of the algorithm by evaluating its performance on mutual information based image registration, shape point set alignment and empirical comparisons to known densities. The present method is also compared to fixed and variable bandwidth kernel density estimators (KDE).

1 Introduction

Density estimation is a well-studied field, encompassing a myriad of techniques and theoretical formulations all with the common goal of utilizing the observed data $X = \{x_i\}_{i=1}^N$ to discover the best approximation to the underlying density that generated them. Methods range from simple histogramming to more statistically efficient kernel based Parzen window techniques [1, 2]. Within the last 20 years, the widespread use of wavelet analysis in applied mathematics and engineering has also made its way into statistical applications. The use of wavelets as a density estimator was first explored in [3]. Wavelet bases have the desirable property of being able to approximate a large class of functions ($\mathbb{L}^2$). Specifically for density estimation, wavelet analysis is often performed on normed spaces that have some notion of regularity like Besov, Hölder and Sobolev. From an empirical point of view, the utility of representing a density in a wavelet basis comes from the fact that they are able to achieve good global approximation properties due to their locally compact nature - a key property when it comes to modeling densities that contain bumps and/or abrupt variations. It is well known that wavelets are localized in both time and frequency and this "compactness" is highly desirable in density estimation as well.

The basic idea behind wavelet density estimation (for one-dimensional data) is to represent the density p as a linear combination of wavelet bases

$$p(x) = \sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \geq j_0,k}^{\infty} \beta_{j,k}\psi_{j,k}(x) \tag{1}$$

where $x \in \mathbb{R}$, $\phi(x)$ and $\psi(x)$ are the *scaling* (a.k.a. father) and *wavelet* (a.k.a. mother) basis functions respectively, and $\alpha_{j_0,k}$ and $\beta_{j,k}$ are scaling and wavelet basis function coefficients; the j-index represents the current level and the k-index the integer translation value. (The translation range of k can be computed from the span of the data and basis function support size [4].) Our goal then is to estimate the coefficients of the wavelet expansion and obtain an estimator $\hat{p}$ of the density. This should be accomplished in a manner that retains the properties of the true density—notably the density should be non-negative and integrate to one. Typically, wavelet density estimators (WDE) are classified as linear and non-linear. The term linear estimator denotes the fact that the coefficients are obtained via a projection of the density's distribution onto the space spanned by the wavelet basis. Non-linear estimators threshold the estimated coefficients, both globally and

locally, to obtain optimal convergence to the true density. These estimators, especially those with thresholding, often cannot guarantee that the resulting $\hat{p}$ from the estimation process satisfies the aforementioned properties.

To guarantee these properties, one typically resorts to estimating $\sqrt{p}$ as

$$\sqrt{p(x)} = \sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j\geq j_0,k}^{\infty} \beta_{j,k}\psi_{j,k}(x) \tag{2}$$

which directly gives $p = \left(\sqrt{p}\right)^2$. Previous work on wavelet density estimation of $\sqrt{p}$ [5, 6], stays within the projection paradigm of trying to estimate the coefficients as an inner product with the corresponding (orthogonal) basis. As such, they have to directly address the estimation of the scalar product involving the square root estimate and the appropriate basis function, e.g. estimating coefficient $\alpha_{j_0,k}$ requires finding an acceptable substitute for $\int_{\mathbb{R}^d} \sqrt{p(x)}\phi_{j_o,k}(x)dx$. In this work, we will show how to completely avoid this paradigm by casting the estimation process in a maximum likelihood framework. The wavelet coefficients of the $\sqrt{p}$ expansion are obtained by minimizing the negative log likelihood over the observed samples with respect to the coefficients. Moreover, asymptotic analysis will illustrate a remarkable property of the Fisher information matrix of the density under this representation—which is $4\mathbf{I}$, where $\mathbf{I}$ is the identity matrix—and this is leveraged in the optimization. This highly structured matrix is also the asymptotic Hessian of our maximum likelihood objective function at the optimal solution point. This tight coupling in our estimation framework will lead to a very efficient, modified Newton's method for computing the wavelet coefficients. We maintain all the desirable properties that inherently come with estimating $\sqrt{p}$, while circumventing the need to establish the previously mentioned substitute. The focus of this paper is to carefully establish these connections and discuss the resulting algorithms for estimating both one and two dimensional densities.

As researchers in the field of image analysis, we have a penchant toward image processing oriented applications. To this end, we provide proof-of-concept demonstrations through information-theoretic shape registration and mutual information (MI) based affine registration of medical imagery [7]. Given a pair of images, one image, designated the source, is considered as similar under affine transformations to a pre-specified target image. MI-based image alignment tries to maximize the mutual information between the image pair, which hopefully occurs when they are optimally aligned. The algorithm requires a density estimation step that computes the joint density between the image

pairs. This density is then used to calculate the mutual information. For shape registration, we adopt the correspondence-free approach as presented in [8] but replace the performance criterion with the Hellinger divergence [9]. For both applications, we replace the typical density estimator, usually a 2D histogram, kernel estimator or mixture model, with our wavelet density estimator and analyze its viability. We also provide anecdotal results of the wavelet density estimator's ability to model true analytic 1D and 2D densities such as Gaussian mixture models.

The rest of this paper is organized in the following manner. In Section 2, we briefly recap multiresolution wavelet analysis and provide a more in depth discussion concerning wavelet density estimation. Section 3 discusses our maximum likelihood framework for estimating the wavelet coefficients for $\sqrt{p}$. It goes on to detail the modified Newton's method used to efficiently compute the coefficients. In Section 4, our method is validated in the application setting of image and shape registration and we showcase the capability to model known densities and compare performance against fixed and variable bandwidth kernel density estimators.

2 Wavelet Theory and its Application to Density Estimation

We now provide a basic introduction to the ideas of wavelet multiresolution theory and move on to discussing how these concepts are carried out in wavelet-based density estimation. We will be focused throughout on clearly communicating the conceptual aspects of the theory, diverting much of the mathematical machinery to the (hopefully) appropriately cited references.

2.1 Multiresolution

For any function $f \in \mathbb{L}^2$ and a starting resolution level j_0, representation in the wavelet basis is given by

$$f(x) = \sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \geq j_0,k}^{\infty} \beta_{j,k}\psi_{j,k}(x), \tag{3}$$

where

$$\begin{aligned} \phi_{j_o,k}(x) &= 2^{j_0/2}\phi(2^{j_0}x - k), \\ \psi_{j,k}(x) &= 2^{j/2}\psi(2^{j}x - k), \end{aligned} \tag{4}$$

are scaled and translated version of the father $\phi(x)$ and mother $\psi(x)$ wavelets. The key idea behind multiresolution theory is a sequence of nested subspaces V_j $j \in \mathbb{Z}$ such that

$$\cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \cdots \tag{5}$$

and which satisfy the properties $\bigcap V_j = \{0\}$ and $\overline{\bigcup V_j} = \mathbb{L}^2$ (completeness). The resolution increases as $j \to \infty$ and decreases as $j \to -\infty$ (some references show this order reversed due to the fact they invert the scale [10]). At any particular level $j+1$, we have the following relationship

$$V_j \bigoplus W_j = V_{j+1} \tag{6}$$

where W_j is a space orthogonal to V_j, i.e. $V_j \bigcap W_j = \{0\}$. The father wavelet $\phi(x)$ and its integer translations form a basis for V_0. The mother wavelet $\psi(x)$ and its integer translates span W_0. These spaces decompose the function into its smooth and detail parts; this is akin to viewing the function at different scales and at each scale having a low pass and high pass version of the function.

