

On Kernel Selection of Multivariate Local Polynomial Modelling and its Application to Image Smoothing and Reconstruction

Z. G. Zhang · S. C. Chan

Received: 2 October 2009 / Revised: 3 May 2010 / Accepted: 4 May 2010 / Published online: 30 May 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract This paper studies the problem of adaptive kernel selection for multivariate local polynomial regression (LPR) and its application to smoothing and reconstruction of noisy images. In multivariate LPR, the multidimensional signals are modeled locally by a polynomial using least-squares (LS) criterion with a kernel controlled by a certain bandwidth matrix. Based on the traditional intersection confidence intervals (ICI) method, a new refined ICI (RICI) adaptive scale selector for symmetric kernel is developed to achieve a better bias-variance tradeoff. The method is further extended to steering kernel with local orientation to adapt better to local characteristics of multidimensional signals. The resulting multivariate LPR method called the steering-kernel-based LPR with refined ICI method (SK-LPR-RICI) is applied to the smoothing and reconstruction problems in noisy images. Simulation results show that the proposed SK-LPR-RICI method has a better PSNR and visual performance than conventional LPR-based methods in image processing.

Keywords Adaptive kernel selection · Bias-variance tradeoff · Image processing · Multivariate local polynomial regression · Multidimensional signal processing · Kernel regression

This work was supported in part by the Hong Kong Research Grant Council (RGC). Parts of this work were presented at the Conference of 2009 IEEE International Symposium on Circuits and Systems [32].

Z. G. Zhang (✉) · S. C. Chan
Department of Electrical and Electronic Engineering,
The University of Hong Kong,
Pokfulam Road,
Hong Kong, China
e-mail: zgzhang@eee.hku.hk

1 Introduction

Local polynomial regression (LPR) is a flexible and efficient nonparametric regression method in statistics. LPR has been widely applied in many research areas such as data smoothing, density estimation, and nonlinear modelling [1, 2]. Given a set of noisy and possibly non-uniformly-spaced samples of a multidimensional signal, the data samples are fitted locally by a multivariate polynomial using the least-squares (LS) criterion with a kernel function $K_{\mathbf{H}}(\mathbf{u}) = \frac{1}{|\mathbf{H}|} K(\mathbf{H}^{-1}\mathbf{u})$, where $K(\mathbf{u}) : \mathbb{R}^d \rightarrow \mathbb{R}$ is a symmetric spherical multi-dimensional basis-kernel and \mathbf{H} is a certain bandwidth matrix to determine the kernel size and other properties.

LPR has been shown to possess an excellent spatial adaptation and reasonably simple implementation. Therefore, it has received considerable attention in the statistics and signal processing communities. Since signals may vary considerably, it is very crucial to choose a proper local kernel to achieve the best bias-variance tradeoff in estimating the local polynomial coefficients for modelling non-stationary signals. Generally, the optimal local kernel should be selected to minimize the mean squared error (MSE) or its integrated and smoothed versions, which are functions of the local kernel. For example, for slow varying parts of a one-dimensional signal, we would like the kernel to have a large scale so that more accurate estimates can be obtained by averaging out the additive noise as much as possible. At fast varying parts of the signal, however, we would like to have a small kernel scale so that excessive bias errors due to the limited order of the fitting polynomial will not occur.

For multivariate LPR, the bias-variance tradeoff problem is more complicated. It is because the bandwidth \mathbf{H} is a matrix in multidimensional signal processing. Not only is

the kernel scale important, but also the kernel shape has significant effects on the estimation results. It explains why there were not many publications concerning the bandwidth matrix selection of multivariate LPR, although the adaptive bandwidth selection for univariate LPR has been studied systematically and intensively [2–10].

Conventionally, the bandwidth matrix selection for multivariate LPR is simplified by using kernels with a specific shape, which is usually symmetric and spherical. More precisely, using a symmetric and spherical basis-kernel $K(\mathbf{u})$ and a bandwidth matrix $\mathbf{H}=h\mathbf{I}$ with h the scale parameter and \mathbf{I} the identity matrix, the kernel $K_{\mathbf{H}}(\mathbf{u})$ is simplified to $K_h(\mathbf{u}) = \frac{1}{h^d}K(\mathbf{u}/h)$, where d is the dimension of the multivariate LPR. As a result, the selection of the scale h is much easier than the selection of a d -dimensional bandwidth matrix \mathbf{H} . With this assumption, Ruppert et al. [10–12] studied the asymptotic bias, variance, and MSE of multivariate LPR and developed the empirical-bias bandwidths selection (EBBS) algorithm for multivariate LPR [13]. Based on Ruppert's work, Fan and Gijbels [2], Yang and Tschernig [14] also developed similar scale selection methods for multivariate local linear regression. These methods are innovative and useful, but they were developed mainly for linear or quadratic regression and did not consider the partial derivatives of the estimates. Also, they were all based on the symmetric kernel assumption.

Unlike above methods, Masry [15–17] considered a more general kernel, which may not be symmetric and spherical. More precisely, the basis-kernel $K(\mathbf{u})$ in the kernel function $K_h(\mathbf{u}) = \frac{1}{h^d}K(\mathbf{u}/h)$ may be asymmetric and of various shapes. A comprehensive study on the bias and variance of multivariate LPR with higher-order polynomials and higher-order derivatives was carried out, and an analytical expression of the optimal scale h was given. However, the optimal scale selection method is generally not directly usable in practice because its analytical form always involves quantities which are somewhat difficult to be estimated.

Instead of computing an optimal scale in an analytical form, empirical methods choose the scale from a finite set of scale parameters in a geometric grid. Recently, Katkovnik et al. [18–20] applied an intersection confidence intervals (ICI) method to the multivariate LPR problem for selecting the optimal local scale and studied its applications in image processing. Given a set of scale parameters in ascending order, the ICI rule determines the optimal scale by comparing the confidence intervals of the estimates with different scale parameters in the scale set. However, the ICI method (and other empirical scale selectors) can only select the local scale from the finite scale set, and therefore the exact scale value cannot be calculated if it is not included in the scale set.

In this paper, we first propose a new refined ICI adaptive scale selection method (RICI) for multidimensional signal

processing using the bias and variance expressions derived from multivariate LPR analysis of Masry [15–17]. The RICI method can determine a more accurate scale parameter beyond the limit of the finite scale set. The main idea of the RICI method is inspired by the work of Stanković [21], which discussed in-depth the bias-variance tradeoff problem and the ICI adaptive scale selection algorithm.

Next, we study the application of the proposed LPR with RICI adaptive scale selection method (LPR-RICI) for smoothing noisy images. Motivated by the iterative steering kernel regression method of Takeda et al. [22], the local kernel shape is iteratively determined from the local orientation of the image data. The locally adaptive steering kernel has a better adaptability to local characteristics of images than the conventional symmetric circular kernel. Coupled with the proposed RICI scale selection, a fully automatic iterative steering-kernel-based LPR with refined ICI method (SK-LPR-RICI) for smoothing multivariate data is proposed. It has a better MSE and perceptual equality than the conventional symmetric circular kernel. Since LPR can be naturally adapted to non-uniformly spaced data points, we also propose to employ the SK-LPR-RICI method for interpolating non-uniformly sampled images, which finds applications in image reconstruction and super-resolution.

The rest of the paper is organized as follows. In Section 2, the basic principle of multivariate LPR and its bias-variance tradeoff problem are revisited. Section 3 introduces the ICI scale selection algorithm and proposed the refined ICI adaptive scale selector. Section 4 is devoted to the applications of LPR to image processing problems such as smoothing and reconstruction, and a fully data-driven iterative steering-kernel-based LPR with refined ICI method will be proposed. This algorithm is also applicable to other 2-dimensional data. Experimental results and comparisons to other conventional methods using both test and real images are presented in Section 5. Finally, conclusions are drawn in Section 6.

2 Multivariate Local Polynomial Regression

In multivariate LPR, we have a set of multidimensional observations: (Y_i, \mathbf{X}_i) , $i=1,2,\dots,n$, where for each i , Y_i is a scalar response variable and $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,d})^T$ is a d -dimensional explanatory variable. We assume the homoscedastic model:

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\varepsilon_i \quad (1)$$

where $m(\mathbf{X}_i)$ is a smooth function specifying the conditional mean of Y_i given \mathbf{X}_i , ε_i is an independent identically distributed (i.i.d.) additive white Gaussian noise (AWGN)

with zero mean and unit variance, and $\sigma^2(\mathbf{X}_i)$ is its conditional variance. The problem is to estimate $m(\mathbf{X}_i)$ and its partial derivatives from the noisy samples Y_i .

