Estimating Density Ridges by Direct Estimation of Density-Derivative-Ratios

Hiroaki Sasaki[1,4] hsasaki@is.naist.jp
Takafumi Kanamori[2,4] kanamori@is.nagoya-u.ac.jp
Masashi Sugiyama[3,4] sugi@k.u-tokyo.ac.jp

[1] Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
[2] Department of Computer Science and Mathematical Informatics, Nagoya University, Aichi, Japan
[3] Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
[4] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

Abstract

Estimation of *density ridges* has been gathering a great deal of attention since it enables us to reveal lower-dimensional structures hidden in data. Recently, *subspace constrained mean shift* (SCMS) was proposed as a practical algorithm for density ridge estimation. A key technical ingredient in SCMS is to accurately estimate the ratios of the density derivatives to the density. SCMS takes a three-step approach for this purpose — first estimating the data density, then computing its derivatives, and finally taking their ratios. However, this three-step approach can be unreliable because a good density estimator does not necessarily mean a good density derivative estimator and division by an estimated density could significantly magnify the estimation error. To overcome these problems, we propose a novel method that directly estimates the ratios without going through density estimation and division. Our proposed estimator has an analytic-form solution and it can be computed efficiently. We further establish a non-parametric convergence bound for the proposed ratio estimator. Finally, based on this direct ratio estimator, we develop a practical algorithm for density ridge estimation and experimentally demonstrate its usefulness on a variety of datasets.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

1 Introduction

Estimating the *ridge* of the data density possesses a wide range of real-world applications, including estimation of filamentary structures formed by galaxies in cosmology [Chen et al., 2016], extraction of curvilinear structures (e.g., blood vessels in eye balls) in medical imaging [You et al., 2011], skeletonization of optical characters for feature extraction and compression [Kégl and Krzyźak, 2002], traffic pattern analysis [Einbeck and Dwyer, 2011], and shape analysis in computer vision [Su et al., 2013] (see Pulkkinen [2015] for more applications). For this reason, density ridge estimation has been gathering a great deal of attention recently [Ozertem and Erdogmus, 2011, Genovese et al., 2014, Chen et al., 2015b,a, Ghassabeh et al., 2013].

Extending the classical concept of *principal curves* [Hastie, 1984, Hastie and Stuetzle, 1989], a practical algorithm for density ridge estimation called *subspace constrained mean shift* (SCMS) [Ozertem and Erdogmus, 2011] has been proposed recently. SCMS is essentially a projected gradient ascent algorithm over an estimated data density: At each iteration, a gradient vector of the estimated density is computed as in *mean shift* clustering [Fukunaga and Hostetler, 1975, Cheng, 1995, Comaniciu and Meer, 2002], and then it is projected to the subspace orthogonal to the density ridge. Along this projected gradient vector, data points are updated toward the density ridge until they converge. See Genovese et al. [2014], Chen et al. [2015b], Chen et al. [2015a] and Ghassabeh et al. [2013] for theoretical properties of SCMS.

Technically, the key ingredients to obtain such projected gradient vectors are *the ratios of density derivatives to the density*. SCMS takes a three-step approach to estimate the ratios: First, estimate the data density

by *kernel density estimation*, then compute its derivatives, and finally take their ratios. However, this three-step approach can be unreliable because a good density estimator does not necessarily mean a good density derivative estimator, and division by an estimated density could significantly magnify the estimation error.

To cope with these problems, we propose a novel estimator called the *least-squares density-derivative-ratios* (LSDDR). In stark contrast to the three-step approach in SCMS, LSDDR neither performs density estimation nor involves division by an estimated quantity; rather, it directly estimates the *density-derivative-ratios*. Previously, a method to directly estimate the log-density derivatives (which is equal to the ratio of the first-order density derivative to the density) has been proposed [Cox, 1985, Sasaki et al., 2014]. LSDDR can be regarded as its generalization to higher-order derivatives. LSDDR has an analytic-form solution and it can be computed efficiently. Furthermore, we establish a non-parametric convergence bound for LSDDR.

Based on this LSDDR, we then develop a new algorithm for density ridge estimation called the *least-squares density ridge finder* (LSDRF). We experimentally demonstrate the advantages of LSDRF over SCMS on a variety of datasets.

2 Density Ridge Estimation

In this section, we formulate the problem of density ridge estimation and review an existing method.

2.1 Problem Formulation

Suppose that independent and identically distributed samples $\mathcal{X} = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)})^\top\}_{i=1}^n$ drawn from an unknown probability distribution with density $p(\boldsymbol{x})$ are available, where $^\top$ denotes the transpose. For positive integer d such that $d < D$, the goal is to estimate from $\mathcal{X}$ the d-dimensional *density ridge* [Eberly, 1996, Ozertem and Erdogmus, 2011, Genovese et al., 2014], which is defined as a collection of points satisfying

$$\mathcal{R} = \{\boldsymbol{x} \in \mathbb{R}^D \mid \mathbf{V}(\boldsymbol{x})\mathbf{V}(\boldsymbol{x})^\top \nabla p(\boldsymbol{x}) = 0, \eta_{d+1}(\boldsymbol{x}) < 0\}, \quad (1)$$

where ∇ denotes the differential operator w.r.t. $\boldsymbol{x}$, $\mathbf{V}(\boldsymbol{x}) = (\boldsymbol{v}_{d+1}, \ldots, \boldsymbol{v}_D)$, and $\boldsymbol{v}_i$ is the eigenvector associated with the eigenvalue $\eta_i(\boldsymbol{x})$ of the Hessian matrix $\nabla\nabla p(\boldsymbol{x})$. We assume that the eigenvalues are sorted in descending order such that $\eta_1(\boldsymbol{x}) \geq \eta_2(\boldsymbol{x}) \geq \cdots \geq \eta_D(\boldsymbol{x})$.

2.2 Subspace Constrained Mean Shift (SCMS)

According to definition (1), a practical algorithm for finding density ridges called *subspace constrained mean shift* (SCMS) was proposed in Ozertem and Erdogmus [2011]. SCMS is based on *mean shift* (MS) clustering [Fukunaga and Hostetler, 1975, Cheng, 1995, Comaniciu and Meer, 2002] and the *inverse local-covariance matrix*.