We will assume $\phi(x)$, $\psi(x)$ and their scaled and translated versions form orthogonal bases for their respective spaces. Under these assumptions, the standard way to calculate the coefficients for (3) is by using the inner product of the space, e.g. the coefficient $\alpha_{j_0,k}$ is obtained by

$$\alpha_{j_0,k} =< f, \phi_{j_0,k} >= \int f(x)\phi_{j_0,k}(x)dx \tag{7}$$

where we have used the $\mathbb{L}^2$ inner product. Most of the existing wavelet-based density estimation techniques exploit this projection paradigm to estimate the coefficients. Replacing the general function f by a density p, the coefficients for (1) can be calculated as

$$\alpha_{j_0,k} = \int p(x)\phi_{j_0,k}(x)dx = \mathcal{E}\left[\phi_{j_0,k}(x)\right] \tag{8}$$

where $\mathcal{E}$ is the expectation operator. Given N samples, this is approximated as the sample average

$$\alpha_{j_0,k} = \frac{1}{N}\sum_{i=1}^{N}\phi_{j_0,k}(x_i). \tag{9}$$

Many density estimation techniques, including ours, require evaluating $\phi(x)$ and $\psi(x)$ at various

domain points in their support region. However, most father and mother wavelets do not have an analytic closed-form expression. The strategy is to use the close coupling between the scaling function and wavelet—find $\phi(x)$ by numerically solving the *dilation equation* and then directly obtain $\psi(x)$ by solving the *wavelet equation*. The dilation equation is given by

$$\phi(x) = 2\sum_{k} l(k)\phi(2x-k) \tag{10}$$

where $l(k)$ are the low pass filter coefficients associated with a particular scaling function family. This can be numerically solved using iterative procedures such as the cascade algorithm [11]. Upon solving for $\phi(x)$, one can immediately get the associated wavelet function by using the high pass filter coefficients $h(k)$ and solving the wavelet equation

$$\psi(x) = 2\sum_{k} h(k)\phi(2x-k). \tag{11}$$

The numerical versions of $\phi(x)$ and $\psi(x)$ will have values at domain points that are integer multiples of $\frac{1}{2^M}$ (where M controls the discretization level). If an x value lands in between these grid points, it is a straightforward process to interpolate and get the desired value. We have adopted a cubic spline interpolation strategy to obtain the intermediate values.

2.2 Wavelet Density Estimation

A practical consideration of using wavelets for density estimation requires careful consideration of several issues. First, one must decide the family of wavelets that will be used as the basis. Though this issue has received considerably less attention in theoretical literature, a pragmatic solution suggests that the choice is closely tied to problem domain. However, one almost always assumes that the bases satisfy the desirable orthonormality property and are compactly supported. Also, multiresolution analysis for function approximation requires the use of bases that have a coupled scaling and wavelet function relationship. This restricts our choice of basis functions to families such as Haar, Daubechies, Coiflets and Symlets. It is worth mentioning that these basis functions are not perfectly symmetric (except for Haar); in fact in classical wavelet analysis, symmetry is a detriment to perfect reconstruction of the signal [10]. Exact symmetry is not a critical requirement for density estimation and families like Symlets exhibit characteristics close to symmetry for higher order

vanishing moments. Throughout this paper, we assume that our bases are orthogonal, compactly supported and have both a scaling and wavelet function.

Second, and perhaps most importantly, we must address efficient computation of the basis coefficients and the impact they have on the properties of the estimated density. The main contribution of this paper is to further advance the theoretical framework for the estimation of these coefficients, while maintaining the important properties of a bona fide density—non-negative in its support and integrating to one.

Finally, it is necessary to consider the practical issue of selecting the basis truncation parameters. Notice that in (2), convergence of the wavelet representation to the true function assumes a starting resolution level j_0 and uses an infinite number of detail resolution levels. In practical computation, it is necessary to develop a principled way of choosing j_0 and also a stopping level j_1 as we cannot have an infinite expansion. These issues are necessarily addressed by model selection methods such as cross validation. Model selection is not the focus of our current work. It is possible to adopt any of the existing model selection methods, as summarized in [12], to appropriately choose these parameters and incorporate it into our framework.

Returning to the second point, classical wavelet density estimation [13, 14] does not try to explicitly ensure that the density is non-negative and usually suffers from negative values in the tails of the density. For example, in the work of Donoho *et. al.* [13], this artifact is introduced by the necessity to threshold the coefficients. Non-linear thresholding obtains better convergence and achieves the optimal minimax rate under the global integrated mean squared error (IMSE) $\mathcal{E}\left[\|\hat{p}-p\|^2\right]$ measure, where $\mathcal{E}$ is the expectation operator. Though these are favorable properties, it is still somewhat unsettling to have a density with negative values. Also, thresholding leads to the problem of having to renormalize the coefficients to maintain the integrable property of the estimated density. Under the usual non-linear estimation process, this is not a straightforward procedure and may require further integration to work out the normalizing constant. Next we will show how it is possible to incorporate the benefits of thresholding the coefficients while maintaining the integrity of the estimated density.

The preferred way to maintain these properties is to estimate the square root of the density $\sqrt{p}$ rather than p. Estimating $\sqrt{p}$ has several advantages: (i) non-negativity is guaranteed by the fact $p = \left(\sqrt{p}\right)^2$ (ii) integrability to one is easy to maintain even in the presence of thresholding, and (iii) the square root is a variance stabilizing transform [15]. The present work also falls into this

category of techniques that estimate $\sqrt{p}$. To our knowledge there are only two previous works [5, 6] that estimate the square root of the density using a wavelet basis expansion.

We begin by representing the square root of the density using

$$\sqrt{p(x)} = \sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \geq j_0,k}^{j_1} \beta_{j,k}\psi_{j,k}(x). \tag{12}$$

Imposing our integration condition, $\int_{\mathbb{R}^d} \left(\sqrt{p}\right)^2 dx = 1$, implies that

$$\sum_{j_0,k} \alpha_{j_0,k}^2 + \sum_{j \geq j_0,k}^{j_1} \beta_{j,k}^2 = 1. \tag{13}$$

Notice that if (13) is not one but some arbitrary constant D, such as when a thresholding scheme changes the weights, it is possible to perform a straightforward renormalization by merely dividing the coefficients by $\sqrt{D}$.