The smooth function $m(\mathbf{X}_i)$ can be approximated locally, say at a given point $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$, by a multivariate polynomial of a certain degree:

$$\begin{aligned} m(\mathbf{X} : \mathbf{x}) &= m(\mathbf{x}) + \{\nabla m(\mathbf{x})\}^T (\mathbf{X} - \mathbf{x}) \\ &\quad + \frac{1}{2} (\mathbf{X} - \mathbf{x}) \{\mathcal{H}_m(\mathbf{x})\}^T (\mathbf{X} - \mathbf{x}) + \dots \\ &= m(\mathbf{x}) + \{\nabla m(\mathbf{x})\}^T (\mathbf{X} - \mathbf{x}) \\ &\quad + \frac{1}{2} \text{vec}^T \{\mathcal{H}_m(\mathbf{x})\} \cdot \text{vec} \{ (\mathbf{X} - \mathbf{x})(\mathbf{X} - \mathbf{x})^T \} + \dots \\ &= \beta_0 + \beta_{\nabla}^T \cdot (\mathbf{X} - \mathbf{x}) + \beta_{\mathcal{H}}^T \cdot \text{vech} \{ (\mathbf{X} - \mathbf{x})(\mathbf{X} - \mathbf{x})^T \} + \dots \end{aligned} \tag{2}$$

where $\beta_0 = m(\mathbf{x})$, $\beta_{\nabla} = \nabla m(\mathbf{x}) = \left[\frac{\partial m(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial m(\mathbf{x})}{\partial x_d} \right]^T$,

$$\beta_{\mathcal{H}} = \text{vech} \{ \mathcal{H}_m(\mathbf{x}) \} = \frac{1}{2} \left[\frac{\partial^2 m(\mathbf{x})}{\partial x_1^2}, 2 \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_2}, \dots, \frac{\partial^2 m(\mathbf{x})}{\partial x_d^2} \right]^T,$$

$\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ is in a neighborhood of \mathbf{x} , ∇^m is the $d \times 1$ gradient operator, \mathcal{H}_m is the $d \times d$ Hessian matrix of $m(\mathbf{x})$, $\text{vec}(\cdot)$ is the vectorization operator which converts a matrix into a column vector lexicographically, and $\text{vech}(\cdot)$ is the half-vectorization operator which converts, for example, $\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$ into $\text{vech}(\mathbf{A}) = [a_1, a_3, a_4]^T$.

Alternatively, a degree p multivariate polynomial in (2) can be written as:

$$m(\mathbf{X} : \mathbf{x}) = \sum_{\mathcal{K}=0}^p \sum_{k_1+\dots+k_d=\mathcal{K}} \beta_{k_1, \dots, k_d} \prod_{j=1}^d (X_j - x_j)^{k_j}, \tag{3}$$

where $\beta = \{ \beta_{k_1, \dots, k_d} : k_1 + \dots + k_d = \mathcal{K} \text{ and } \mathcal{K} = 0, \dots, p \}$ is the vector of coefficients with length $q = \prod_{j=1}^d [(d+j)/j]$.

Since ε_i is i.i.d. and Gaussian distributed, the maximum likelihood (ML) estimation of the coefficient vector β at location \mathbf{x} can be obtained by solving the following weighted least-squares (WLS) regression problem:

$$\widehat{\beta}(\mathbf{x}; \mathbf{H}) = \arg \min_{\beta} \{ J(\mathbf{x}; \mathbf{H}) \}, \tag{4}$$

where

$$J(\mathbf{x}; \mathbf{H}) = \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \left[Y_i - \sum_{\mathcal{K}=0}^p \sum_{k_1+\dots+k_d=\mathcal{K}} \beta_{k_1, \dots, k_d} \prod_{j=1}^d (X_{i,j} - x_j)^{k_j} \right]^2, \tag{5}$$

is a LS cost function and $K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) = \frac{1}{|\mathbf{H}|} K(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}))$ is a d -variate non-negative kernel function which emphasizes neighboring observations around \mathbf{x} in estimating $\beta(\mathbf{x})$. The bandwidth matrix \mathbf{H} determines the weights of neighboring

samples around \mathbf{x} to be used in estimating $\beta(\mathbf{x})$. For separable windows, we have

$$K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) = \prod_{j=1}^d K_{h_j}(X_{i,j} - x_j) = \frac{1}{h_1 \dots h_d} \prod_{j=1}^d K_{h_j} \left(\frac{1}{h_j} (X_{i,j} - x_j) \right), \tag{6}$$

Further, for identical bandwidth, $h_1 = \dots = h_d = h$. It is the most popular and simplest way to simplify the bandwidth parameter from a matrix \mathbf{H} to a scalar h .

In [15–17], the kernel considered is of the form $K_h(\mathbf{u}) = \frac{1}{h^d} K(\mathbf{u}/h)$, where the non-negative basis-kernel function $K(\mathbf{u})$ may not be circular or symmetric, and h is the scale parameter. To facilitate the development in the following sections, we assume that the basis-kernel $K(\mathbf{u})$ is known a priori. Its extension to a locally orientation adaptive basis-kernel will be discussed in Section 4.

The WLS solution to (4) is:

$$\widehat{\beta}(\mathbf{x}; h) = (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \mathbb{Y}, \tag{7}$$

where $\mathbb{X} = \begin{bmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T & \text{vech} \{ (\mathbf{X}_1 - \mathbf{x})(\mathbf{X}_1 - \mathbf{x})^T \} & \dots \\ 1 & (\mathbf{X}_2 - \mathbf{x})^T & \text{vech} \{ (\mathbf{X}_2 - \mathbf{x})(\mathbf{X}_2 - \mathbf{x})^T \} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \text{vech} \{ (\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T \} & \dots \end{bmatrix}$, and $\mathbb{W} = \text{diag} \{ K_h(\mathbf{X}_1 - \mathbf{x}), \dots, K_h(\mathbf{X}_n - \mathbf{x}) \}$,

$$\mathbb{Y} = [Y_1, Y_2, \dots, Y_n]^T$$

$K_h(\mathbf{X}_n - \mathbf{x})$ is the weighting matrix.

From the estimated coefficient vector $\widehat{\beta}(\mathbf{x}; h)$, we can obtain the estimated \mathbf{k} -th order derivative of $m(\mathbf{x})$ as follows

$$\widehat{m}^{(\mathbf{k})}(\mathbf{x}) = \frac{\partial^{\mathcal{K}} m(\mathbf{x})}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} = \prod_{j=1}^d (k_j!) \cdot \widehat{\beta}_{k_1, \dots, k_d}(\mathbf{x}; h), \tag{8}$$

where $\mathbf{k} = (k_1, \dots, k_d)$ is a vector of non-negative integers and $\mathcal{K} = \sum_{j=1}^d k_j$. In particular, $m(\mathbf{x})$ is estimated from $\widehat{\beta}_{0, \dots, 0}(\mathbf{x})$, that is $\widehat{m}(\mathbf{x}) = \widehat{m}^{(0, \dots, 0)}(\mathbf{x}) = \widehat{\beta}_{0, \dots, 0}(\mathbf{x})$.

As mention earlier, a key problem in LPR is to select an appropriate scale parameter h so as to achieve the best bias-variance tradeoff in estimating $m^{(\mathbf{k})}(\mathbf{x})$. To this end, we need first of all to examine the analytical expressions for the conditional bias, variance and MSE of the polynomial coefficient estimate $\widehat{\beta}$. From (7), it can be shown that:

$$\begin{aligned} \text{Bias}(\widehat{\beta}(\mathbf{x}; h) | \widetilde{\mathbf{X}}) &= E(\widehat{\beta}(\mathbf{x}; h)) - \beta(\mathbf{x}; h) \\ &= (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} (\mathbf{m} - \mathbb{X} \beta(\mathbf{x}; h)), \end{aligned} \tag{9}$$

$$\text{Var}(\widehat{\beta}(\mathbf{x}; h) | \widetilde{\mathbf{X}}) = (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \Sigma \mathbb{X} (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1}, \tag{10}$$

$$\text{MSE}(\widehat{\beta}(\mathbf{x}; h) | \widetilde{\mathbf{X}}) = \text{Bias}^2(\widehat{\beta}(\mathbf{x}; h) | \widetilde{\mathbf{X}}) + \text{Var}(\widehat{\beta}(\mathbf{x}; h) | \widetilde{\mathbf{X}}), \tag{11}$$

where $\tilde{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, $\mathbf{m} = [m(\mathbf{X}_1), m(\mathbf{X}_2), \dots, m(\mathbf{X}_n)]^T$, and $\Sigma = \text{diag}\{K_h(\mathbf{X}_i - \mathbf{x})\sigma^2(\mathbf{X}_i)\}$.

While the variance is not difficult to estimate from the observations and the kernel coefficients, the estimation of the bias poses a significant problem because it depends on the estimate $\hat{\beta}(\mathbf{x}; h)$ to be determined. Denote $\mathbf{k}! = \prod_{j=1}^d k_j!$, and let $f(\mathbf{x})$ be the probability density function (p.d.f.) of \mathbf{x} . The asymptotic bias and variance of the \mathbf{k} -th partial derivative $\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)$ are [15–17]:

$$\text{Bias}\{\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} = \sqrt{\mathbb{B}}h^{p+1-\mathcal{K}}, \tag{12}$$

$$\text{Var}\{\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} = \mathbb{V}h^{-(d+2\mathcal{K})}, \tag{13}$$

where $\mathbb{B} = (\mathbf{k}!(\mathbf{M}^{-1}\mathbf{B}\mathbf{m}_{p+1}(\mathbf{x}))_{\mathcal{K}})^2$, $\mathbb{V} = (\mathbf{k}!)^2\sigma^2(\mathbf{x})(\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1})_{\mathcal{K}, \mathcal{K}}/nf(\mathbf{x})$, $\mathbf{B} = [M_{0,p+1}, M_{1,p+1}, \dots, M_{p,p+1}]^T$, $\mathbf{m}_{p+1}(\mathbf{x}) = \text{vech}(\hat{m}^{(\mathbf{k})}(\mathbf{x}))$ is the regression result obtained with a higher order $p + 1$, $(\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1})_{\mathcal{K}, \mathcal{K}}$ is the $(\mathcal{K}, \mathcal{K})$ diagonal element of the constant matrix $\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1}$, $(\mathbf{M}^{-1}\mathbf{B}\mathbf{m}_{p+1}(\mathbf{x}))_{\mathcal{K}}$ is the \mathcal{K} -th element of the vector $\mathbf{M}^{-1}\mathbf{B}\mathbf{m}_{p+1}(\mathbf{x})$, and the (i, j) -th element of the matrices \mathbf{M} and Γ are $M_{i,j} = \int \mathbf{u}^{(i+j)}K(\mathbf{u})d\mathbf{u}$ and $\Gamma_{ij} = \int \mathbf{u}^{(i+j)}K^2(\mathbf{u})d\mathbf{u}$ respectively. The details can be found in [15–17]. Similar asymptotic analytical results on bias and variance of multivariate LPR were also reported by Ruppert and Ward [10], and Fan and Gijbels [2].