MS is a mode-seeking clustering method based on kernel density estimation (KDE):

$$\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x}) = \frac{1}{nZ}\sum_{i=1}^n K_{\mathrm{KDE}}\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{h^2}\right),$$

where Z is the normalizing constant, K_{KDE} is a kernel function for KDE, and h denotes its bandwidth. MS updates data points toward the nearest mode (i.e., a local maximum) of $\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x})$ by

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{x}) = \frac{\sum_{i=1}^n \boldsymbol{x}_i G_{\mathrm{KDE}}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{h^2}\right)}{\sum_{i=1}^n G_{\mathrm{KDE}}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{h^2}\right)},$$

where $G_{\mathrm{KDE}}(t) = -K'_{\mathrm{KDE}}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}K_{\mathrm{KDE}}(t)$. $\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{x})$ is called the *mean shift vector* which is shown to be parallel to $\nabla\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x})$ [Comaniciu and Meer, 2002]:

$$\begin{aligned}\nabla\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x}) &= \frac{2}{nh^2Z}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{x}) G_{\mathrm{KDE}}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{h^2}\right) \\ &= \alpha(\boldsymbol{x})\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{x}),\end{aligned}$$

where $\alpha(\boldsymbol{x}) = \frac{2}{nh^2Z}\sum_{i=1}^n G_{\mathrm{KDE}}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{h^2}\right)$. The above equation indicates that MS performs gradient ascent with adaptive step size $1/\alpha(\boldsymbol{x})$.

The basic idea of SCMS is to perform MS-like gradient ascent on the subspace which is orthogonal to the density ridge. SCMS obtains such a subspace as the span of the eigenvectors of the Hessian matrix of the log-density, which is called the *inverse local-covariance matrix* [Ozertem and Erdogmus, 2011]:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}) &= -\nabla\nabla\log p(\boldsymbol{x}) \\ &= -\frac{\nabla\nabla p(\boldsymbol{x})}{p(\boldsymbol{x})} + \frac{\nabla p(\boldsymbol{x})\nabla p(\boldsymbol{x})^\top}{p(\boldsymbol{x})^2}. \quad (2)\end{aligned}$$

As theoretically shown in Genovese et al. [2014], the log-density is used instead of the (non-log) density because it has some advantages: In practice, a projector to the subspace is obtained by applying *principal component analysis* (PCA) to the estimated inverse local-covariance matrix $\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{KDE}}(\boldsymbol{x})$ where $p(\boldsymbol{x})$ in (2) is

Input: A data point $\boldsymbol{x}$.

Step 1 Initialize $t = 0$ and $\boldsymbol{y}(t) = \boldsymbol{x}$.

Step 2 Evaluate the mean shift vector $\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{y}(t))$.

Step 3 Evaluate the inverse local-covariance matrix using $\widehat{p}_{\mathrm{KDE}}(\boldsymbol{y}(t))$.

Step 4 Construct a projector by applying PCA to $\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{KDE}}(\boldsymbol{x})$.

Step 5 Update $\boldsymbol{y}(t)$ as $\boldsymbol{y}(t+1) = \boldsymbol{y}(t) + \widehat{\mathbf{V}}_{\mathrm{KDE}}\widehat{\mathbf{V}}^{\top}_{\mathrm{KDE}}\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{y}(t))$.

Step 6 Stop if $\|\boldsymbol{y}(t+1) - \boldsymbol{y}(t)\|_2 < \epsilon$. Otherwise, $t \leftarrow t+1$ and go back to Step 2.

Output: $\widehat{\boldsymbol{y}} = \boldsymbol{y}(t)$

Figure 1: An algorithm of subspace constrained mean shift [Ozertem and Erdogmus, 2011, Ghassabeh et al., 2013]. $\|\cdot\|_2$ is the ℓ_2 norm and ϵ denotes a small positive constant. As an input data point, a data sample itself is typically used.

replaced with $\widehat{p}_{\mathrm{KDE}}(\boldsymbol{x})$. Then, the projected gradient update rule of SCMS is given as

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \widehat{\mathbf{V}}_{\mathrm{KDE}}\widehat{\mathbf{V}}^{\top}_{\mathrm{KDE}}\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{x}). \tag{3}$$

The entire algorithm of SCMS is summarized in Figure 1. The convergence of the SCMS algorithm is proved in Ghassabeh et al. [2013].

In SCMS, one of the key challenges is to accurately estimate the inverse local covariance matrix (2). SCMS takes a three-step approach, i.e., estimate $p(\boldsymbol{x})$ by KDE, compute its derivatives, and plug them into (2). However, this approach can perform poorly because a good density estimator does not necessarily mean a good density derivative estimator. In addition, division by an estimated density could significantly magnify the estimation error. A more appropriate way would be to directly estimate the ratios in (2) without going through density estimation and division. Following this idea, we next propose a novel direct estimator of the ratios.

3 Direct Estimation of Density-Derivative-Ratios

In this section, we propose a direct estimator of density-derivative-ratios. Then, the estimator is theoretically analyzed.

3.1 Least-Squares Density-Derivative-Ratios

Here, our tentative goal is to estimate the ratio of the k-th order partial derivative of $p(\boldsymbol{x})$ to $p(\boldsymbol{x})$ from $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$\frac{\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})}{p(\boldsymbol{x})}, \tag{4}$$

where $\partial^{k,\boldsymbol{j}} = \frac{\partial^k}{\partial (x^{(1)})^{j_1} \partial (x^{(2)})^{j_2} \ldots \partial (x^{(D)})^{j_D}}$, $\boldsymbol{j} = (j_1, j_2, \ldots, j_D)$ and $j_1 + j_2 + \cdots + j_D = k$ for $j_i \in \{0, 1, \ldots, k\}$. For instance, when $k = 1$ (or $k = 2$), $\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})/p(\boldsymbol{x})$ is a single element of $\nabla p(\boldsymbol{x})/p(\boldsymbol{x})$ (or of $\nabla\nabla p(\boldsymbol{x})/p(\boldsymbol{x})$).