In the previous two works [5, 6], estimation of $\alpha_{j_0,k}$ and $\beta_{j,k}$ is motivated by applying the previously discussed projection method. We are now working with $\sqrt{p}$, however, which changes (8) to

$$\begin{aligned} \alpha_{j_0,k} &= \int \sqrt{p(x)}\phi_{j_0,k}(x)dx \\ &= \int \frac{p(x)}{\sqrt{p(x)}}\phi_{j_0,k}(x)dx \\ &= \mathcal{E}\left[\frac{\phi_{j_0,k}(x)}{\sqrt{p(x)}}\right]. \end{aligned} \tag{14}$$

The $\beta_{j,k}$ coefficients are defined by analogy. In [5], the authors propose a suitable substitute to the empirical estimator $\frac{1}{N}\sum_{i=1}^{N} \phi_{j_0,k}(x)/\sqrt{p(x)}$, but the coefficient estimation is sensitive to the pre-estimator of $p(x)$. In the work by Penev and Dechevsky [6], the coefficient computation is based on order statistics of the sample data. As we illustrate in the next section, the method we present avoids these issues by casting the density estimation problem in a maximum likelihood setting. The maximum likelihood model also ensures the asymptotic consistency of our estimated coefficients. To complete the spectrum of non-negative density estimation techniques, it is worth mentioning that Walter [16] presents an alternative using a clever construction of non-negative wavelets; exploration of this method, however, is beyond the present scope.

3 Maximum Likelihood for Wavelet Density Estimation

We now discuss how to cast wavelet density estimation in a maximum likelihood framework. Often maximum likelihood is designated a parametric technique and reserved for situations where we are able to assume a functional form for the density. Thus, current research typically categorizes wavelet density estimation as a non-parametric estimation problem. Treating the coefficients as the parameters we wish to estimate, however, allows us to move the problem into the parametric realm. This interpretation is also possible for other formulations such as kernel density estimation, where maximum likelihood is applied to estimate the kernel bandwidth parameter. For estimating the wavelet coefficients, adopting the maximum likelihood procedures will lead to a constrained optimization problem. We then investigate the connections between estimating $\sqrt{p}$ and the Fisher information of the density. Exploring this connection will allow us to make simplifying assumptions about the optimization problem, resulting in an efficient modified Newton's method with good convergence properties. We will present derivations for both 1D and 2D density estimation, extensions to higher dimensions would follow a similar path.

3.1 1D Constrained Maximum Likelihood

Let $X = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}$ represent N i.i.d. samples from which we will estimate the parameters of the density. As is often customary, we will choose to minimize the negative log likelihood rather than maximize the log likelihood. The negative log likelihood objective is given by

$$\begin{aligned} -\log p(X; \{\alpha_{j_0,k}, \beta_{j,k}\}) &= -\tfrac{1}{N} \log \textstyle\prod_{i=1}^N \left[\sqrt{p(x_i)}\right]^2 \\ &= -\tfrac{1}{N} \textstyle\sum_{i=1}^N \log \Big[\sum_{j_0,k} \alpha_{j_0,k}\phi_{j_0,k}(x_i) \\ &\qquad + \textstyle\sum_{j\geq j_0,k}^{j_1} \beta_{j,k}\psi_{j,k}(x_i)\Big]^2. \end{aligned} \tag{15}$$

Trying to minimize (15) directly w.r.t. $\alpha_{j_0,k}$ and $\beta_{j,k}$ would result in a density estimator which does not integrate to one. To enforce this condition we require the following equality constraint

$$h(\{\alpha_{j_0,k}, \beta_{j,k}\}) = \left[\sum_{j_0,k} \alpha_{j_0,k}^2 + \sum_{j\geq j_0,k} \beta_{j,k}^2\right] - 1 = 0. \tag{16}$$

This constraint can be incorporated via a Lagrange parameter λ to obtain the following constrained objective function

$$\begin{aligned}\mathcal{L}(X, \{\alpha_{j_0,k}, \beta_{j,k}\}, \lambda) \quad &= -\log\, p(X; \{\alpha_{j_0,k}, \beta_{j,k}\}) \\ &\quad +\lambda h(\{\alpha_{j_0,k}, \beta_{j,k}\})\end{aligned} \tag{17}$$

The constraint (16) dictates that the solution to (17) lives on a unit hypersphere; it can be solved using standard constrained optimization techniques. Before presenting our particular solution, we explore the properties of the Fisher information matrix associated with this problem.

3.2 The Many Faces of Fisher Information

The classic form of the Fisher information matrix is given by

$$g_{uv}(\theta) = \int p(\mathbf{x}; \theta) \frac{\partial}{\partial \theta^u} \log p(\mathbf{x}; \theta) \frac{\partial}{\partial \theta^v} \log p(\mathbf{x}; \theta) d\mathbf{x}, \tag{18}$$

where the (u, v) index pair denotes the row, column entry of the matrix and consequently the appropriate parameter pair. Intuitively, one can think of the Fisher information as *a measure of the amount of information present in the data about a parameter* θ; for wavelet density estimation $\theta = \{\alpha_{j_0,k}, \beta_{j,k}\}$ and the (u, v)-indexing is adjusted to be associated with the appropriate level, translation index pair, i.e. $\{u = (j, k), v = (l, m)\}$where j and l are the level indices and k and m are the translation indices. For the current setting, where we are estimating $\sqrt{p}$, the Fisher information has a more pertinent form

$$\begin{aligned}\tilde{g}_{uv} \quad &= \quad \int \frac{\partial \sqrt{p(x|\theta)}}{\partial \theta^u} \frac{\partial \sqrt{p(x|\theta)}}{\partial \theta^v} dx \\ &\quad \Rightarrow g_{uv} = 4\tilde{g}_{uv}.\end{aligned} \tag{19}$$

Hence, $\tilde{g}_{uv}$ computed using the square root of the density differs only by a constant factor from the true Fisher information. Using this algebraic relationship, the Fisher information of the wavelet density estimator can be calculated by substituting the wavelet expansion of $\sqrt{p(x; \theta)}$ given in (2). This gives

$$\begin{aligned}\tilde{g}_{uv} \quad &= \quad \int \phi_u(x)\psi_v(x)dx \\ &= \quad \begin{cases} 1 & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases},\end{aligned} \tag{20}$$

where we have leveraged the orthogonal property of the wavelet basis functions. This is an identity matrix and the Fisher information of $p(x;\theta)$ can be written as $g_{uv} = 4\mathbf{I}$.

There is another algebraic manipulation that allows us to compute the Fisher information using the Hessian of the log likelihood, specifically

$$g_{uv} = -\mathcal{E}\left[\nabla\nabla^T \log p(x;\theta)\right] = -\mathcal{E}\left[H\right], \tag{21}$$

where ∇ is the gradient operator w.r.t. the parameters and H is the Hessian matrix of the multi-parameter negative log likelihood. Recalling that equation (15) is the negative log likelihood, we can immediately make the connection that (17)'s asymptotic Hessian should be $H_{\mathcal{L}} = g_{uv} = 4\mathbf{I}$. To verify this, let

$$H_{\mathcal{L}} = H_{nll} + \lambda H_h \tag{22}$$

where H_{nll} and H_h are the Hessian of the negative log likelihood (15) and constraint equation (16), respectively. Equation (22) is the Hessian of the Lagrangian which is typical of constrained minimization problems. We illustrate the computation of the asymptotic H_{nll} by providing results for a particular coefficient β; other coefficients are calculated in a similar manner. The second partial derivative of (15) is

$$\frac{\partial^2}{\partial\beta_{h,l}\partial\beta_{p,m}}[-\log\, p] = 2\frac{\psi_{h,l}(x)\psi_{p,m}(x)}{p\left(x;\theta=\{\alpha_{j_0,k},\beta_{j,k}\}\right)} \tag{23}$$

and taking its expected value, we get

$$\begin{array}{rcl} \mathcal{E}\left[\frac{\partial^2}{\partial\beta_{h,l}\partial\beta_{p,m}}[-\log\, p]\right] & = & 2\int \psi_{h,l}(x)\psi_{p,m}(x)dx \\ & = & 2\delta_{hp}\delta_{lm}. \end{array} \tag{24}$$