It can be seen from (12) and (13) that the bias term will increase with the increase of the scale parameter h since that the underlying data at \mathbf{x} cannot be modeled well by the local polynomial with the given order. On the other hand, a larger scale parameter will reduce the variance of the estimator. Hence, there exists a locally optimal scale $h^{opt}(\mathbf{x})$ for estimating the \mathbf{k} -th partial derivative $m^{(\mathbf{k})}(\mathbf{x})$, and $h^{opt}(\mathbf{x})$ should minimize an appropriate measure such as the MSE as follows:

$$\begin{aligned} \text{MSE}\{\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} &= \text{Bias}^2\{\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} + \text{Var}\{\hat{m}^{(\mathbf{k})}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} \\ &= \mathbb{B}h^{2(p+1-\mathcal{K})} + \mathbb{V}h^{-(d+2\mathcal{K})}. \end{aligned} \tag{14}$$

By minimizing (14), an analytical formula for $h^{opt}(\mathbf{x})$ can be determined as follows:

$$h^{opt}(\mathbf{x}) = \left(\frac{(2\mathcal{K} + d)\mathbb{V}}{2(p - \mathcal{K} + 1)\mathbb{B}}\right)^{\frac{1}{d+2p+1}}. \tag{15}$$

However, some of the quantities in (15) are difficult to be calculated, which makes the optimal scale difficult to be estimated accurately. Instead of computing an optimal scale in an analytical form, empirical methods choose the local scale from a finite set of scale parameters in a geometric grid. In the next section, an empirical scale selection method, called the intersection of confidence interval rule is introduced.

3 Refined ICI Scale Selection Algorithm

The intersection of confidence interval (ICI) rule is an empirical method to select the adaptive scale from a set of scales in ascending order. The ICI rule was proposed by Goldenshluger and Nemirovski in [23], and Katkovnik et al. [18–20] applied the ICI rule to univariate and bivariate LPR for selecting the locally adaptive bandwidth or scale. The ICI method can be regarded as a special implementation of the Lepski’s approach [24], which compared a series of estimates with different scales from the scale set and the optimal scale was the largest one whose estimate did not differ significantly from the estimate of a smaller scale.

Starting with a finite set of scale parameters in a geometric grid,

$$\tilde{\mathbf{H}} = \{h_j\mathbf{I}|h_j = h_0a^j, j = 1, \dots, J\}, \tag{16}$$

where $a > 1$ is a step factor, $h_0 > 0$ is the base scale, and J is the size of the scale set. The minimum scale in the scale set should be small enough so that sharp changes can be preserved, while the maximum scale should be large enough to remove the additive noise adequately. The maximum scale is usually chosen as $n/\log_{10}(n)$ to avoid high complexity and possible over-smoothing [2].

3.1 ICI Rule

In the ICI rule, the optimal scale is determined by comparing the confidence intervals of the estimates with different scales from $\tilde{\mathbf{H}}$. The ICI rule needs only to estimate the variance, which can be easily estimated from (10), and it does not involve the computation of the bias and MSE values.

To see how the algorithm works, we first note that, for each \mathbf{x} , the estimate $\hat{m}(\mathbf{x}; h)$ is a random variable around the true value $m(\mathbf{x})$ with bias and variance as shown in (12) and (13). For simplicity, in this section we shall drop the index \mathbf{x} and \mathbf{k} to save notation. It has been shown in [2, 20, 21, 23] that

$$\{\hat{m}(h) - [m + \text{Bias}(\hat{m}(h))]\}/\sqrt{\text{Var}(\hat{m}(h))} \sim \mathcal{N}(0, 1) \tag{17}$$

is Gaussian distributed with zero mean and unit variance. In other words, $\widehat{m}(h)$ is Gaussian distributed with mean $Mean(\widehat{m}(h)) = m + Bias(\widehat{m}(h))$ and standard deviation $Std(\widehat{m}(h)) = \sqrt{Var(\widehat{m}(h))}$.

The confidence intervals of the estimate $\widehat{m}(h)$ can also be constructed from the bias and variance as

$$[Mean(\widehat{m}(h)) - \kappa \cdot Std(\widehat{m}(h)), Mean(\widehat{m}(h)) + \kappa \cdot Std(\widehat{m}(h))]. \tag{18}$$

As a result, the probability that $\widehat{m}(h)$ lies in the confidence interval $|\widehat{m}(h) - Mean(\widehat{m}(h))| \leq \kappa \cdot Std(\widehat{m}(h))$, $Pr(\kappa)$, can be computed readily. Usually, the parameter κ is selected so that $Pr(\kappa)$ will be close to one and a reasonable estimation of the possible range of values of $\widehat{m}(h)$ can be achieved. For example, typical values are $Pr(1.65)=0.90$, $Pr(1.96)=0.95$, and $Pr(2.58)=0.99$. A larger κ will lead to a wider confidence interval and thus a smaller bandwidth selected, while a smaller κ will obtain a larger ICI-selected bandwidth. Instead of choosing a fixed and appropriate value for κ , it is also possible to optimize this value by assigning a local threshold κ to each locations. In [18], Katkovnik used the generalized cross-validation (GCV) to determine the value of κ for each \mathbf{x} . It is also based on an empirical method, where the desired value of κ is chosen from a finite set of threshold parameters $\Gamma = \{\kappa_1, \kappa_2, \dots, \kappa_K\}$. Each threshold parameter $\kappa \in \Gamma$ is used in the ICI rule to calculate a series of adaptive estimates. The optimal κ is chosen as the one which minimizes the GCV criterion:

$$\kappa^{opt}(\mathbf{x}) = \arg \min_{\kappa} \sum_i \left(\frac{Y_i - \widehat{m}(\mathbf{X}_i; h_{j^+|\kappa}(\mathbf{x}))}{1 - K_{h_{j^+|\kappa}(\mathbf{x})}(\mathbf{X}_i - \mathbf{x})} \right)^2, \tag{19}$$

where $h_{j^+|\kappa}(\mathbf{x})$ is the optimal scale with the threshold parameter κ . This method may lead to a better result at the expense of significantly higher computational complexity. It is because we need to compute a $\widehat{m}(h_j)$ for each $\kappa \in \Gamma$ and each $h_j \in \mathbf{H}$, not to mention the GCV computation and comparison.

The mean value of $\widehat{m}(h)$ is difficult to be estimated since it involves the true value m . The standard deviation of $\widehat{m}(h)$, $Std(\widehat{m}(h))$, can be computed with the help of (10). The quantity $\sigma^2(\mathbf{x})$ in (10) is estimated from the homoscedasticity assumption of the additive noise. The robust noise variance estimator given in [18–20] can be used:

$$\widehat{\sigma} = \{median(|Y_i - Y_{i-1}|)\} / (\sqrt{2} \cdot 0.6745). \tag{20}$$

where $median(\cdot)$ denotes the sample median operation.

Next, consider a series of confidence intervals $D_j = [L_j, U_j]$ from the estimated $\widehat{m}(h_j)$ with different values of scale h_j from the scale set \mathbf{H} :

$$U_j = \widehat{m}(h_j) + (\kappa + \Delta\kappa) \cdot Std(\widehat{m}(h_j)), \tag{21}$$

$$L_j = \widehat{m}(h_j) - (\kappa + \Delta\kappa) \cdot Std(\widehat{m}(h_j)), \tag{22}$$

where $\Delta\kappa$ is a threshold parameter used to adjust the width of the confidence interval. We notice that the confidence intervals (21) and (22) are biased because the estimated value $\widehat{m}(h_j)$ is used instead of the true value $m(\mathbf{X}_i)$. Another constant $\Delta\kappa$ is introduced to reduce the influence of the estimation bias, and its effect and selection will be discussed later in the refined ICI method.

For a small value of h , we shall expect the bias will be small and the confidence interval will gradually decrease with increasing value of h while the center of the interval remains more or less fixed. When h is increased to a point where the observations cannot be satisfactorily modeled, a large bias will result and the center of the interval will change significantly with respect to the others, while its length will still be small. As a result, the confidence interval will no longer intersect those with smaller values of h .

Motivated by this observation, the ICI scale selection method computes and examines the following quantities from the confident intervals in order to detect this sudden change:

$$\begin{aligned} \bar{L}_j &= \max[\bar{L}_{j-1}, L_j], \text{ for } j = 1, 2, \dots, J \\ \underline{U}_j &= \min[\underline{U}_{j-1}, U_j], \text{ for } j = 1, 2, \dots, J \\ \bar{L}_0 &= \underline{U}_0 = 0. \end{aligned} \tag{23}$$

It can be seen that \bar{L}_j is the largest upper bound of the confidence interval for scale evaluated up to h_j , while \underline{U}_j is the corresponding lower bound. The largest value of these j for which $\underline{U}_j \geq \bar{L}_j$, denoted by j^+ , gives the desirable scale h_{j^+} , because above which the confidence intervals no longer intersect with each other. In the ICI rule, it has been shown in [21] that if the two confidence intervals D_i and D_j intersects, then:

$$|\widehat{m}(h_i) - \widehat{m}(h_j)| \leq (\kappa + \Delta\kappa) \cdot [Std(\widehat{m}(h_i)) + Std(\widehat{m}(h_j))]. \tag{24}$$

3.2 Refined ICI Method

We now consider the selection of $\Delta\kappa$ and give a refined version of the conventional ICI rule. From (21), (22), and (24), it can be seen that the threshold $\Delta\kappa$ can significantly affect the performance of the ICI algorithm. When $\Delta\kappa$ is large, the confidence interval D_j becomes wide, and it may lead to a large value of h_{j^+} . This will result in over-smoothing of the observations. On the contrary, when $\Delta\kappa$ is small, the interval D_j will become narrow, and it may result in a very small value of h_{j^+} so that the additive noise cannot be removed effectively.