Our main idea is to directly fit a model $r_{k,\boldsymbol{j}}(\boldsymbol{x})$ to $\frac{\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})}{p(\boldsymbol{x})}$ under the squared-loss:

$$\begin{aligned} &J_{k,\boldsymbol{j}}(r_{k,\boldsymbol{j}}) \\ &= \int \left\{ r_{k,\boldsymbol{j}}(\boldsymbol{x}) - \frac{\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})}{p(\boldsymbol{x})} \right\}^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - C_{k,\boldsymbol{j}} \\ &= \int \left\{ r_{k,\boldsymbol{j}}(\boldsymbol{x}) \right\}^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - 2 \int r_{k,\boldsymbol{j}}(\boldsymbol{x}) \partial^{k,\boldsymbol{j}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{5}$$

where $C_{k,\boldsymbol{j}} = \int \left\{ \frac{\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})}{p(\boldsymbol{x})} \right\}^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$. The first term in (5) can be naively estimated from samples, but it seems challenging to estimate the second term because it includes the derivative of the unknown density. However, similarly to Sasaki et al. [2015], repeatedly applying *integration by parts* allows us to transform the second term as

$$\begin{aligned} &\int r_{k,\boldsymbol{j}}(\boldsymbol{x}) \left\{ \partial^{k,\boldsymbol{j}} p(\boldsymbol{x}) \right\} \mathrm{d}\boldsymbol{x} \\ &= (-1)^k \int \left\{ \partial^{k,\boldsymbol{j}} r_{k,\boldsymbol{j}}(\boldsymbol{x}) \right\} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{6}$$

where we assumed that as for all j, $|x^{(j)}| \to \infty$, the product of $\partial^{k_1,\boldsymbol{j}_1} r_{k_1,\boldsymbol{j}}(\boldsymbol{x})$ and $\partial^{k_2,\boldsymbol{j}_2} p(\boldsymbol{x})$ approaches zero for any pairs of k_1 and k_2 satisfying $k_1 + k_2 = k - 1$. As a result, the right-hand side of (6) can be easily estimated from samples. Then, an empirical version of (5) is given by

$$\tilde{J}_{k,\boldsymbol{j}}(r_{k,\boldsymbol{j}}) = \sum_{i=1}^{n} r_{k,\boldsymbol{j}}(\boldsymbol{x}_i)^2 - 2(-1)^k \partial^{k,\boldsymbol{j}} r_{k,\boldsymbol{j}}(\boldsymbol{x}_i). \tag{7}$$

To estimate $r_{k,\boldsymbol{j}}$, we employ a linear-in-parameter model:

$$r_{k,\boldsymbol{j}}(\boldsymbol{x}) = \sum_{i=1}^{n} \theta^{(i)}_{k,\boldsymbol{j}} \psi^{(i)}_{k,\boldsymbol{j}}(\boldsymbol{x}) = \boldsymbol{\theta}^{\top}_{k,\boldsymbol{j}} \boldsymbol{\psi}_{k,\boldsymbol{j}}(\boldsymbol{x}),$$

where we set $\psi^{(i)}_{k,\boldsymbol{j}}(\boldsymbol{x}) = \partial^{k,\boldsymbol{j}} K(\boldsymbol{x}, \boldsymbol{x}_i)$ and $K(\cdot,\cdot)$ is a smooth kernel function such as the Gaussian kernel. Substituting the model into (7) and adding the

ℓ_2-regularizer yield the following quadratic objective function:

$$\widetilde{J}_{k,\boldsymbol{j}}(\boldsymbol{\theta}_{k,\boldsymbol{j}}) = \boldsymbol{\theta}_{k,\boldsymbol{j}}^\top \mathbf{G}_{k,\boldsymbol{j}} \boldsymbol{\theta}_{k,\boldsymbol{j}} - 2(-1)^k \boldsymbol{\theta}_{k,\boldsymbol{j}}^\top \boldsymbol{h}_{k,\boldsymbol{j}} + \lambda_{k,\boldsymbol{j}} \boldsymbol{\theta}_{k,\boldsymbol{j}}^\top \boldsymbol{\theta}_{k,\boldsymbol{j}}, \quad (8)$$

where

$$\mathbf{G}_{k,\boldsymbol{j}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}_{k,\boldsymbol{j}}(\boldsymbol{x}_i)\boldsymbol{\psi}_{k,\boldsymbol{j}}(\boldsymbol{x}_i)^\top,$$
$$\boldsymbol{h}_{k,\boldsymbol{j}} = \frac{1}{n}\sum_{i=1}^{n} \partial^{k,\boldsymbol{j}}\boldsymbol{\psi}_{k,\boldsymbol{j}}(\boldsymbol{x}_i).$$

The minimizer of (8) can be computed analytically as

$$\widehat{\boldsymbol{\theta}}_{k,\boldsymbol{j}} = \operatorname*{argmin}_{\boldsymbol{\theta}_{k,\boldsymbol{j}}} \widetilde{J}_{k,\boldsymbol{j}}(\boldsymbol{\theta}_{k,\boldsymbol{j}}) = (-1)^k \left(\mathbf{G}_{k,\boldsymbol{j}} + \lambda_{k,\boldsymbol{j}}\mathbf{I}\right)^{-1} \boldsymbol{h}_{k,\boldsymbol{j}},$$

where $\mathbf{I}$ denotes the identity matrix. Finally, a density-derivative-ratio estimator is given by

$$\widehat{r}_{k,\boldsymbol{j}}(\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}_{k,\boldsymbol{j}}^\top \boldsymbol{\psi}_{k,\boldsymbol{j}}(\boldsymbol{x}).$$

We call this method the *least-squares density-derivative ratios* (LSDDR). Note that when $k = 1$, LSDDR is reduced to an existing log-density derivative estimator [Cox, 1985, Sasaki et al., 2014]. Therefore, LSDDR can be regarded as its generalization to higher-order derivatives.