(Note: for readability we have let $p \equiv p(X;\{\alpha_{j_0,k},\beta_{j,k}\})$.) The Hessian of (16) is constant and is consequently equal to its expected value, i.e.

$$\begin{array}{rcl} \frac{\partial^2}{\partial\beta_{h,l}\partial\beta_{p,m}}h & = & \mathcal{E}\left[\frac{\partial^2}{\partial\beta_{h,l}\partial\beta_{p,m}}h\right] \\ & = & 2\delta_{hp}\delta_{lm}. \end{array} \tag{25}$$

Both (24) and (25) are $2\mathbf{I}$ matrices. Hence, referring back to (22), in order for $H_{\mathcal{L}} = 4\mathbf{I}$ we require $\lambda = 1$ at the optimal solution point; the proof of which is obtained through algebraic manipulation of the Lagrangian's, eq. (17), first-order necessary conditions. Intuitively, what we have shown is

that under an orthonormal expansion of the square root of density, the Fisher information matrix essentially specifies a hypersphere [17].

3.3 Efficient Minimization using a Modified Newton's Method

In light of the discussion in the previous section, we proceed to design an efficient optimization method to iteratively solve for the coefficients. A Newton's method solution to (17) would result in the following update equations at iteration τ

$$\begin{bmatrix} x^{\tau+1} \\ \lambda^{\tau+1} \end{bmatrix} = \begin{bmatrix} x^{\tau} \\ \lambda^{\tau} \end{bmatrix} - \begin{bmatrix} H_{\mathcal{L}}^{\tau} & A^{\tau} \\ (A^{\tau})^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} l^{\tau} \\ h^{\tau} \end{bmatrix} \tag{26}$$

where $x^{\tau} = (\alpha_{j_0,k}^{\tau}, \beta_{j,k}^{\tau})$, $A^{\tau} = \nabla h(x^{\tau})$, $l^{\tau} = [\nabla nll(x^{\tau}) + \lambda^{\tau} \nabla h(x^{\tau})]$, $h^{\tau} = h(x^{\tau})$, and $H_{\mathcal{L}}^{\tau} = H_{\mathcal{L}}(x^{\tau})$ and $nll(\theta^{\tau}) \stackrel{\text{def}}{=} -\log p(X;\theta)$. When the Hessian is positive definite throughout the feasible solution space Ω, it is possible to directly solve for x^{τ} and λ^{τ} [18]. For (17), it is certainly true that $x^T H_{\mathcal{L}} x > 0$ over Ω. In order to avoid the computationally taxing $H_{\mathcal{L}}$ update at each iteration, we adopt a modified Newton's method[18]. Modified Newton techniques replace $H_{\mathcal{L}}$ by B, where B is a suitable approximation to $H_{\mathcal{L}}$. Here we can take advantage of the fact we know $H_{\mathcal{L}}$ at the optimal solution point; hence, we let $B = H_{\mathcal{L}}^* = 4\mathbf{I}$. In practice, this method is implemented by solving the system

$$\begin{aligned} \lambda^{\tau+1} &= C^{-1}\left[h^{\tau} - F\nabla nll(x^{\tau})\right], \\ d^{\tau} \quad &= -B^{-1} \quad \left[I - A^{\tau}C^{-1}F\right]\nabla nll(x^{\tau}) \\ &\quad -B^{-1}A^{\tau}C^{-1}h^{\tau}. \end{aligned} \tag{27}$$

where $C = (A^{\tau})^T B^{-1} A^{\tau}$, $F = (A^{\tau})^T B^{-1}$ and the coefficient updates are given by $x^{\tau+1} = x^{\tau} + d^{\tau}$. Notice that these equations can be simplified even further by taking advantage of the simple structure of B to avoid explicit matrix inverses and making it very efficient for implementation, i.e. set $B^{-1} = \frac{1}{4}\mathbf{I}$. This method depends on having a unit step size and has convergence properties comparable to the standard Newton's method [18]. It was also shown in [19, 20] that using the known Hessian at the optimal solution points has the effect of doubling the convergence area thus making it robust to various initializations.

Algorithm 1 Wavelet density estimation using modified Newton's method.

1. Initialize $x^0 = \{\alpha_{j_0,k}, \beta_{j,k}\}$. We typically set all values to be equal subject to the constraint $\sum_{j_0,k} \alpha_{j_0,k}^2 + \sum_{j \geq j_0,k} \beta_{j,k}^2 = 1$.
2. Perform modified Newton updates in (27) to get coefficient increments d^τ.
3. Update coefficients according to $x^{\tau+1} = x^\tau + d^\tau$.
4. Repeat Steps 2 and 3 until convergence which gives a minimum $\hat{x}$ to our objective (17).
5. Use estimated set of coefficients $\hat{x} = \left\{\hat{\alpha}_{j_0,k}, \hat{\beta}_{j,k}\right\}$ to construct $\hat{p} = (\sqrt{p})^2$ as in (12) for 1D or (28) for 2D.

3.4 2D Density Estimation

Extensions to bivariate, wavelet density estimation are made possible by using the tensor product method to construct 2D wavelet basis functions from their 1D counterparts [10]. The notation becomes noticeably more complicated and requires careful attention during implementation. Let $(x_1, x_2) = \mathbf{x} \in \mathbb{R}^2$ and now the $\sqrt{p(\mathbf{x})}$ expansion is given by

$$\sqrt{p(\mathbf{x})} = \sum_{j_0,\mathbf{k}} \alpha_{j_0,\mathbf{k}} \phi_{j_0,\mathbf{k}}(\mathbf{x}) + \sum_{j \geq j_0,\mathbf{k}}^{j_1} \sum_{w=1}^{3} \beta_{j,\mathbf{k}}^w \psi_{j,\mathbf{k}}^w(\mathbf{x}) \tag{28}$$

where $(k_1, k_2) = \mathbf{k} \in \mathbb{Z}^2$ is a multi-index. The tensor products are

$$\begin{aligned} \phi_{j_0,\mathbf{k}}(\mathbf{x}) &= 2^{j_0} \phi(2^{j_0} x_1 - k_1) \phi(2^{j_0} x_2 - k_2) \\ \psi_{j,\mathbf{k}}^1(\mathbf{x}) &= 2^j \phi(2^j x_1 - k_1) \psi(2^j x_2 - k_2) \\ \psi_{j,\mathbf{k}}^2(\mathbf{x}) &= 2^j \psi(2^j x_1 - k_1) \phi(2^j x_2 - k_2) \\ \psi_{j,\mathbf{k}}^3(\mathbf{x}) &= 2^j \psi(2^j x_1 - k_1) \psi(2^j x_2 - k_2). \end{aligned} \tag{29}$$

Again our goal is to estimate the set of coefficients $\left\{\alpha_{j_0,\mathbf{k}}, \beta_{j,\mathbf{k}}^w\right\}$. As in the univariate case, we repeat the necessary steps by first creating the objective function that incorporates the negative log likelihood and the Lagrange parameter term to handle the equality constraint. Then the minimization procedure follows according to Section 3.3. The resulting equations are exactly the same form with straightforward adjustments during implementation of (27) to incorporate the 2D nature of the indices and wavelet basis. The algorithm to perform one or two dimensional wavelet density estimation using our modified Newton's method is presented in Algorithm 1.