The selection and adjustment of $\Delta\kappa$ have been discussed in [20] and [23]. By substituting the optimal scale (15) into

the asymptotic bias and variance expressions in (12) and (13), we have:

$$\frac{Bias^2\{\widehat{m}^{(k)}(\mathbf{x}; h^{opt})|\widetilde{\mathbf{X}}\}}{Var\{\widehat{m}^{(k)}(\mathbf{x}; h^{opt})|\widetilde{\mathbf{X}}\}} = \frac{(2\mathcal{K} + d)}{2(p - \mathcal{K} + 1)}. \tag{25}$$

To simplify notations, let

$$b_m(h) = Bias\{\widehat{m}(h)|\widetilde{\mathbf{X}}\}, \tag{26}$$

$$\sigma_m^2(h) = Var\{\widehat{m}(h)|\widetilde{\mathbf{X}}\}, \tag{27}$$

$$\beta = 2(p + 1 - \mathcal{K}), \text{ and } \nu = 2\mathcal{K} + d. \tag{28}$$

Then, one gets $b_m^2(h) = \mathbb{B}h^\beta$, $\sigma_m^2(h) = \mathbb{V}h^{-\nu}$, and $MSE(h) = \mathbb{B}h^\beta + \mathbb{V}h^{-\nu}$. The optimal scale can then be written as

$$h^{opt} = \left(\frac{\beta}{\nu} \cdot \frac{\mathbb{V}}{\mathbb{B}}\right)^{\frac{1}{\beta+\nu}}. \tag{29}$$

In [23], $\Delta\kappa$ is chosen as κ because the bias and standard deviation are assumed to be equal. In [20], Katkovnik showed that $\Delta\kappa = (\beta/\nu)\kappa$ was an appropriate choice because it can minimize the upper bound of the estimation error. That is, $\Delta\kappa$ should be selected as the ratio of the squared bias and the variance at the optimal scale h^{opt} .

It can be seen that the above method can only select the optimal scale $h^{opt} = h_{j^+}$ from the finite scale set $\widetilde{\mathbf{H}}$. Generally, the exact optimal scale is not included in the scale set $\widetilde{\mathbf{H}}$. In [21], the computation of a more exact optimal scale for the general ICI scale selection problem with given bias and variance formulas was considered. The importance of the result is that explicit knowledge of \mathbb{B} and \mathbb{V} in (12) and (13) is not required. Here, we shall employ this method to further refine the ICI adaptive scale for the LPR problem using the bias and variance formulas in (12) and (13).

By assuming that $h^{opt}\mathbf{I}$ is close to the conventional ICI-based optimal bandwidth $h_{j^+}\mathbf{I} \in \widetilde{\mathbf{H}}$, we can write $h_{j^+} = a^\eta h^{opt}$, where a is the step factor and η is a variable to be determined for refining the optimal scale. The entries in the scale set $\widetilde{\mathbf{H}}$ of (16) can thus be rewritten as

$$\widetilde{\mathbf{H}} = \left\{ h_j \mathbf{I} | h_j = h_{j^+} a^{(j-j^+)} = h^{opt} a^{(j-j^++\eta)}, j = 1, \dots, J \right\}. \tag{30}$$

Substituting (30) into (12) and (13), the bias and standard deviation can be written as:

$$b_m(h_j) = \sqrt{\mathbb{B}h_j^\beta} = \sqrt{\nu/\beta} \sigma_m(h^{opt}) a^{\frac{\beta(j-j^++\eta)}{2}}, \tag{31}$$

$$\sigma_m(h_j) = \sqrt{\mathbb{V}/h_j^\nu} = \sigma_m(h^{opt}) a^{-\frac{\nu(j-j^++\eta)}{2}}. \tag{32}$$

In the ICI rule, the optimal h_{j^+} is the largest scale where the confidence intervals D_{j^+-1} and D_{j^+} intersect, but not D_{j^+} and D_{j^++1} . Given an appropriate value of κ such that $\Pr(\kappa) \approx 1$. From the p.d.f. of (17), we have high confidence that $\widehat{m}(h_j)$ will lie in the confidence interval:

$$m + b_m(h_j) - \kappa\sigma_m(h_j) \leq \widehat{m}(h_j) \leq m + b_m(h_j) - \kappa\sigma_m(h_j). \tag{33}$$

Making use of (33) and substituting (16) to (24), one gets, after slight manipulation, the condition for D_{j^+-1} and D_{j^+} to intersect as:

$$b_m(h_{j^+-1}) + \Delta\kappa\sigma_m(h_{j^+-1}) \geq b_m(h_{j^+}) + \Delta\kappa\sigma_m(h_{j^+}), \tag{34}$$

and the condition for D_{j^+} and D_{j^++1} not to intersect as:

$$b_m(h_{j^+}) + (2\kappa + \Delta\kappa)\sigma_m(h_{j^+})b_m(h_{j^++1}) - (2\kappa + \Delta\kappa)\sigma_m(h_{j^++1}). \tag{35}$$

Substituting the estimated bias (31) and variance (32) into the inequalities (34) and (35), the threshold parameter $\Delta\kappa$ and the refining parameter η in the expressions can be determined as follows:

$$\Delta\kappa = \frac{2\kappa}{a^{(\beta+\nu)/2} - 1}, \tag{36}$$

$$\eta = \frac{2}{\beta + \nu} \log_a \left(\Delta\kappa \sqrt{\frac{\beta}{\nu} \frac{1 + a^{\nu/2}}{1 - a^{-\beta/2}}} \right). \tag{37}$$

Thus, given an appropriate value of the threshold parameter κ , say 1.96, in LPR problem, $\Delta\kappa$ and η can be obtained from (36) and (37). Hence, a refined adaptive scale is $h^{opt} = h_{j^+} a^{-\eta}$.

Since the scale set is constructed in a geometric order, we still need to add a logarithmic shift $\Delta\eta$ to η for compensating for the difference between the arithmetic mean and the geometric mean. This shift value can be approximated as [21]:

$$\Delta\eta = \frac{2[\log_a((1 + a^{(\beta+\nu)/2})/2)]}{\beta + \nu} - \frac{1}{2}. \tag{38}$$

Finally, the desired optimal scale is determined as:

$$h^{opt} = h_{j^+} a^{-(\eta+\Delta\eta)}. \tag{39}$$

We now summarize the refined ICI (RICI) adaptive scale selection method as follows:

- 1) Set up a finite scale set $\widetilde{\mathbf{H}}$ as in (16);
- 2) Compute $\widehat{m}(h_j)$ for $h_j \mathbf{I} \in \widetilde{\mathbf{H}}$;
- 3) Select the threshold κ and compute $\Delta\kappa$ from (36);

- 4) Find h_{j^+} from the set $\tilde{\mathbf{H}}$ with the ICI rule;
- 5) Compute the refined parameters η and $\Delta\eta$ from (37) and (38);
- 6) Calculate the refined optimal scale $h^{opt} = h_{j^+} a^{-(\eta+\Delta\eta)}$.

Note that in step 3), the threshold parameter κ can be chosen as some reasonable values such as 1.96. If κ is selected by (19), the result may be better but the complexity is higher.

4 Data-Driven LPR for Image Processing

Recently, bivariate LPR has received considerable attention in image processing field. Katkovnik et al. [18–20] successfully applied LPR with ICI bandwidth selection to de-noising, de-blurring, and edge detection of images. Takeda et al. developed a steering kernel regression method for image de-noising, interpolation, and fusion [22]. In this section, we shall extend the adaptive scale selection method developed in Section 3 to kernel that can adapt to the local orientation of an image. This steering-kernel-based LPR-RICI (SK-LPR-RICI) will then be applied to smoothing and interpolation of noisy images.

An $N_1 \times N_2$ gray-scale image can be viewed as the observations of a 2-dimensional (2D) function where the samples are taken along the 2D explanatory variable $\mathbf{x} = (x_1, x_2)$ at $x_1 = 1, 2, \dots, N_1$ and $x_2 = 1, 2, \dots, N_2$ with $Y(x_1, x_2)$ the intensity of the image pixel. In principle, the locations of the samples can be non-uniformly spaced. This makes LPR very useful in general interpolation and smoothing problems. The smoothed result is obtained from the estimated $\hat{m}(x_1, x_2) = \beta_{0,0}$ by applying LPR with a suitably chosen 2D kernel $K_{\mathbf{H}}(\cdot)$ to the original image. The quantities $\beta_{1,0} = \partial m(x_1, x_2) / \partial x_1$ and $\beta_{0,1} = \partial m(x_1, x_2) / \partial x_2$ represent respectively the first-order partial derivatives of an image along the x_1 and x_2 coordinates. When \mathbf{x} is evaluated at a series of 2D grid points, we obtain a smoothed surface representation of the data from the noisy image Y . Therefore, the smoothing and interpolation of digital images can be performed through LPR.

4.1 Kernel Selection

The Gaussian-type kernel is employed in our analysis work and experiments on image processing because it can be transformed to produce a steering kernel. Given a bandwidth matrix \mathbf{H} , the Gaussian kernel is calculated as:

$$\mathbf{K}_{\mathbf{H}}(\mathbf{u}) = \frac{1}{2\pi|\mathbf{H}|} \exp\left\{-\mathbf{u}^T (\mathbf{H}^{-1})^T \mathbf{H}^{-1} \mathbf{u} / 2\right\} \quad (40)$$

Since the Gaussian kernel is not of compact support, it should be truncated to a sufficient kernel size $N_k \times N_k$ to

reduce the arithmetic complexity, which will be discussed later in this section.