3.2 Theoretical Analysis

Next, we perform theoretical analysis of LSDDR. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) associated with the kernel $K(\cdot,\cdot)$. In our analysis, we assume that the true density-derivative-ratio is contained in $\mathcal{H}$:

$$r^*_{k,\boldsymbol{j}}(\boldsymbol{x}) := \frac{\partial^{k,\boldsymbol{j}} p(\boldsymbol{x})}{p(\boldsymbol{x})} \in \mathcal{H}.$$

According to Zhou [2008], $\partial^{k,\boldsymbol{j}} K(\boldsymbol{x},\cdot)$ belongs to $\mathcal{H}$ under regularity conditions. Thus, the linear-in-parameter model employed for $r_{k,\boldsymbol{j}}$ also belongs to $\mathcal{H}$. In this analysis, we define a slightly modified version of the LSDDR estimator $\widehat{r}_{k,\boldsymbol{j}}$ as the optimal solution of

$$\min_{r\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}r(\boldsymbol{x}_i)^2 - (-1)^k \partial^{k,\boldsymbol{j}} r(\boldsymbol{x}_i)\right\} \quad \text{subject to } \|r\|_{\mathcal{H}}^2 \le M_n^2, \quad (9)$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in $\mathcal{H}$ and M_n is a constant depending on the sample size n. Note that the optimal solution of (9) can be expressed by the kernelized LSDDR estimator with regularization parameter $\lambda_{k,\boldsymbol{j}}$.

In the following theorem, we establish the convergence rate of LSDDR. The accuracy of the estimator is evaluated by the L_2-norm under the distribution P defined as

$$\|h\|_P^2 = \int |h(\boldsymbol{x})|^2 \mathrm{d}P(\boldsymbol{x}).$$

Theorem 1 (Convergence rate of LSDDR)**.** *Let us assume that the kernel function $K(\boldsymbol{x},\boldsymbol{x}')$ of $\mathcal{H}$ is smooth and that there exists a constant $c_K > 0$ such that*

$$\partial^{\boldsymbol{j}'}\partial^{'\boldsymbol{j}'} K(\boldsymbol{x},\boldsymbol{x}) := \frac{\partial^{|\boldsymbol{j}'|}}{\partial^{j'_1}x_1 \cdots \partial^{j'_D}x_D}\frac{\partial^{|\boldsymbol{j}'|}}{\partial^{j'_1}x'_1 \cdots \partial^{j'_D}x'_D} K(\boldsymbol{x},\boldsymbol{x}')\Big|_{\boldsymbol{x}'=\boldsymbol{x}} \le c_K \quad (10)$$

holds for any $|\boldsymbol{j}'| \le k+\ell$, where $|\boldsymbol{j}'|$ is the sum of the elements in $\boldsymbol{j}' = (j'_1,\ldots,j'_D)$ and ℓ is a natural number greater than $D/2$. Suppose that M_n is of the poly-logarithmic order of n. Then,

$$\|\widehat{r}_{k,\boldsymbol{j}} - r^*_{k,\boldsymbol{j}}\|_P^2 = O_p(c_n/n^{1/(2+D/\ell)}),$$

where $O_P(\cdot)$ denotes the probabilistic order and c_n is of the poly-logarithmic order of n.

Due to lack of space, we only provide a sketch of proof below. The full proof is given in the supplementary material.

Sketch of the proof of Theorem 1. Let $L(r)$ be the objective function of (9). Suppose that n is sufficiently large so that $\|r^*_{k,\boldsymbol{j}}\|_{\mathcal{H}} \le M_n$ holds. Then, the inequality $L(\widehat{r}_{k,\boldsymbol{j}}) \le L(r^*_{k,\boldsymbol{j}})$ leads to

$$\frac{1}{2}\|\widehat{r} - r^*_{k,\boldsymbol{j}}\|_P^2 \le \frac{1}{2}\int\{(\widehat{r})^2 - (r^*_{k,\boldsymbol{j}})^2\}d(P-P_n) - (-1)^k\int\{\partial^{k,\boldsymbol{j}}\widehat{r} - \partial^{k,\boldsymbol{j}} r^*_{k,\boldsymbol{j}}\}d(P-P_n),$$

where the equality, $\int (r^*_{k,\boldsymbol{j}})^2 \mathrm{d}P = (-1)^k \int \partial^{k,\boldsymbol{j}} r^*_{k,\boldsymbol{j}} \mathrm{d}P$, is used. Using Proposition 4 in Cucker and Smale [2002], we find that the convergence rate of $\|\widehat{r} - r^*_{k,\boldsymbol{j}}\|_P^2$ is closely related to the complexity of the function sets,

$$\mathcal{F} = \{r^2 - (r^*_{k,\boldsymbol{j}})^2 \,|\, \|r\|_{\mathcal{H}} \le M_n\},$$
$$\mathcal{G} = \{\partial^{k,\boldsymbol{j}} r - \partial^{k,\boldsymbol{j}} r^*_{k,\boldsymbol{j}} \,|\, \|r\|_{\mathcal{H}} \le M_n\}.$$

Here, the complexity of the function set $\mathcal{F}'$ is measured by the covering number $\mathcal{N}_\infty(\mathcal{F}',\varepsilon)$, which is defined as the minimal number $n \in \mathbb{N}$ such that there exist n disks with radius ε covering $\mathcal{F}'$, i.e.,

$$\mathcal{N}_\infty(\mathcal{F}',\varepsilon) = \min\{n \in \mathbb{N} \,|\, \exists \mathcal{S} = \{g_1,\ldots,g_n\} \subset \mathcal{F}', \forall f \in \mathcal{F}', \exists g \in \mathcal{S}, \|f-g\|_\infty < \varepsilon\}.$$

Under the assumption (10) with $|\boldsymbol{j}'| = 0$, the standard technique shown in Cucker and Smale [2002] can be used to obtain an upper bound of $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$. In order to obtain an upper bound of $\mathcal{N}_\infty(\mathcal{G}, \varepsilon)$, the formula on the derivative in the RKHS [Zhou, 2008], i.e.,

$$|\partial^{k',\boldsymbol{j}'} r(\boldsymbol{x})| \leq \|r\|_{\mathcal{H}} \sqrt{\partial^{\boldsymbol{j}'} \partial^{\prime \boldsymbol{j}'} K(\boldsymbol{x}, \boldsymbol{x})},$$

is used up to the order $k' \leq k + \ell$. Then, the uniform convergence rates of $\int (r^2 - (r^*_{k,\boldsymbol{j}})^2) d(P - P_n)$ and $\int (\partial^{k,\boldsymbol{j}} r - \partial^{k,\boldsymbol{j}} r^*_{k,\boldsymbol{j}}) d(P - P_n)$ are respectively evaluated by these covering numbers. Eventually, we obtain the upper bound of $\|\widehat{r} - r^*_{k,\boldsymbol{j}}\|_P^2$. □