4 Experimental Results

As our work is a general density estimation technique, it is applicable to a whole host of applications that rely on estimating densities from observational data. We are still in the early stages of identifying and exploring the best-of-class applications that are well suited to take advantage of the underlying flexibility that comes with wavelet-based analysis. The present experimental evaluation of the proposed methods was conducted on both synthetic and real data. We measure performance by validation against true, analytical densities (both 1D and 2D) and illustrate proof-of-concept scenarios for real data applications that require density estimation as part of their solution. Specifically, the two applications we showcase here are shape alignment and mutual-information based image registration. Though our method has some advantages over contemporary wavelet density estimators, there are still practical considerations that all, including our, wavelet-based solutions bump against. These considerations are peppered throughout our analysis of the experimental results.

4.1 One and Two Dimensional Density Approximation

The approximation power of the present method was validated against the class of known densities as presented in [21] and [22], where the authors provide constructions of several densities (which are all generated as appropriate mixtures of Gaussians) that analytically exercise representative properties of real densities, such as spatial variability, asymmetry and concentrated peaks. Most other wavelet density estimators have also showcased their results on a small subset of these densities. In order to provide comprehensive, robust analysis of our method, we selected the following 13 one dimensional densities to analyze approximation capabilities: Gaussian, skewed unimodal, strongly skewed unimodal, kurtotic unimodal, outlier, bimodal, separated bimodal, skewed bimodal, trimodal, claw, double claw, asymmetric claw, and asymmetric double claw. The reader is referred to [21] for a visual depiction of all 13 densities. During preliminary empirical evaluation, we noticed a trend that best results were observed when using a single-level, scaling function representation of the density. This was further confirmed via private communication with G. G. Walter and also discussed in [23]. We performed the estimation over a range of scale values, i.e. j_0 in equation (12), from $j_0 = -1$ to 5. (Note: we initially used a cross validation method, see [12], to automatically select j_0 but opted to test over a range to be more thorough). We also used three different families of wavelet basis, with multiple orders within a family, to approximate each of the densities—Daubechies of

	WDE			KDE	
	Best Basis	j_0^*	ISE	FixedBW†ISE	Var.BW†ISE
Gaussian	SYM10	-1	3.472E-04	3.189E-04	5.241E-03
Skewed Uni.	DB7	1	3.417E-04	1.970E-04	7.551E-03
Str. Skewed Uni.	SYM7	3	2.995E-03	6.947E-02	7.610E-03
Kurtotic Uni.	COIF2	2	2.399E-03	2.869E-02	1.388E-02
Outlier	SYM10	2	1.593E-03	3.911E-03	3.962E-02
Bimodal	COIF5	0	5.973E-04	2.084E-04	4.223E-03
Sep. Bimodal	SYM7	1	5.354E-04	6.237E-03	5.419E-03
Skewed Bimodal	DB10	1	8.559E-04	1.461E-03	3.885E-03
Trimodal	COIF3	1	9.811E-04	1.439E-03	3.787E-03
Claw	SYM10	2	1.511E-03	3.692E-02	7.014E-03
Dbl. Claw	COIF1	2	2.092E-03	1.795E-03	5.283E-03
Asym. Claw	DB3	3	2.383E-03	1.373E-02	6.490E-03
Asym. Dbl. Claw	COIF1	2	2.250E-03	4.759E-03	4.940E-03

Table 1: 1D Density Estimation. Optimal start level j_0^* was selected by taking lowest ISE for $j_0 \in [-1, 5]$. (†BW is bandwidth.)

order 1-10, Symlets of order 4-10, and Coiflets of order 1-5. The analyses across families provide some guidance as to the approximation performance capabilities of different bases. All densities were estimated using 2,000 samples drawn from the known analytic densities. The approximation error of each test was measured using the integrated squared error (ISE) between the known p and estimated $\hat{p}$ densities, i.e. $\int_{\mathbb{R}} (p - \hat{p})^2 \, dx$. This was computed by discretizing the 1D support of the density at equally spaced points and then summing area measures. The best wavelet basis, order and starting level j_0 were selected based on this error measure. We executed 15 iterations of our modified Newton's method with many densities converging in fewer (8 to 10) iterations – 15 iterations on 2000 samples takes approximately 30 seconds with a Matlab implementation. We compared the wavelet reconstructed densities with kernel density estimators (KDE) [2]. All experiments were conducted using a Gaussian kernel, a reasonable choice since the true densities are mixtures of Gaussians, with the smoothing parameter (bandwidth of kernel) selected two ways: (1) automatically as described in [2, Ch. 3] and (2) a nearest-neighbor variable bandwidth, see [2, Ch. 5]. The freely available KDE Toolbox for Matlab [24] by A. Iheler provides fast, robust implementation of these kernel methods. We refer to the automatically selected bandwidth as *fixed* since it is the same for all kernels. The 1D results are summarized in Table 1.

For the two dimensional evaluation, we selected the following five difficult densities, i.e. ones that exhibited more variations or closely grouped Gaussians, from [22]: bimodal IV, trimodal III, kurtotic, quadrimodal and skewed. The experimental procedure was similar to that of the one

	WDE			KDE	
	Best Basis	j_0^*	ISE	Fixed BW† ISE	Var. BW† ISE
Bimodal IV	SYM7	1	6.773E-03	1.752E-02	8.114E-03
Trimodal III	COIF2	1	6.439E-03	6.621E-03	1.037E-02
Kurtotic	COIF4	0	6.739E-03	8.050E-03	7.470E-03
Quadrimodal	COIF5	0	3.977E-04	1.516E-03	3.098E-03
Skewed	SYM10	0	4.561E-03	8.166E-03	5.102E-03

Table 2: 2D Density Estimation. Optimal start level j_0^* was selected by taking lowest ISE for $j_0 \in [-1, 3]$. (†BW is bandwidth.)

dimensional densities and results are summarized in Table 2. Again our modified Newton's method was able to converge with fewer than 15 iterations. In all 2D test cases, our wavelet density estimator was able to outperform both the fixed and variable bandwidth kernel density estimators. Overall, in both the 1D and 2D cases the wavelet bases were able to accurately represent the true densities. Of the families tested, there was no clear-cut winner as to which basis was better than another. The performance of a particular basis depended on the shape of the true density. Some applications may prefer Symlets or Coiflets as they are "more symmetric" than other bases. In comparison to the kernel estimators, our method provided better results on the more difficult densities and performed only slightly worse on slowly varying ones. Also, there were instances where variable bandwidth KDE provided a lower ISE but visually exhibited more peaks than the true density. In the future, we plan to do further analysis to better evaluate this bias-variance tradeoff when selecting the best density approximation. Examples of estimated densities and some of their properties are illustrated in Figures 6 and 5.