When the bandwidth matrix is chosen as $\mathbf{H} = h\mathbf{I}$, a symmetric Gaussian kernel results, which is used in most LPR methods for image processing because of its simplicity. For the symmetric basis-kernel, the locally adaptive scale parameter h at a point \mathbf{x} will control the amount of smoothing to be performed at that point. When h is small, image details such as edges and textures will be preserved. However, possible additive noise may not be removed effectively. On the contrary, a large-scale kernel has better de-noising properties at the expense of possibly blurring of the image details. Therefore, a locally adaptive scale selection method is very crucial to LPR in image processing. The ICI rule with a symmetric circular kernel has been shown to be an effective tool for adaptive scale selection [18–20].

Recently, Takeda et al. [22] proposed a steering kernel, which is obtained by elongating and rotating a symmetric Gaussian kernel so that it can adapt to the dominant orientation locally. In [22], the orientation and scale of the steering kernel was determined using principal components analysis (PCA). The resultant iterative steering kernel regression (ISKR) method was shown to have a better performance than the conventional symmetric kernel, especially along image edges. In the ISKR method, the local scaling parameter was obtained as $h_i = h_0 \gamma_i$, where h_0 and γ_i are respectively the global smoothing parameter and the local scaling parameter. The local scaling factor γ_i was estimated using the local energy distribution of the image, while h_0 was chosen as to minimize the root mean squared error (RMSE) value for known test images or to yield most visually appealing results for real images. Consequently, the scale selection process is not fully automatic and the criterion for selecting the global smoothing parameter for real images is somewhat subjective.

To address this problem, we propose a fully data-driven adaptive scale selection method using the RICI rule proposed in Section 3. Specifically, we will extend the RICI rule for symmetric kernel to the steering kernel and obtain a new steering-kernel-based LPR with refined ICI bandwidth selection (SK-LPR-RICI). To this end, we propose to construct the local steering kernel $K_h(\mathbf{u}) = \frac{1}{h^2} K(\mathbf{u}/h)$ at each sample \mathbf{x} of an image by a steering basis-kernel $K(\mathbf{u})$ and a scale parameter h . The shape of the local basis-kernel $K(\mathbf{u})$ will be determined first based on the local dominant orientation estimation around \mathbf{x} which can be achieved by the PCA method in [22]. The details of the PCA method is omitted to save place, and please be referred to [22] for details. Then, the RICI rule is employed to choose the optimal scale parameter h for the steering basis-kernel $K(\mathbf{u})$.

4.2 Steering-Kernel-Based LPR with RICl Method

Once the locally adaptive steering basis-kernel is determined, the RICl adaptive scale selection method can be used with slight modification to determine the locally adaptive scale parameter h . More precisely, with different scale parameter h_j from the scale set $\tilde{\mathbf{H}}$, we use the steering kernel $K_{h_j}(\mathbf{X} - \mathbf{x}; \mathbf{H}_{S_j}) = \frac{1}{h_j^2} K((\mathbf{X} - \mathbf{x})/h_j; \mathbf{H}_{S_j})$ with a bandwidth matrix $h_j \mathbf{H}_{S_j}$ to fit a p -order polynomial at \mathbf{x} and obtain a series of smoothed estimates of $\hat{m}(\mathbf{x}; h_j)$. The traditional ICi rule is then invoked to select a preliminary scale parameter h_{j^*} . For $d=2$ and $\mathbf{k}=(0,0)$, the bias and variance of the estimator are

$$\text{Bias}\{\hat{m}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} = \sqrt{\mathbb{B}}h^{p+1}, \quad (41)$$

$$\text{Var}\{\hat{m}(\mathbf{x}; h)|\tilde{\mathbf{X}}\} = \mathbb{V}h^{-2}. \quad (42)$$

The refined scale h^{opt} can then be determined by the RICl method. A LPR using this refined scale parameter is then performed on the image to yield the final estimate $\hat{m}(\mathbf{x}; h^{opt}) = \hat{\beta}_{0,0}(\mathbf{x}; h^{opt})$.

Inspired by the ISKR method, the local steering matrix estimation and LPR with RICl rule can be executed iteratively to achieve better de-noising results, if necessary. The proposed iterative SK-LPR-RICl method can be summarized as follows:

Step 1 Initial gradient estimation

With an initial bandwidth matrix $\mathbf{H}^{(0)} = h^{(0)}\mathbf{I}$, calculate the initial values of $\hat{m}^{(0)}(\mathbf{x}; \mathbf{H}^{(0)})$ and the gradient estimation $\hat{\beta}_{0,1}^{(0)}(\mathbf{x}; \mathbf{H}^{(0)})$ and $\hat{\beta}_{1,0}^{(0)}(\mathbf{x}; \mathbf{H}^{(0)})$.

Step 2 Steering kernel estimation (l-th iteration)

The first-order gradient estimates $\hat{\beta}_{0,1}^{(l-1)}(\mathbf{x}; \mathbf{H}^{(l-1)})$ and $\hat{\beta}_{1,0}^{(l-1)}(\mathbf{x}; \mathbf{H}^{(l-1)})$ are used to calculate the steering bandwidth matrix $\mathbf{H}_{S,\mathbf{x}}^{(l)}$ and the steering basis-kernel $K_{\mathbf{x}}^{(l)}(\mathbf{u}; \mathbf{H}_{S,\mathbf{x}})$.

Step 3 Local scale parameter estimation (l-th iteration)

Obtain a series of LPR results using kernels with the bandwidth matrix $h \cdot \mathbf{H}_{S,\mathbf{x}}^{(l)}$ with h belonging to $\tilde{\mathbf{H}}$. Determine the scale parameter $h^{opt,(l)}$ using the RICl rule.

Step 4 LPR with estimated steering kernel (l-th iteration)

A LPR is performed with the estimated bandwidth matrix $\mathbf{H}^{(l)} = h^{opt,(l)}\mathbf{H}_{S,\mathbf{x}}^{(l)}$ to obtain $\hat{m}^{(l)}(\mathbf{x}; \mathbf{H}^{(l)})$, along with its partial derivatives $\hat{\beta}_{1,0}^{(l)}(\mathbf{x}; h^{opt,(l)})$ and $\hat{\beta}_{0,1}^{(l)}(\mathbf{x}; h^{opt,(l)})$.

Generally, more iterations will reduce the estimation variance, but the bias will increase. As a result, when the additive noise is small, the number of iterations should also be small to avoid over-smoothing for the image. When the image is contaminated by large amount of noise, more iterations can improve the de-noising results. Similarly,

given prior knowledge of additive noise, the initial scale parameter $h^{(0)}$ should be selected as a large value when noise component is significantly large, while a small value of $h^{(0)}$ is more appropriate when the additive noise is small.

The computational complexity of the SK-LPR-RICl method is mainly determined by the polynomial order p and the kernel size N_K . For each data sample, the complexities of computing $\mathbb{X}^T \mathbb{W} \mathbb{X}$ and $\mathbb{X}^T \mathbb{W} \mathbb{Y}$ in (7) are $O(q^2 N_K^2)$ and $O(q N_K^2)$ respectively, and the matrix inversion needs a complexity of $O(q^3)$, where $q = \prod_{j=1}^p [(d+j)/j]$ is the length of the coefficient β . If $q \ll N_K$, the complexity of the WLS solution is $O(N_K^2)$. As a result, a large p , which means a large q , will lead to a very heavy computational complexity. Hence, a linear ($p=1$) regression is generally used, and the kernel size N_K is given a small value of, say 9 or 11. Since the SK-LPR-RICl method involves the SVD, which has a complexity of $O(N_K^3)$, it has a higher arithmetic complexity than the symmetric-kernel-based LPR. As for the ISKR method, if there is no *a priori* knowledge about the global smoothing parameter h_0 , each h_0 from $\tilde{\mathbf{H}}$ has to be tested and more importantly the quality has to be judged manually. Hence, the complexity of ISKR and the proposed iterative SK-LPR-RICl method are comparable.

4.3 LPR for Image Smoothing and Reconstruction

Since LPR is applicable to uniformly as well as non-uniformly sampled observations, it is very useful in smoothing and interpolation of generally sampled data set. One useful application is the reconstruction of a high-resolution (HR) image from multiple shifted low-resolution (LR) images in super-resolution imaging [25–28]. Since the LR images may be shifted arbitrarily from each other, the image samples are non-uniformly sampled. Using LPR, we can construct a HR image from the non-uniform spaced image samples. More information about super-resolution and image registration can be found in [25–30]. Here, we only focus on the application of the SK-LPR-RICl method to the HR imaging from LR images without blurring. The image samples may be corrupted by noise due to imprecise registration.

As shown before, LPR is a practical de-noising and interpolation technique for image processing. Therefore, it can be used to interpolate noisy uniformly spaced images and remove possible artifacts due to data compression. This will be illustrated further by examples in Section 5. However, it is worth noting that the proposed LPR-based methods are only capable of handling the artifacts that can be modeled as an additive Gaussian white noise. It will be interesting to improve the LPR-based methods to deal with other unwanted effects like illumination or blurring, but it is outside the scope the paper and we will pursue such direction in further study.

5 Experimental Results

In this section, we will evaluate the proposed RICl method for image smoothing, coding artifacts removal, and super-resolution imaging. Peak signal-to-noise ration (PSNR), $PSNR = 10 \cdot \log_{10}(255^2/MSE)$, is used here as the performance criterion for evaluation. The PSNR values shown in the tables are averages of 100 independent Monte-Carlo realizations, and one realization is illustrated in figures.