3.3 Model Selection by Cross-Validation

In practice, the performance of LSDDR depends on the choice of hyper-parameters such as parameters in $\psi^{(i)}_{k,\boldsymbol{j}}(\boldsymbol{x})$ and the regularization parameter $\lambda_{k,\boldsymbol{j}}$. Such hyper-parameters can be chosen by cross-validation with respect to the squared-loss criterion in a straightforward way as follows:

1. Divide the samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^T$.

2. Obtain the estimator $\widehat{r}^{(t)}_{k,\boldsymbol{j}}(\boldsymbol{x})$ from $\mathcal{X} \setminus \mathcal{X}_t$, and then compute $\widetilde{J}_{k,\boldsymbol{j}}$ from the hold-out samples as

$$\mathrm{CV}(t) = \frac{1}{|\mathcal{X}_t|} \sum_{\boldsymbol{x} \in \mathcal{X}_t} \left[\left\{ \widehat{r}^{(t)}_{k,\boldsymbol{j}}(\boldsymbol{x}) \right\}^2 - 2(-1)^k \partial^{k,\boldsymbol{j}} \widehat{r}^{(t)}_{k,\boldsymbol{j}}(\boldsymbol{x}) \right],$$

   where $|\mathcal{X}_t|$ denotes the number of elements in $\mathcal{X}_t$.

3. Choose the model that minimizes $\mathrm{CV} = \frac{1}{T} \sum_{t=1}^T \mathrm{CV}(t)$.

4 Application in Density Ridge Estimation

In this section, based on LSDDR proposed in Section 3, we develop a novel density ridge estimator called the *least-squares density ridge finder* (LSDRF). For LSDDR, we employ the Gaussian kernel with bandwidth $\sigma_{k,\boldsymbol{j}}$.[1]

The algorithm of LSDRF essentially follows the same line as SCMS (see Figure 1), i.e., projected gradient ascent is performed:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \widehat{\mathbf{V}}_{\mathrm{LS}} \widehat{\mathbf{V}}_{\mathrm{LS}}^\top \widehat{\boldsymbol{m}}_{\mathrm{LS}}(\boldsymbol{x}).$$

[1] If the sample size n is large, we may only use a subset of data samples as Gaussian centers in LSDDR.

However, compared with the update rule (3) used in SCMS, $\widehat{\mathbf{V}}_{\mathrm{KDE}}$ and $\widehat{\boldsymbol{m}}_{\mathrm{MS}}(\boldsymbol{x})$ are replaced with $\widehat{\mathbf{V}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{m}}_{\mathrm{LS}}(\boldsymbol{x})$, respectively.[2]

$\widehat{\mathbf{V}}_{\mathrm{LS}}$ is obtained by applying PCA to an estimate of the inverse local-covariance matrix obtained based on LSDDR, not KDE:

$$\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{LS}}(\boldsymbol{x}) = -\widehat{\boldsymbol{H}}_{\mathrm{LS}}(\boldsymbol{x}) + \widehat{\boldsymbol{g}}_{\mathrm{LS}}(\boldsymbol{x}) \widehat{\boldsymbol{g}}_{\mathrm{LS}}(\boldsymbol{x})^\top, \tag{11}$$

where the elements in $\widehat{\boldsymbol{g}}_{\mathrm{LS}}(\boldsymbol{x})$ and $\widehat{\boldsymbol{H}}_{\mathrm{LS}}(\boldsymbol{x})$ are the LSDDR solutions $\widehat{r}_{1,\boldsymbol{j}}(\boldsymbol{x})$ and $\widehat{r}_{2,\boldsymbol{j}}(\boldsymbol{x})$, respectively.

$\widehat{\boldsymbol{m}}_{\mathrm{LS}}(\boldsymbol{x})$ is given by

$$\begin{aligned}
[\widehat{\boldsymbol{m}}_{\mathrm{LS}}(\boldsymbol{x})]_\ell &= \frac{\sum_{i=1}^n \widehat{\theta}^{(i)}_{1,\ell} [\boldsymbol{x}_i]_\ell K(\boldsymbol{x}, \boldsymbol{x}_i)}{\sum_{i=1}^n \widehat{\theta}^{(i)}_{1,\ell} K(\boldsymbol{x}, \boldsymbol{x}_i)} - [\boldsymbol{x}]_\ell \\
&= \frac{1}{\sum_{i=1}^n \widehat{\theta}^{(i)}_{1,\ell} K(\boldsymbol{x}, \boldsymbol{x}_i)} [\widehat{\boldsymbol{g}}_{\mathrm{LS}}(\boldsymbol{x})]_\ell,
\end{aligned} \tag{12}$$

where $[\boldsymbol{x}]_\ell$ denotes the ℓ-th element in $\boldsymbol{x}$. This vector comes from another mode-seeking clustering method called *least-squares log-density gradients* (LSLDG) clustering [Sasaki et al., 2014], which was experimentally shown to work much better than MS especially for higher-dimensional data.