4.2 WDE for Registration and Shape Alignment

Information-theoretic approaches have been applied to a variety of image analysis and machine learning problems [25, 26, 27]. These techniques typically require estimating the underlying density from which the given data are generated. We applied our new density estimation procedure to two information-theoretic methods that utilize the Shannon entropy of the data. We begin with the well established image registration method based on mutual information [7]. This multi-modal registration method iteratively optimizes the mutual information (MI) between a pair of images over the assumed parameter space of their differing transformation. We next implement an adaptation of a more contemporary method which minimizes the Jensen-Shannon divergence in order to align two

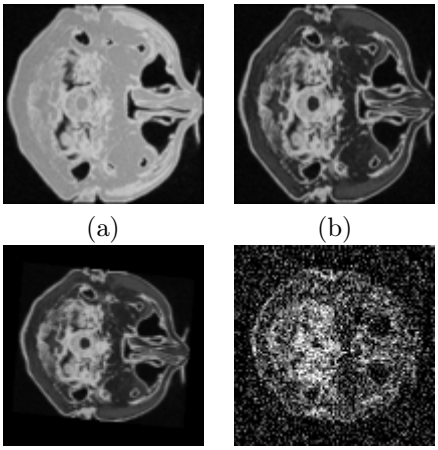


Figure 1: Registration using mutual information. (a,b) Registration image pair. (c) Affine warp applied to (b) without noise. Target image (c) with noise.

point-set representation of shapes [8]. Both of these methods require estimating two dimensional densities.

4.2.1 Registration Using Mutual Information

For the MI registration experiments, we used slices from the Brainweb simulated MRI volumes for a normal brain [28]. The goal was to recover a global affine warp between an image pair. (Note: To expedite experiments, we did not include translations in the affine warp.) We follow the affine warp decomposition used in [29], which results in a four parameter search space (θ,ϕ,s,t). In order to minimize experimental variability, we manually imposed a known affine transformation between the source and target image. The optimal parameter search was conducted using a coarse-to-fine search strategy over a bounded, discretized range of the parameter space. Calculating the mutual information performance criterion between image pairs requires estimating the joint density between them. This is typically accomplished using a simple 2D histogram or Parzen window estimator (i.e. the kernel estimators evaluated in §4.1). We replace these methods with our wavelet based density estimator, leaving the rest of the algorithm unchanged. Figure 1 shows the source and target images used in the trials and the results are listed in Table 3. In the absence of noise, we were able to perfectly recover the transformation parameters using our method whereas KDE (both fixed and variable bandwidth) failed to estimate the optimal parameters. In the noise trial, we were able to correctly recover two, s and t, out of four parameters with the other two, θ and ϕ, values only

	Truth	WDE		Fixed BW KDE		Var. BW KDE	
		$\sigma^2 = 0$	$\sigma^2 = 0.05$	$\sigma^2 = 0$	$\sigma^2 = 0.05$	$\sigma^2 = 0$	$\sigma^2 = 0.05$
θ	10	10	9.8	9.8	10.2	10	10.4
ϕ	-5	-5	-5.4	-5	-4.8	-4.6	-5
s	0.3	0.3	0.3	0.3	0.3	0.2	0.3
t	-0.1	-0.1	-0.1	-0.1	-0.1	0	-0.1

Table 3: Mutual information registration results. The left-most column are affine parameters (see text). With no noise, $\sigma^2 = 0$, the parameters were exactly recovered when using WDE whereas KDE estimates were slightly off. With noise added, the recovered values were close to the truth but there was not a significant advantage using KDE versus our method.

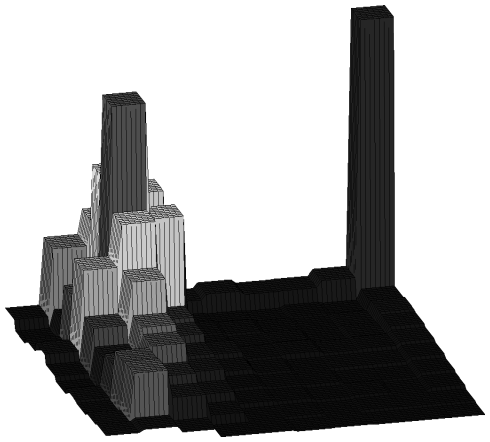
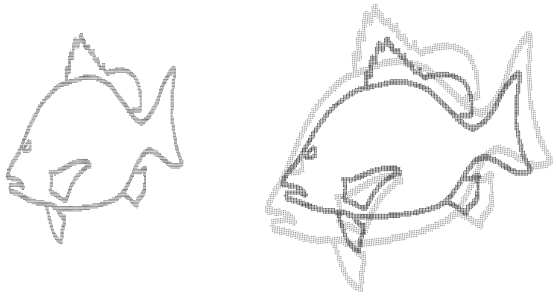


Figure 2: Example of joint density estimation from two images utilized in registration experiments; scaling and wavelet functions from Haar basis using levels $j_0 = -3$ to $j_1 = -2$.

missing the ground truth by one and two discretization steps, respectively (see Table 3). The KDE performed about the same as our method in these noise trials. All experiments were conducted using the Daubechies order 1 (db1) family, with a multi-resolution basis starting at level $j_0 = -3$ stopping at $j_1 = -2$. This means that both scaling and wavelet functions were used in the density estimation, see Figure 2 for an example. It is also possible to use other basis families. Because our optimization does not use a step size parameter, we did encounter some cases where choosing a bad starting level caused convergence issues. This can be remedied by utilizing any standard optimization method that incorporates a line search to control the descent direction. Currently we are using both qualitative analysis (visual inspection) and a 2D version of the cross-validation procedure, as referenced above in the approximation experiments, in order to select the start (j_0) and stop (j_1) levels.



(a) (b)

Figure 3: Fish point sets. (a) Dog snapper represented by 3,040 points. (b) Overlay of source and target shape (lighter shade) used in Hellinger divergence based registration.

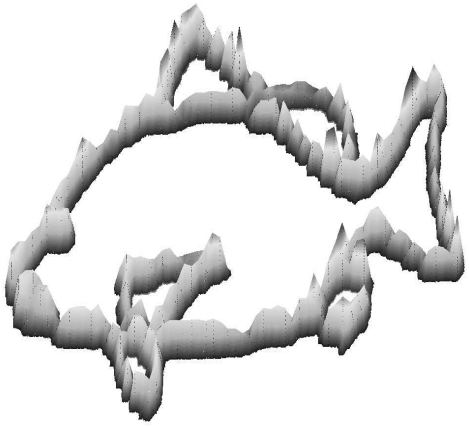


Figure 4: Example of 2D density estimated from fish point set using Coiflet 4, only scaling functions at level $j_0 = 3$.