5.1 Image De-Noising

We now apply the proposed SK-LPR-RICl method to suppress the additive Gaussian white noise in gray-scale images. Two 256×256 8-bit images, *Test* and *Lena* are used in the simulation, and the noise-free original images are shown in Figs. 1a and 2a. The image *Test* contains distinct edges and corners, and the performances around image edges are easier to be visualized and compared. *Lena* also contains a good mixture of details, flat area, and

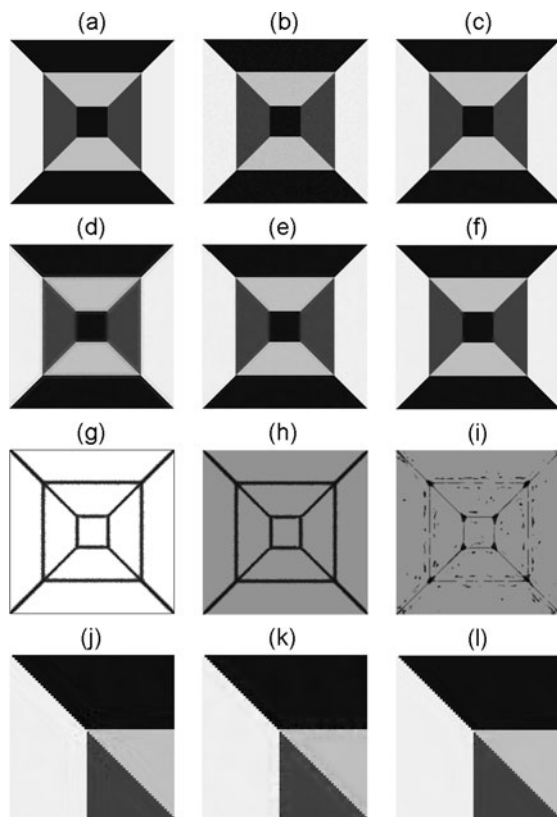


Figure 1 Comparison of different adaptive scale selection methods for smoothing the image *Test*: **a** original noise-free image, **b** noisy image, **c** ISKR, **d** LPR-ICI, **e** LPR-RICI, **f** SK-LPR-RICI, **g** adaptive local scale with ICI rule, **h** adaptive local scale with RICl rule, **i** adaptive local scale with RICl rule based on steering kernel, **j** enlarged image of ISKR, **k** enlarged image of LPR-RICI, **l** enlarged image of SK-LPR-RICI.

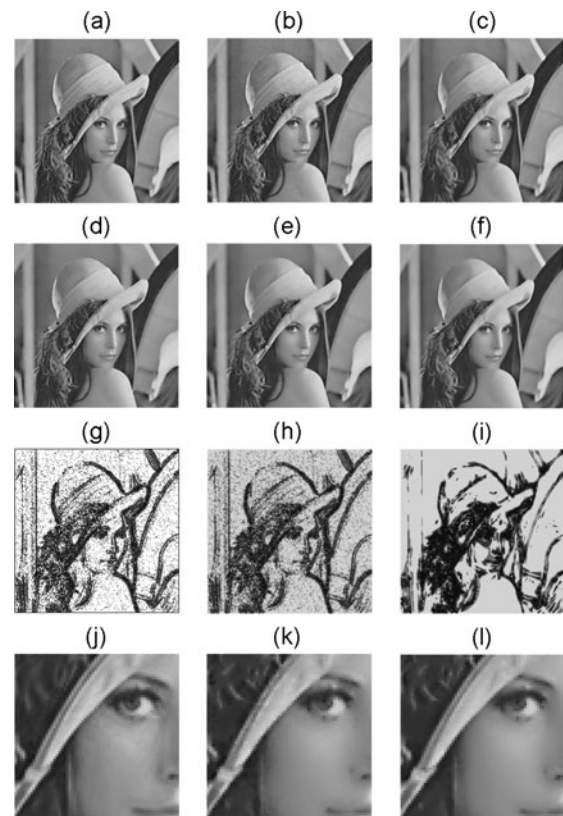


Figure 2 Comparison of different adaptive scale selection methods for smoothing the image *Lena*: **a** original noise-free image, **b** noisy image, **c** ISKR, **d** LPR-ICI, **e** LPR-RICI, **f** SK-LPR-RICI, **g** adaptive local scale with ICI rule, **h** adaptive local scale with RICl rule, **i** adaptive local scale with RICl rule based on steering kernel, **j** enlarged image of ISKR, **k** enlarged image of LPR-RICI, **l** enlarged image of SK-LPR-RICI.

texture, which makes it a commonly used test image. Gaussian kernels are used here, and the scale set of (16) is chosen as $\tilde{H} = \{h_j : |h_j = 1/4, 1/2, 1, 2, 4\}$. The kernel size parameter N_K is set to 11.

Four different adaptive scale selection methods are evaluated:

- 1) *LPR-ICI*: it uses the traditional ICI adaptive scale selector introduced by Katkovnik in [18], for the symmetric circular kernel.
- 2) *LPR-RICI*: it uses the refined ICI adaptive scale selector method proposed in Section 3, and a symmetric circular kernel.
- 3) *Iterative SK-LPR-RICI*: it uses the iterative steering-kernel-based LPR with the refined ICI adaptive scale selector, as proposed in Section 4.
- 4) *ISKR*: it uses the iterative steering kernel regression method introduced by Takeda et al. [22].

For the ISKR method, the global smoothing parameter h_0 is chosen from scale parameters in the set \tilde{H} . The number of iterations in the SK-LPR-RICI and ISKR

method is set to 3. We shall estimate the smooth function $\widehat{m}(\mathbf{X}_i)$ using linear regression with $p=1$. As a result, we have $\beta = 2(p+1-\mathcal{K}) = 4$ and $\nu = 2\mathcal{K} + d = 2$ in the bias and variance expressions (31) and (32). More precisely, the estimate bias is proportional to the squared scale while the variance is inversely proportional to the squared scale. The threshold parameter κ used for the ICI-based methods is selected as 1.96 so that the probability that $\widehat{m}(\mathbf{X}_i)$ lies in the confidence interval is 95%. With these parameter and Eqs. 36–39, we can calculate the parameters for the RIC method: $\Delta\kappa=0.56$, $\eta=0.55$, $\Delta\eta=0.22$, and $h^{opt} = 0.59h_{j^+}$.

The corresponding smoothing results are shown in Figs. 1 and 2 where the additive noise is Gaussian distributed with zero mean and variance $\sigma^2=25$. To evaluate the best possible performance of the ISKR, we assume that the “true image” is available so that the best global smoothing scale h_0 can be found by a one-dimensional search. We found that the MSE curve as a function of h_0 is convex near the global minimum, and, therefore, a bisection search is employed to seek for the best value of h_0 from the bandwidth set $\widetilde{\mathbf{H}}$.

The smoothing results of ISKR method in Figs. 1c and 2c are obtained with the optimal global smoothing parameter h_0 from $\widetilde{\mathbf{H}}$ and the optimal number of iteration l . For the *Test* and *Lena* images, the optimal global smoothing parameters are $h_0=1.27$ and $h_0=0.65$, respectively. The number of iteration for images *Test* and *Lena* are both $l=1$ in the ISKR method. The results of the iterative SK-LPR-RICI method illustrated in Figs. 1f and 2f, are obtained with the same number of iteration. It also gives the best PSNR value.

We can see from Figs. 1c–f and 2c–f that all the four adaptive scale selection methods have satisfactory results: the noise in flat area is suppressed and the edges are preserved quite well. By comparing the PSNR values in Table 1 (a), we find that the SK-LPR-RICI method has the highest PSNR. Some specific areas are enlarged in Figs. 1j–l and 2j–l for better visualization. It shows that the SK-LPR-RICI method also have a better visual performance.

Next, we evaluate the performances of various scale selection methods under different noise variances. The

PSNR values are summarized in Table 2. It can be seen from Table 2 that:

1. Generally, the PSNR values of the four adaptive scale selection methods in ascending order are: LPR-ICI, LPR-RICI, ISKR, SK-LPR-RICI. Note that the best performance in the ISKR method cannot in general be realized because h_0 is chosen to minimize the MSE, which is unavailable in practice.
2. The best global smoothing parameter h_0 for the ISKR method varies with the noise variance. For small noise level, h_0 is relatively small, while for large noise level, h_0 increases accordingly.
3. More iterations generally cause more smoothing of the images. Therefore, the number of iteration should increase with the noise level of the images, if it is known in prior. More iterations should only be applied to rather noisy images in order to avoid over-smoothing.

Finally, the computational complexities of these methods under test are compared. The detailed analysis of one LPR operation in a pixel can be found in Section 4.2. The LPR-ICI method has the lowest complexity because it does not need to refine the ICI-based bandwidths and to calculate the steering kernels. The SK-LPR-RICI and ISKR methods have higher complexities than LPR-ICI and LPR-RICI methods because the parameters to construct the steering kernels have to be estimated. The SK-LPR-RICI method has a higher complexity than ISKR because it includes additional ICI bandwidth selection and refinement. In summary, the computational complexities of the four methods in ascending order are: LPR-ICI, LPR-RICI, ISKR, SK-LPR-RICI.

5.2 Coding Artifacts Removal

LPR can also be used to suppress the quantization noise in image and video compression. This is because the artifacts or noise produced during the compression and quantization can be roughly modeled as Gaussian

Table 1 PSNR comparisons of variable basis-kernel and scale selection methods for image processing ($\sigma^2=25$).