5 Experiments

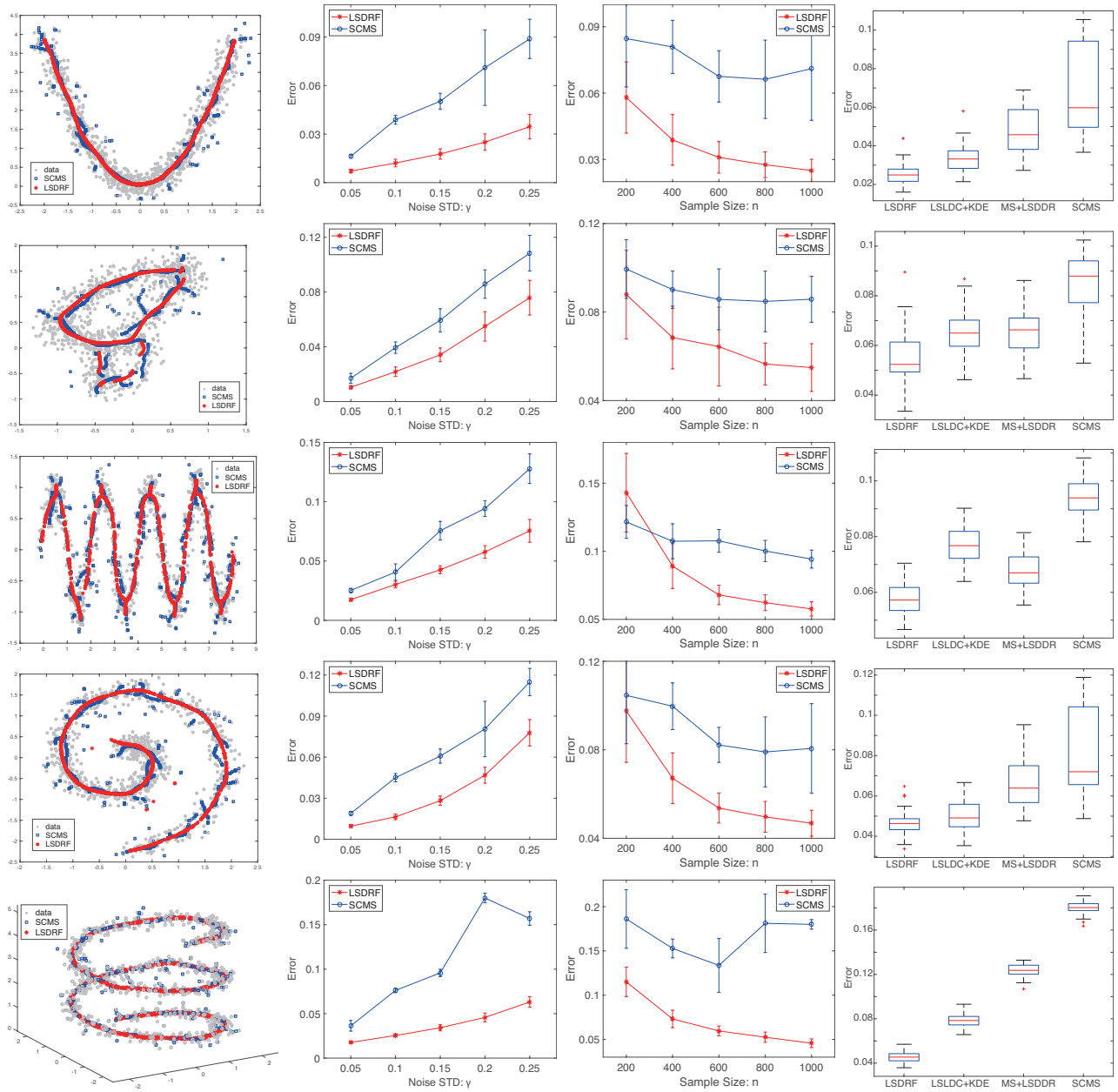
In this section, we experimentally demonstrate the usefulness of LSDRF.

5.1 Illustration on Simulated Data

First, we investigate the performance of LSDRF and compare it with SCMS on a variety of simulated datasets.[3] The i-th observation of data was generated according to $x_i^{(j)} = f^{(j)}(t_i) + n_i^{(j)}$, where t_i was taken from some range at regular intervals, $f^{(j)}(\cdot)$ denotes some fixed function, and $n_i^{(j)}$ was the Gaussian noise with mean 0 and standard deviation γ. The bandwidth h of the Gaussian kernel in SCMS was determined by least-squares cross-validation using ten candidates from $10^{-1.5}$ to 10^0 at regular intervals in logarithmic scale. For LSDDR, model selection was performed by five-hold cross-validation using ten candidates from 10^{-1} (or 10^{-4}) to $10^{0.5}$ (or 10^0) for $\sigma_{k,\boldsymbol{j}}$

[2] To avoid numerical instability, we stop updating a data point $\boldsymbol{x}$ if $\sum_{i=1}^n \widehat{\theta}^{(i)}_{1,\ell} K(\boldsymbol{x}, \boldsymbol{x}_i)$ is less than a very small positive constant.

[3] Most of the datasets are generated using a MATLAB package made by Jakob Verbeek, which is available at `http://lear.inrialpes.fr/people/verbeek/code/kseg_soft.tar.gz`.



(a) Density ridge estimation (b) Error against the noise standard deviation γ. (c) Error against the sample size n. (d) Performance comparison in four settings.

Figure 2: Density ridge estimation on simulated data. In (b) and (c), each point and error bar denote the average and standard deviation of estimation errors over 50 runs, respectively. We set $(n, \gamma) = (1000, 0.15)$ for (a), $n = 1000$ for (b) and (d), and $\gamma = 0.2$ for (c).

(or $\lambda_{k,\boldsymbol{j}}$) at regular intervals in logarithmic scale. The estimation error was measured by

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n} \min_{l} \|\widehat{\boldsymbol{y}}_i - \boldsymbol{f}(t_l)\|_2,$$

where $\widehat{\boldsymbol{y}}_i$ denotes an estimate of the density ridge obtained from $\boldsymbol{x}_i$ and $\boldsymbol{f}(\cdot) = (f^{(1)}(\cdot), f^{(2)}(\cdot), \ldots, f^{(D)}(\cdot))^\top$.

Two density ridge estimates from LSDRF and SCMS are visualized in Figures 2(a). LSDRF provides more accurate and smooth ridge estimates than SCMS on all datasets. Figures 2(b) show the noise tolerance property of both methods. As the noise standard deviation γ increases, the performance of both LSDRF and SCMS gets worse, but LSDRF still works better than SCMS in all five cases. Figures 2(c) further indi-

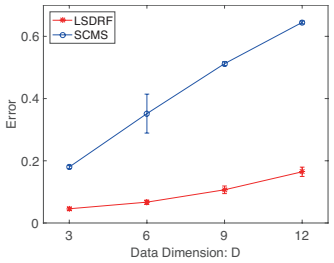


Figure 3: Performance comparison for higher-dimensional data. We set $(n, \gamma) = (1000, 0.2)$.

cates that the estimation error of LSDRF more quickly decreases with respect to the sample size n.