4.2.2 Shape Alignment Under Hellinger Divergence

Next we applied this density estimation method to shape analysis. The applications of landmark and point-set based shape analysis are often cast in a probabilistic framework which requires a density estimation procedure. The compact, localized nature of the wavelet basis allows one to model a rich class of shapes, with intricate structures and arbitrary topology. We qualitatively illustrate this by estimating the density corresponding to a dog snapper fish shape consisting of 3,040 points, Figure 3(a). The density estimation was carried out using a Coiflet 4 basis with only scaling basis functions starting at level $j_0 = 3$. The estimated density is shown in Figure 4. Notice how the wavelet basis captures the detailed structures such as the fins and closely hugs the spatial support region of the original points. We use our estimation method in a correspondence-free registration framework, as described in [8, 30], to recover an affine transformation between two point sets. In [8],

	Truth	WDE	Fixed BW KDE	Var. BW KDE
θ	10	10	9.8	9.8
ϕ	-5	-5	-5.2	-5.2
s	0.3	0.3	0.2	0.2
t	-0.1	-0.1	-0.1	-0.1

Table 4: Hellinger divergence shape alignment. The WDE recovers all of the transformation parameters exactly.

their goal was to determine a probabilistic atlas using Jensen-Shannon divergence. Here, we wish to showcase the wavelet density estimator for probabilistic shape matching, as in [30]. However, rather than using the Kullback-Leibler divergence measure as in [30], we elect to use the Hellinger divergence instead [9]:

$$\begin{aligned} D_{\mathrm{H}}(p_1, p_2) &= \int_{\mathbb{R}^2} (\sqrt{p_1} - \sqrt{p_2})^2 dx \\ &= 2 - 2 \sum_{j_0,k} \alpha^{(1)}_{j_0,k} \alpha^{(2)}_{j_0,k} \\ &\quad -2 \sum_{j \geq j_0,k}^{j_1} \beta^{(1)}_{j,k} \beta^{(2)}_{j,k} \end{aligned} \tag{30}$$

where $(\alpha^{(1)}, \beta^{(1)})$ and $(\alpha^{(2)}, \beta^{(2)})$ are the wavelet parameters of p_1 and p_2 respectively. The Hellinger divergence is also closely related to the geodesic distance on a sphere where each point on the sphere is a wavelet density. The advantage in using the Hellinger divergence (or the geodesic distance) over the Kullback-Leibler divergence is that the divergence is in closed form and does not need to be estimated from the data using a law of large numbers-based approach as in [8, 30]. In order to control experimental variability, we used a brute force coarse-to-fine search over the affine parameters. The target shape's density, p_1 in (30), is estimated once at the beginning using our method. At each iteration of the affine parameters, the source point set is deformed by the current affine parameters and a new p_2 is estimated with the wavelet density estimator using these transformed points. The Hellinger divergence error criterion in (30) is minimized when the two densities are best aligned and this in turn gives the optimal parameters of the affine transformation. Following a strategy similar to those described in the MI experiments, we were able to successfully recover the affine transformation. The ground truth affine parameters were the same as in the MI tests. In these experiments, the KDE—using fixed and variable bandwidths—again failed to estimate all of the parameters correctly (see Table 4). Some of the KDE's inaccuracies could be attributed to the fact that (30) is available in closed form under our representation but has to be numerically computed for the KDE. Figure 3(b) illustrates the source and target point set used in this matching experiment.

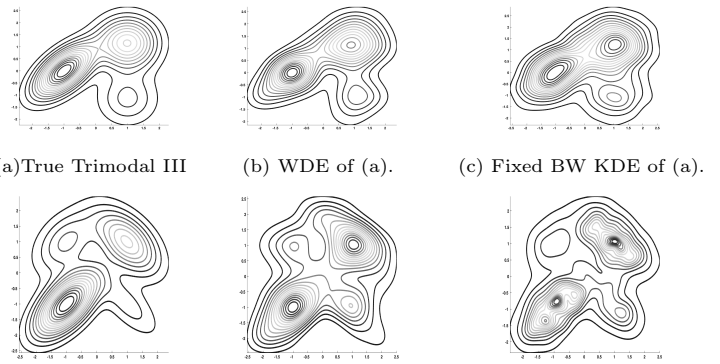
In this section, we have detailed both the approximation power and practical utility of our

proposed wavelet density estimator. The estimator is able to accurately represent a large class of parametric and non-parametric densities, a well-known trait of wavelet bases. Our method robustly satisfies the integrability and non-negativity constraints desired from density estimators with the added localization benefits inherent to wavelet expansions. This allowed us to seamlessly plug in our technique into several applications that critically depend on assessing densities from sample data.

5 Conclusions

In this paper, we have presented a new technique for non-negative, density estimation using wavelets. The non-negativity and unit integrability properties of bona fide densities are preserved through directly estimating $\sqrt{p}$ which allows one to obtain the desired density through the simple transformation $p = \left(\sqrt{p}\right)^2$. In sharp contrast to previous work, our method casts the estimation process in maximum likelihood framework. This overcomes some of the drawbacks of methods that require good pre-estimators for the density we are trying to find. The maximum likelihood setting consequently resulted in a constrained objective function whose minimization yielded the required basis function coefficients for our wavelet expansion. We were able to develop a efficient modified Newton method to solve the constrained problem by analyzing the relationship to the Fisher information matrix under the wavelet basis representation. We showed that the Hessian matrix at the solution point of the maximum likelihood objective function had a highly structured and simple form, allowing us to avoid matrix inverses typically required in Newton-type optimization. Verification of our proposed method was first empirically demonstrated by testing this method's capability to accurately reproduce known densities. Success was illustrated across a range of densities and wavelet families and validated against kernel density estimators. We also applied the estimation process to two image analysis problems: mutual information image registration and density estimation for point-set shape alignment. Both illustrated the successful operation of our method.

As a general density estimation procedure, this method could be applied to numerous applications. The compact, localized nature of wavelets and the large class of functions they are capable of representing make them an excellent choice for density estimation. Another property of particular interest to us is the regularity of wavelet families. We hope to explore this in future work, where we can take advantage of the differentiability characteristics to design and optimize information-theoretic objective functions typical in image analysis applications. It is also possible to investigate



(a)True Trimodal III (b) WDE of (a). (c) Fixed BW KDE of (a).

(d)True Quadrimodal (e) WDE of (d). (f) Variable BW KDE of (d).

Figure 5: 2D Density estimation comparison: WDE versus KDE contours. See Table 2 for WDE estimation parameters.

other optimization techniques to solve the present objective, such as preconditioned conjugate gradient, quasi-Newton or trust region methods. Finally, we plan to leverage the useful property that the Hellinger divergence and the geodesic distance between two wavelet densities (in the same family) are available in closed form in shape matching and indexing applications.

Acknowledgments

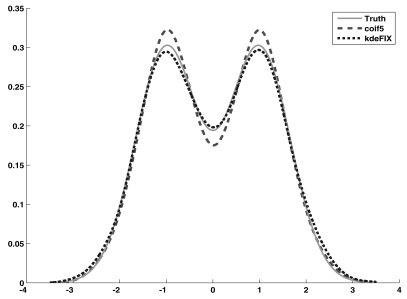
This work is partially supported by NSF IIS-0307712 and NIH R01NS046812. We want to thank Spiridon Penev and Gilbert Walter for literature material and helpful suggestions.

References

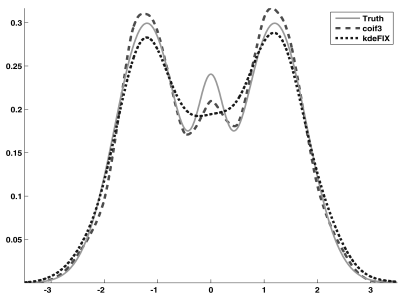
[1] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley-Interscience, 2001.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis.* Chapman and Hall/CRC, 1986.