(a). De-noising (Gaussian noise)	LPR-ICI	LPR-RICI	SK-LPR-RICI	ISKR		
<i>Test</i>	38.58	43.58	49.00 ($l=3$)	48.79 ($h_0=1.27, l=1$)		
<i>Lena</i>	38.48	38.64	41.32 ($l=3$)	41.28 ($h_0=0.65, l=1$)		
(b). De-noising (Coding artifacts)	LPR-ICI	LPR-RICI	SK-LPR-RICI	ISKR		
<i>Lena</i>	36.97	37.69	37.97	37.96 ($h_0=1$)		
(c). Super-resolution	LPR-ICI	LPR-RICI	SK-LPR-RICI	ISKR	Bicubic spline	IBP
<i>Lena</i>	37.23	39.57	43.09	42.75 ($h_0=1$)	35.82	36.97

Table 2 PSNR comparisons of variable basis-kernel and scale selection methods for image de-noising (against different noise variance).

	Noise variance	LPR-ICI	LPR-RICI	SK-LPR-RICI	ISKR					
					$h_0=1/4$	$h_0=1/2$	$h_0=1$	$h_0=2$	$h_0=4$	h_0^{opt}
Test	$\sigma^2=25$	38.58	43.58	49.00 ($l=3$)	42.58 ($l=1$)	45.94 ($l=1$)	47.83 ($l=1$)	48.07 ($l=1$)	43.22 ($l=1$)	48.79 ($h_0=1.27, l=1$)
	$\sigma^2=100$	34.97	37.67	45.07 ($l=3$)	34.86 ($l=1$)	39.33 ($l=1$)	42.46 ($l=1$)	44.33 ($l=2$)	42.02 ($l=2$)	44.96 ($h_0=2.18, l=2$)
	$\sigma^2=225$	34.71	35.67	42.35 ($l=3$)	34.19 ($l=3$)	34.43 ($l=3$)	39.11 ($l=3$)	41.92 ($l=3$)	40.14 ($l=3$)	42.36 ($h_0=2.43, l=3$)
	$\sigma^2=400$	34.19	34.84	39.28 ($l=3$)	33.37 ($l=3$)	32.54 ($l=3$)	36.74 ($l=3$)	39.14 ($l=3$)	38.89 ($l=3$)	39.33 ($h_0=2.89, l=3$)
	$\sigma^2=625$	33.60	34.08	37.99 ($l=3$)	32.65 ($l=3$)	31.28 ($l=3$)	34.70 ($l=3$)	37.34 ($l=3$)	37.50 ($l=3$)	38.10 ($h_0=3.15, l=3$)
Lena	$\sigma^2=25$	38.48	38.64	41.32 ($l=1$)	39.74 ($l=1$)	41.05 ($l=1$)	40.69 ($l=1$)	38.16 ($l=1$)	36.46 ($l=1$)	41.28 ($h_0=0.65, l=1$)
	$\sigma^2=100$	35.33	35.88	39.10 ($l=1$)	34.61 ($l=1$)	36.09 ($l=1$)	38.66 ($l=1$)	37.78 ($l=1$)	36.38 ($l=1$)	39.08 ($h_0=1.22, l=1$)
	$\sigma^2=225$	34.09	34.91	37.38 ($l=1$)	33.97 ($l=1$)	33.60 ($l=1$)	36.57 ($l=2$)	37.12 ($l=2$)	36.13 ($l=2$)	37.41 ($h_0=1.72, l=2$)
	$\sigma^2=400$	33.60	33.90	36.94 ($l=3$)	33.60 ($l=3$)	32.30 ($l=3$)	35.00 ($l=3$)	36.61 ($l=3$)	36.00 ($l=3$)	37.05 ($h_0=2.23, l=3$)
	$\sigma^2=625$	32.27	33.20	36.28 ($l=3$)	33.24 ($l=3$)	31.46 ($l=3$)	33.64 ($l=3$)	36.11 ($l=3$)	35.69 ($l=3$)	36.37 ($h_0=2.71, l=3$)

distributed. This is supported by observations in [31], which reported that the JPEG compression noise is Gaussian distributed with zero mean and a covariance matrix determined by the quantization table and DCT transform matrix.

To evaluate the effectiveness of the LPR algorithms, we compressed the image *Lena* by JPEG with a compression ratio 50%, as shown in Fig. 3b. Only one iteration is employed in the ISKR and iterative SK-LPR-RICI methods because the quantization noise is not very severe. Other experimental settings are the same as the previous de-noising experiment. The compression artifacts mainly occur near sharp edges and have block-like shapes, which are illustrated in Fig. 3. It can be seen that the blocking artifacts are suppressed satisfactorily by the LPR-based methods. The proposed SK-LPR-RICI method can achieve the best results among these LPR-based methods, and their PSNR values are summarized in Table 1 (b).

5.3 Super-Resolution Imaging

In this section, the applications of the proposed SK-LPR-RICI method to super-resolution of test gray-scale images and real color photos are presented.

- 1) Gray-scale images: A set of four LR (128×128) images are generated from the original HR (256×256) *Lena* by decimating it by a factor of 2 both horizontally and vertically and then with a shift of zero pixels, one pixel to right, one pixel to bottom, and one pixel to right and bottom, respectively. Gaussian noise of zero mean and variance $\sigma^2=25$ is added to the LR images. The scale set \tilde{H} , the threshold parameter κ , and the kernel size N_K used in this experiment follow their values in previous de-noising experiment.

Experiments on test image *Lena* are carried to evaluate the performances of the SK-LPR-RICI method with other traditional reconstruction methods based on bicubic spline interpolation, the iterative back-projection (IBP) method [28,

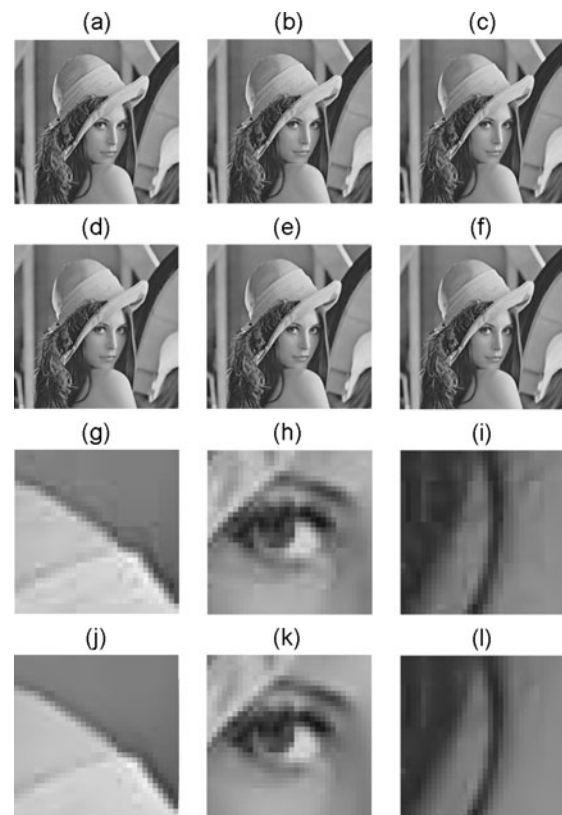


Figure 3 Comparison of different adaptive scale selection methods for smoothing the JPEG coded image *Lena*: **a** original noise-free image, **b** JPEG compressed image, **c** ISKR, **d** LPR-ICI, **e** LPR-RICI, **f** SK-LPR-RICI, **g-i** enlarged images of the original image, **j-l** enlarged images of SK-LPR-RICI.

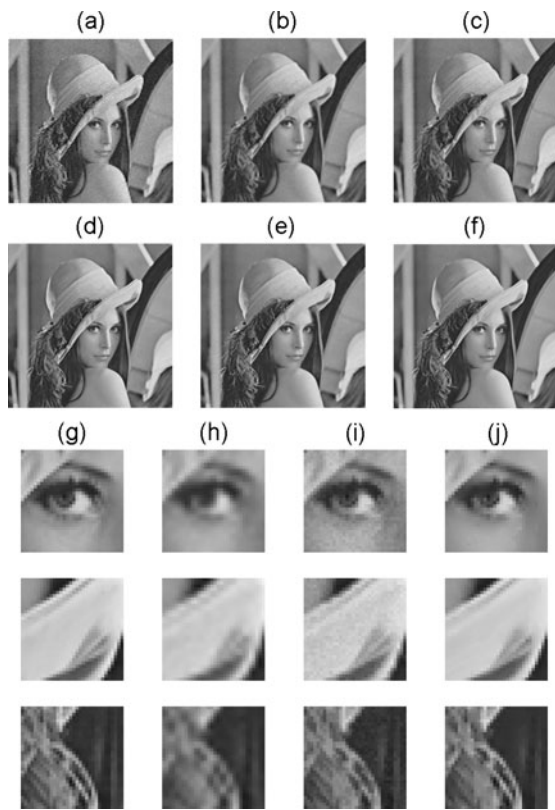


Figure 4 Comparison of different super-resolution methods for the image *Lena*: **a** one low-resolution noisy image, **b** bicubic spline interpolation, **c** IBP, **d** ISKR, **e** LPR-RICI, **f** SK-LPR-RICI, **g** enlarged images of the original HR image, **h** enlarged images of the bicubic spline interpolation, **i** enlarged images of IBP, **j** enlarged images of SK-LPR-RICI.

29], and ISKR. The IBP method minimizes the LS error between the HR image to be estimated and a set of known LR images using LS estimation method with regularization. In LPR, we model the each pixel in the true image as a local polynomial and solve a series of local LS estimation problem using the samples from the known LR images, instead of the whole image. As a result, compared with the IBP method,

the LPR-based reconstruction methods can better adapt to local orientations in the image and achieve a better performance at the expense of a much higher complexity.

As shown in the results of Fig. 4, the IBP method in (c) has a better result than the bicubic interpolation method in (b). But, the edges are blurred as compared with the LPR-based methods in (d)–(f). The main difficulty of the LS regularization method is to choose a global regularization parameter to suppress the additive noise while preserving image edges. The PSNR values of Fig. 4b–f are listed in Table 1 (c). The proposed SK-LPR-RICI method has a higher PSNR value than other methods tested.