Next, we investigate how LSDRF and SCMS perform in higher-dimensional cases. For this experiment, to create higher-dimensional data, Gaussian variables with mean 0 and standard deviation $\gamma = 0.2$ were appended to the three-dimensional spiral data in Figure 2. The result shown in Figure 3 indicates that LSDRF is more useful than SCMS especially for higher-dimensional data.

LSDRF makes two major modifications from SCMS: (a) $\widehat{\boldsymbol{m}}_{\mathrm{LS}}$ from a previously proposed mode-seeking clustering method [Sasaki et al., 2014] and (b) $\widehat{\boldsymbol{\Sigma}}_{\mathrm{LS}}$ from LSDDR. To understand how these modifications improved the performance in ridge estimation, we performed experiments with the following additional methods:

- LSDRF: $\widehat{\boldsymbol{m}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{LS}}$ are used.
- (LS+KDE): $\widehat{\boldsymbol{m}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{KDE}}$ are used.
- (MS+LSDDR): $\widehat{\boldsymbol{m}}_{\mathrm{MS}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{LS}}$ are used.
- SCMS: $\widehat{\boldsymbol{m}}_{\mathrm{MS}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{KDE}}$ are used.

Figures 2(d) show that (LS+KDE) and (MS+LSDDR) improve SCMS, while LSDRF performs better than (LS+KDE) and (MS+LSDDR). Therefore, both $\widehat{\boldsymbol{\Sigma}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{m}}_{\mathrm{LS}}$ surely contribute to improving the performance of SCMS. However, the amount of performance improvement obtained from $\widehat{\boldsymbol{m}}_{\mathrm{LS}}$ or from $\widehat{\boldsymbol{\Sigma}}_{\mathrm{LS}}$ seems dependent on datasets.

5.2 Density Ridge Visualization on Real-World Datasets

Finally, we apply LSDRF to real-world datasets. As in Pulkkinen [2015], we employed the following two datasets:

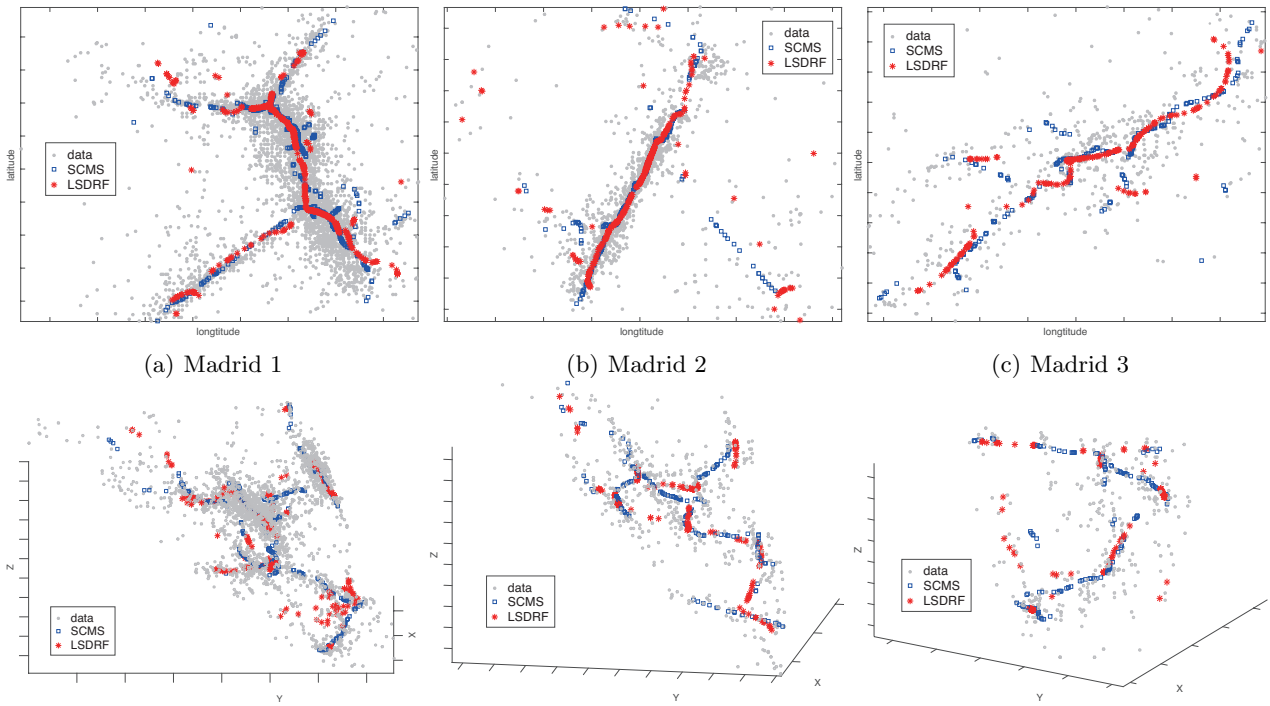
- *New Madrid earthquake* dataset: This seismological dataset was downloaded from the Center for Earthquake Research and Information.[4] The dataset contains positional information for earthquakes around the New Madrid seismic zone from 1974 to 2016, providing 11,131 samples. The three regions in Figures 4(a,b,c) were extracted according to (a) $(-90.2, -89.25)$, (b) $(-92.5, -92.15)$ and (c) $(-85.5, -83.5)$ degrees for the latitude range. For the longitude range, (a) $(36, 36.8)$, (b) $(35.2, 35.4)$ and (c) $(34.5, 36.5)$ degrees were selected. The total number of the original data samples and reduced data samples in each region was (a) $(n, n') = (5902, 500)$, (b) $(n, n') = (1548, 300)$ and (c) $(n, n') = (594, 200)$.

- *Shapley galaxy* dataset: This dataset was downloaded from the Center for Astrostatistics at Pennsylvania State University.[5] The dataset contains information about the three-dimensional sky angles and recession velocity of 4,215 galaxies. As done in Pulkkinen [2015], we transformed the data samples into the three-dimensional Cartesian coordinates based on the fact that the recession velocity is proportional to the radial distance [Drinkwater et al., 2004]. The three regions in Figures 4(a,b,c) were extracted according to a velocity range: (a) $(6000, 20000)$ km/s, (b) $(1500, 6000)$ km/s and (c) $(6000, 10500)$ km/s, respectively. The total number of the original data samples and reduced data samples in each region was (a) $(n, n') = (2849, 500)$, (b) $(n, n') = (595, 200)$ and (c) $(n, n') = (351, 150)$.