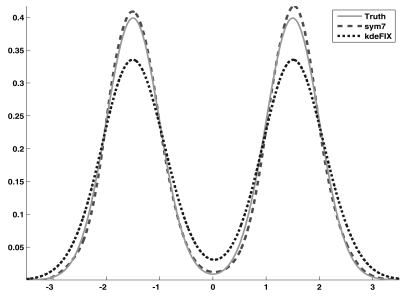
[3] P. Doukhan, "Formes de Töeplitz associées à une analyse multiechélle," *C.R. Acad. Sci. Paris Sér. A*, vol. 306, pp. 663–666, 1988.



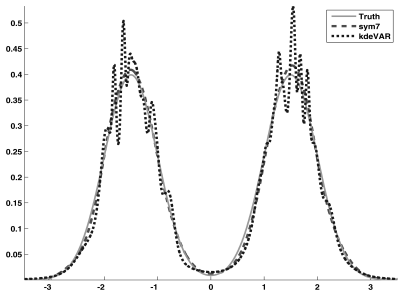
(a) Fixed BW KDE performs slightly better than our WDE on this bimodal density.



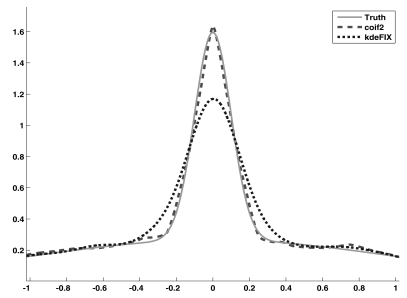
(b) The WDE captures the middle peak of this trimodal density. The fixed BW KDE misses it.



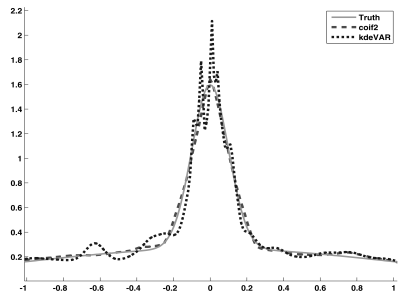
(c) Fixed BW KDE under estimates peaks of this separated bimodal density.



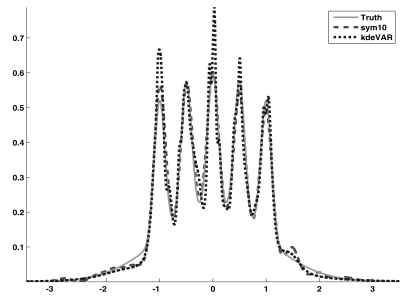
(d) Variable BW KDE has lower ISE than fixed BW KDE in (c) but incorrectly gives several peaks.



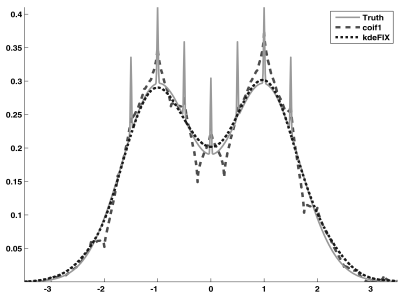
(e) WDE captures the main peak area of this kurtotic density, fixed BW KDE fails. (Zoomed around peak.)



(f) Variable BW KDE also fails on kurtotic density estimating several peaks. (Zoomed around peak.)



(g) The WDE captures all peaks of this claw density while the variable BW KDE over shoots peaks 1, 3 and 4.



(h) The fixed BW KDE has a lower ISE than WDE on this double claw density but it misses all the sharp peaks.

Figure 6: 1D Density estimation comparison: WDE versus KDE. True analytic density is solid line, WDE is dashed line and KDE is dotted line. See Table 1 for WDE estimation parameters.

[4] M. Vannucci, "Nonparametric density estimation using wavelets," ISDS, Duke University, Tech. Rep. DP 95-26, September 1995, available at http://www.isds.duke.edu.

[5] A. Pinheiro and B. Vidakovic, "Estimating the square root of a density via compactly supported wavelets," vol. 25, no. 4, pp. 399–415, 1997.

[6] S. Penev and L. Dechevsky, "On non-negative wavelet-based density estimators," *Journal of Nonparametric Statistics*, vol. 7, pp. 365–394, 1997.

[7] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[8] F. Wang, B. Vemuri, A. Rangarajan, I. Schmalfuss, and S. Eisenschenk, "Simultaneous nonrigid registration of multiple point sets and atlas construction," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 551–563.

[9] R. Beran, "Minimum Hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5, no. 3, pp. 445–463, 1977.

[10] I. Daubechies, *Ten Lectures on Wavelets*, ser. CBMS-NSF Reg. Conf. Series in Applied Math. SIAM, 1992.

[11] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1997.

[12] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: John Wiley and Sons, 1999.

[13] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24(2), pp. 508–539, 1996.

[14] W. Hardle, G. Kerkyacharian, D. Pickard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*, ser. Lecture Notes in Statistics 129. New York: Springer-Verlag, 1998.

[15] D. C. Montgomery, *Design and Analysis of Experiments*. Wiley, 2004.

[16] G. G. Walter, *Wavelets and Other Orthogonal Systems with Applications*. Boca Raton: CRC Press Inc., 1994.

[17] G. Lebanon, "Riemannian geometry and statistical machine learning," Ph.D. dissertation, Carnegie Mellon University, Pittsburg, PA, 2005.

[18] D. Luenberger, *Linear and Nonlinear Programming.* Reading, MA: Addison–Wesley, 1984.

[19] H. Burkhardt and N. Diehl, "Simultaneous estimation of rotation and translation in image sequences," *Proc. Eur. Signal Processing Conf.*, pp. 821–824, 1986.

[20] B. Vemuri, S. Huang, S. Sahni, C. M. Leonard, C. Mohr, R. Gilmore, and J. Fitzsimmons, "An efficient motion estimator with application to medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 79–98, March 1998.

[21] S. J. Marron and M. P. Wand, "Exact mean integrated squared error," *The Annals of Statistics*, vol. 20, no. 2, pp. 712–736, 1992.

[22] M. P. Wand and M. C. Jones, "Comparison of smoothing parameterizations in bivariate kernel density estimation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 520–528, 1993.

[23] G. G. Walter and J. K. Ghorai, "Advantages and disadvantages of density estimation with wavelets," *Computing Science and Statistics: Graphics and Visualization*, pp. 234–243, 1992.

[24] A. Ihler, "Kernel density estimation toolbox for matlab," 2003. [Online]. Available: http://ttic.uchicago.edu/∼ihler/code/kde.php

[25] A. Peter and A. Rangarajan, "Shape matching using the Fisher-Rao Riemannian metric: Unifying shape representation and deformation," *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1164–1167, 2006.

[26] T. Cootes and C. Taylor, "A mixture model for representing shape variation," in *Proceedings of British Machine Vision Conference*, 1997, pp. 110–119.

[27] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. Wiley, 2000, pp. 265–319.

[28] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, 1998.

[29] J. Zhang and A. Rangarajan, "Affine image registration using a new information metric," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 848–855.

[30] Y. W. K. Woods and M. McClain, "Information-theoretic matching of two point sets," *IEEE Trans. Image Processing*, vol. 11, no. 8, pp. 868–872, 2002.