2). Color images: We now consider the reconstruction of real color photos. In color images, each pixel consists of the red (R), green (G), and blue (B) components in the RGB color space. Therefore, it is possible to extend the gray-scale super-resolution methods to color images by treating each channel in the color space as an individual image and employing the gray-scale super-resolution methods to each channel independently. However, as pointed out in [27], this method will produce undesirable color artifacts because it ignores the correlation between different channels. In this paper, we first convert the image to the HSI (Hue, Saturation, and Intensity) color space. The adaptive scale selection methods will be applied to the intensity channel to obtain the local basis-kernel and scale parameters. A LPR using the resulting basis-kernel and scale parameters is then applied to the R, G, and B channels of the image, respectively. An iterative maximum a posteriori (MAP) color super-resolution method proposed in [26, 27] is also included in the experiment for comparison. The MAP method also solves a global LS problem with regularization to minimize the estimation error between the HR and LR images.

Six LR (456 × 289) color photos were captured by a commercial digital camera (Canon IXUS 600). The SK-

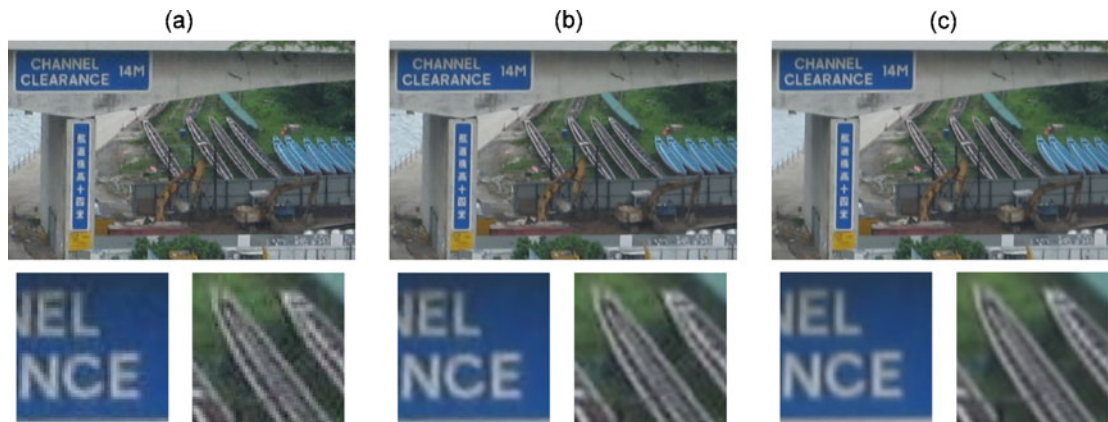


Figure 5 Comparison of different super-resolution methods for real color photos: **a** one LR image and its enlarged images, **b** iterative MAP method and its enlarged images, **c** SK-LPR-RICI and its enlarged images.

LPR-RICI super-resolution method is applied to increase the spatial resolution by a factor of 3 to yield a 1368×867 HR image. A frequency-domain based motion estimation and image registration technique developed in [30] is employed to determine the relative shifts of these LR images.

For comparison, we display the LR photo in the same size as the HR image by resizing each pixel to a 3×3 block. Some areas of the LR and resultant HR images are enlarged so that these details can be seen clearly. From visual inspection of the results in Fig. 5, it can be observed that the SK-LPR-RICI method is able to preserve image edges while remove the image noise effectively.

Overall, based on the experimental results in this section, it can be concluded that the proposed SK-LPR-RICI method is a useful tool in smoothing and reconstruction of noisy images. It can preserve local characteristics well and has a systematic algorithm to automatically calculate the parameters involved in the method.

6 Conclusion

A study on the multivariate LPR for analysis of multidimensional signals was presented. A new refined ICI adaptive bandwidth selection method was first developed for the multivariate LPR and it was shown to have a better bias-variance tradeoff than the traditional ICI rule. This method is further extended to the case of a steering kernel with local orientation for better adaptation to local characteristics of multidimensional signals. The data-adaptive scale selection method was then applied to smoothing and reconstruction of noisy images. Simulation results on test and true images showed that the proposed SK-LPR-RICI method had a better PSNR and visual performance than the conventional LPR-based methods in image processing.

Acknowledgment This work was supported in part by the Hong Kong Research Grant Council (RGC). The authors would like to thank Mr. K. S. Lee for his assistance for this paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.
- Fan, J., & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 57(2), 371–394.
- Fan, J., Gijbels, I., Hu, T. C., & Huang, L. S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6(1), 113–127.
- Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice. *Journal of the Royal Statistical Society: Series B*, 53(2), 453–464.
- Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B*, 53(3), 683–690.
- Hall, P., Sheather, S. J., Jones, M. C., & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78(2), 263–269.
- Chan, S. C., & Zhang, Z. G. (2004). Robust local polynomial regression using M-estimator with adaptive bandwidth. *Proc. IEEE International Symposium of Circuits and Systems*, 333–336.
- Chan, S. C., & Zhang, Z. G. (2004). Multi-resolution analysis of non-uniform data with jump discontinuities and impulsive noise using robust local polynomial regression. *Proc. IEEE International Conference of Acoustics, Speech and Signal Processing*, 769–772.
- Zhang, Z. G., Chan, S. C., Ho, K. L., & Ho, K. R. (2008). On bandwidth selection in local polynomial regression analysis and its application to multi-resolution analysis of non-uniform data. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 52(3), 263–280.
- Ruppert, D., & Wand, M. P. (1995). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22(3), 1346–1370.
- Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432), 1257–1270.
- Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92(439), 1049–1062.
- Yang, L., & Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society: Series B*, 61(4), 793–815.
- Masry, E. (1996). Multivariate polynomial regression for time series; uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(1), 571–599.
- Masry, E. (1996). Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and Their Applications*, 65(1), 81–101.
- Masry, E. (1999). Multivariate regression estimation of continuous-time processes from sampled data: local polynomial fitting approach. *IEEE Transactions on Information Theory*, 45(6), 1939–1953.
- Katkovnik, V. (1999). A new method for varying adaptive bandwidth selection, *IEEE Transactions on Signal Processing*, 47(9), 2567–2571.
- Katkovnik, V., Egiazarian, K., & Astola, J. (2005). A spatially adaptive nonparametric regression image deblurring. *IEEE Transactions on Image Processing*, 14(10), 1469–1478.
- Katkovnik, V., Egiazarian, K., & Astola, J. (2006). *Local approximation techniques in signal and image processing*. Bellingham: SPIE.
- Stanković, L. J. (2004). Performance analysis of the adaptive algorithm for bias-to-variance tradeoff. *IEEE Transactions on Signal Processing*, 52(5), 1228–1234.
- Takeda, H., Farsiu, S., & Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2), 349–366.
- Goldenshluger, A., & Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Mathematical Methods of Statistics*, 6(2), 135–170.
- Lepski, O. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, 35(3), 454–466.
- Chaudhuri, S. (2001). *Super-resolution imaging*. Norwell: Kluwer.

26. Farsiu, S., Robinson, D., Elad, M., & Milanfar, P. (2004). Fast and robust multi-frame superresolution. *IEEE Transactions on Image Processing*, 13(10), 1327–1344.
27. Farsiu, S., Elad, M., & Milanfar, P. (2006). Multi-frame demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing*, 15(1), 141–159.
28. Peleg, S., Keren, D., & Schweitzer, L. (1987). Improving image resolution using subpixel motion. *Pattern Recognition Letters*, 5(3), 223–226.
29. Irani, M., & Peleg, S. (1991). Improving resolution by image registration. *CVGIP. Graphical Models and Image Processing*, 53(3), 231–239.
30. Vandewalle, P., Süsstrunk, S., & Vetterli, M. (2006). A frequency domain approach to registration of aliased images with application to super-resolution. *EURASIP Journal on Applied Signal Processing*. doi:10.1155/ASP/2006/71459.
31. Li, D., & Leow, M. (2003). Closed-form quality measures for compressed medical images: statistical preliminaries for transform coding. *Proc. of the 25th IEEE Annual International Conference on Engineering in Medicine and Biology Society*, pp. 837–840.
32. Zhang, Z. G., Chan, S. C., & Zhu, Z. Y. (2009). A new two-stage method for restoration of images corrupted by Gaussian and impulse noises using local polynomial regression and edge preserving regularization. *Proc. IEEE International Symposium on Circuits and Systems*, 948–951.



Z. G. Zhang received the B.Sc. and M.Eng. degrees in Electrical and Electronic Engineering from Tianjin University and the University of Science and Technology of China in 2000 and

2003, respectively. He obtained the Ph.D. degree in 2008 from the Department of Electrical and Electronic Engineering at the University of Hong Kong, where he is currently a Postdoctoral Fellow. His research interests are in general area of statistical signal processing and digital signal processing, and in particular adaptive filtering, nonparametric regression, time–frequency analysis, and biomedical signal processing.



S. C. Chan received the B.Sc. (Eng) and Ph.D. degrees from the University of Hong Kong in 1986 and 1992, respectively. He joined City Polytechnic of Hong Kong in 1990 as an assistant Lecturer and later as a University Lecturer. Since 1994, he has been with the Department of Electrical and Electronic Engineering, the University of Hong Kong, and is now a Professor. He was a visiting researcher in Microsoft Corporation, Redmond, USA, Microsoft, Beijing, China, University of Texas at Arlington and Nanyang Technological University, Singapore. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed AD converter architecture, bioinformatics and image-based rendering. Dr. Chan is currently a member of the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society and the Associate Editors of *IEEE Transactions on Circuits and Systems I* and *Journal of Signal Processing Systems*. He was Chairman of the IEEE Hong Kong Chapter of Signal Processing 2000–2002 and organizing committee member of IEEE ICASSP 2003 and ICIP 2010.