In each dataset, we focused on three regions containing prominent features, and standardized data samples in each region by subtracting the mean value and dividing by standard deviation in a dimension-wise manner. For performance comparison, we computed the log-likelihood of density ridge estimates, which is defined by $\mathcal{L} = \frac{1}{n'} \sum_{l=1}^{n'} \log \widehat{p}_{\mathrm{KDE}}(\widehat{\boldsymbol{y}}_l)$: The kernel centers in $\widehat{p}_{\mathrm{KDE}}$ were set at data samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ in each region, while density ridges estimates $\{\widehat{\boldsymbol{y}}_l\}_{l=1}^{n'}$ were obtained from $n' (< n)$ data samples randomly chosen from $\{\boldsymbol{x}_i\}_{i=1}^{n}$. If $\mathcal{L}$ is larger, the performance can be interpreted to be better because ridges are defined on relatively high density areas. Unlike the last experiment, we used the following adaptive-bandwidth Gaussian kernel in LSDDR: The bandwidth parameter of each Gaussian kernel was set at the Euclidean distance to the m-nearest sample from the Gaussian center where m was cross-validated. The regularization parameter λ_k, which is supposed to be common to all $\boldsymbol{j}$ in this section, was also determined by cross-validation as in

[4] http://www.memphis.edu/ceri/seismic/

[5] http://astrostatistics.psu.edu/datasets/Shapley_galaxy.html



(a) Madrid 1 (b) Madrid 2 (c) Madrid 3

(d) Shapley 1 (e) Shapley 2 (f) Shapley 3

Figure 4: Density ridge estimation to the (a,b,c) New Madrid earthquake and (d,e,f) Shapley galaxy datasets. Three regions were extracted from each dataset according to (a,b,c) a range of latitude and longitude, and (d,e,f) a range of recession velocity.

Section 3.3. The adaptive-bandwidth Gaussian kernel was also used for SCMS, and a similar cross-validation procedure was performed.

Ridges estimated by LSDRF are often smooth and seem to well-match the ridges of the underlying data (Figure 4). Table 1 quantitatively substantiates that LSDRF overall performs better than SCMS.

6 Conclusion

In this paper, we proposed a new algorithm for density ridge estimation. Our main contribution was the *least-squares density-derivative-ratios* (LSDDR) estimator, which avoids density estimation and division by an estimated density. We theoretically established a non-parametric convergence bound for LSDDR and experimentally demonstrated the superior performance of the density ridge estimator constructed based on LSDDR.

Acknowledgements

HS was supported by KAKENHI 15H06103, TK was supported by KAKENHI 16K00044 and MS acknowledges the JST CREST program

Table 1: The average and standard deviation of the log-likelihood of density ridges over 50 runs. A larger value means a better result. Numbers in the parentheses are standard deviations. The best and comparable methods judged by the unpaired t-test at the significance level 5% are described in boldface.

New Madrid earthquake		
	LSDRF	SCMS
Madrid 1	**-0.611(0.109)**	**-0.632(0.056)**
Madrid 2	**-0.004(0.108)**	-0.051(0.086)
Madrid 3	**-1.097(0.146)**	-1.238(0.086)
Shapley galaxy		
	LSDRF	SCMS
Shapley 1	**0.125(0.101)**	0.038(0.082)
Shapley 2	-1.285(0.113)	**-1.216(0.089)**
Shapley 3	**-1.252(0.197)**	-1.550(0.086)

References

Y.-C. Chen, C. R. Genovese, S. Ho, and L. Wasserman. Optimal ridge detection using coverage risk. In *Advances in Neural Information Processing Systems 28*, pages 316–324. 2015a.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015b.

Y.-C. Chen, S. Ho, P. Freeman, C. Genovese, and L. Wasserman. Cosmic web reconstruction through density ridges: Method and algorithm. *Monthly Notices of the Royal Astronomical Society*, 454(1): 1140–1156, 2016.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

D. D. Cox. A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Annals of the Institute of Statistical Mathematics*, 37(1):271–288, 1985.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

M. J. Drinkwater, Q. A. Parker, D. Proust, E. Slezak, and H. Quintana. The large scale distribution of galaxies in the shapley supercluster. *Publications of the Astronomical Society of Australia*, 21(1):89–96, 2004.

D. Eberly. *Ridges in Image and Data Analysis*. Springer, 1996.

J. Einbeck and J. Dwyer. Using principal curves to analyse traffic patterns on freeways. *Transportmetrica*, 7(3):229–246, 2011.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

Y. A. Ghassabeh, T. Linder, and G. Takahara. On some convergence properties of the subspace constrained mean shift. *Pattern Recognition*, 46(11): 3140–3147, 2013.

T. Hastie. *Principal Curves and Surfaces*. PhD thesis, Stanford University, 1984.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

B. Kégl and A. Krzyźak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1): 59–74, 2002.

U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286, 2011.

S. Pulkkinen. Ridge-based method for finding curvilinear structures from noisy data. *Computational Statistics & Data Analysis*, 82:89–109, 2015.

H. Sasaki, A. Hyvärinen, and M. Sugiyama. Clustering via mode seeking by direct estimation of the gradient of a log-density. In *Machine Learning and Knowledge Discovery in Databases Part III- European Conference, ECML/PKDD 2014*, volume 8726, pages 19–34, 2014.

H. Sasaki, Y. K. Noh, and M. Sugiyama. Direct density-derivative estimation and its application in KL-divergence approximation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 809–818, 2015.

J. Su, A. Srivastava, and F. Huffer. Detection, classification and estimation of individual shapes in 2D and 3D point clouds. *Computational Statistics & Data Analysis*, 58:227–241, 2013.

S. You, E. Bas, D. Erdogmus, and J. Kalpathy-Cramer. Principal curved based retinal vessel segmentation towards diagnosis of retinal diseases. In *IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*, pages 331–337. IEEE, 2011.

D. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, 2008.