Sparse Approximation of Kernel Density Estimators via a Genetic Algorithm: Data-Size Compression

Kiheiji NISHIDA
Kyoto Sangyo University

§1.Introduction

Suppose that $\{\mathbf{X}_i\}_{i=1}^N$ is a d-dimensional i.i.d. sample of size N generated from the true density function $f(\mathbf{x})$ on $\mathbb{R}^d$, where $\mathbf{x}^\top = (x_1, x_2, ..., x_d)$ and $\mathbf{X}_i^\top = (X_{i1}, X_{i2}, ..., X_{id})$, and let $\mathbf{X}$ denote the corresponding $N \times d$ data matrix. We estimate $f(\mathbf{x})$ using the multivariate kernel density estimator (KDE); its general representation is written as

$$\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i) \quad \equiv \quad \boldsymbol{\alpha}^\top \mathbf{k}_{\mathbf{H}}\big(\mathbf{x}|\mathbf{X}\big), \tag{1}$$

where $\mathbf{H}$ is a symmetric and positive definite d-dimensional bandwidth matrix, $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}}\mathbf{x})$ is a non-negative real valued bounded kernel function, $\boldsymbol{\alpha}^\top = (\alpha_1, \alpha_2, ..., \alpha_N)$ is the vector of the weighting coefficient parameters assigned to the data points $\mathbf{X}_i$ satisfying $\boldsymbol{\alpha}\mathbf{1}_N = 1$, and the vector $\mathbf{k}_{\mathbf{H}}\big(\mathbf{x}|\mathbf{X}\big)^\top = \big(K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_1), K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_2), ..., K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_N)\big)$. There are two main approaches to perform KDE: the Parzen window density estimator (PE) in Parzen (1962) and the Reduced Set Density Estimator (RSDE) approach in Girolami and He (2003). In the PE approach, the weighting coefficient parameters are set to be $\boldsymbol{\alpha}_{\mathrm{PE}}^\top \equiv (1/N, 1/N, ..., 1/N)$, and the main focus is on efficiently estimating the bandwidth $\mathbf{H}$. In contrast, the RSDE approach treats the data and the bandwidth as given and optimizes the coefficient $\boldsymbol{\alpha}$ so that they best fit the data in terms of Integrated Squared Error (ISE)

written as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \widehat{ISE}\left(\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}}(\cdot), f(\cdot)\middle|\mathbf{X}\right) \\
&= \min_{\boldsymbol{\alpha}} \int_{\mathbb{R}^d} \left[\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}}(\mathbf{t}) - f(\mathbf{t})\right]^2 d\mathbf{t} \\
&= \min_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \mathbf{Q}\, \boldsymbol{\alpha} - 2\,\mathbf{q}^\top \boldsymbol{\alpha} + R(f), \qquad (2) \\
&\quad \text{s.t.} \quad \boldsymbol{\alpha} \geq \mathbf{0_N}, \quad \boldsymbol{\alpha}^\top \mathbf{1_N} = 1,
\end{aligned}$$

where the (i,j)th element of the matrix $\mathbf{Q}$ is $Q_{ij} = \int_{\mathbb{R}^d} K_{\mathbf{H}}(\mathbf{t} - \mathbf{X}_i) K_{\mathbf{H}}(\mathbf{t} - \mathbf{X}_j)\, d\mathbf{t}$, the ith element of the vector $\mathbf{q}$ is $q_i = \int_{\mathbb{R}^d} f(\mathbf{t})\, K_{\mathbf{H}}(\mathbf{t} - \mathbf{X}_i)\, d\mathbf{t}$ and $R(f)$ is the integral of the squared function f over $\mathbb{R}^d$. The equation (2) becomes a quadratic function of the coefficients $\{\alpha_i\}_{i=1}^N$, since the data, bandwidth, and kernel are fixed and all terms involving $\mathbf{t}$ are integrated out, leaving an expression that depends on $\boldsymbol{\alpha}$ only through their products.

RSDE requires two steps for implementation. In the first stage, the bandwidth matrix $\mathbf{H} = h^2 \mathbf{I}_d$ is optimized using the Least Squares Cross-Validation (LSCV) in Rudemo (1982) and Bowman (1984) with the coefficients fixed at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{PE}}$. We denote this bandwidth by $\mathbf{H}_{\mathrm{R}}$. In the second stage, RSDE optimizes $\widehat{ISE}(\mathbf{H}_{\mathrm{R}}, \boldsymbol{\alpha})$ under the constraints $\boldsymbol{\alpha} \geq \mathbf{0_N}, \boldsymbol{\alpha}^\top \mathbf{1_N} = 1$ with respect to $\boldsymbol{\alpha}$ while keeping the bandwidth $\mathbf{H}$ fixed at $\mathbf{H}_{\mathrm{R}}$, and we denote the resulting coefficients by $\boldsymbol{\alpha}_{\mathrm{R}}$. This optimization problem is a convex quadratic programming problem, and its solution is guaranteed to be a corner solution. Consequently, any data points whose corresponding coefficients become zero are excluded from the KDE, so that the estimator is effectively expressed in a sparse form with respect to data size. In its implementation, Girolami and He (2003) propose two algorithms: Multiplicative Updating of the weighting coefficients and Sequential Minimal Optimization for RSDE.

This two-stage scheme of RSDE arises because $\widehat{ISE}(\mathbf{H}, \boldsymbol{\alpha})$ is *non-jointly convex*: it is *convex* in $\boldsymbol{\alpha}$ but *non-convex* in $\mathbf{H}$. For such non-jointly convex problems, alternating minimization is known to converge to a coordinate-wise minimum under mild regularity conditions (Tseng 2001; Grippo and Sciandrone 2000). However, the resulting stationary point does not necessarily coincide with the global minimizer that would be obtained through joint optimization over $(\mathbf{H}, \boldsymbol{\alpha})$. It follows that the RSDE value of $\widehat{ISE}(\mathbf{H}_{\mathrm{R}}, \boldsymbol{\alpha}_{\mathrm{R}})$ may be further improved by performing joint optimization over $(\mathbf{H}, \boldsymbol{\alpha})$, but such a joint optimization is generally difficult. Hence, we adopt a *metaheuristic* optimization algorithm, such as a genetic algorithm (GA) (e.g. Haupt and Haupt 2004), which does not guarantee the global optimum but provides a reasonable solution, to obtain a sparse KDE optimized over $(\mathbf{H}, \boldsymbol{\alpha})$.

§2.The proposed GA

The proposed GA is described as follows. Let $V(\mathbf{D}, \mathbf{H})$ be a fitness function that assesses the KDE denoted as $\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}}(\mathbf{x}|\mathbf{D})$, where $\mathbf{D} = \{\mathbf{X}_i\}_{i=1}^{p}$ denotes the set of p kernel centers and $\mathbf{H}$ is the bandwidth matrix.

Step 1: Initial generation $g = 1$:

1. Define the size of subsample b, number of subsamples B, final generation number G. Then, $\{\beta_i\}_{i=1}^{b} = \{1/b\}_{i=1}^{b}$

2. Make the number of B subsamples of size b with replacement from the original sample of size N, where B is an even number. Each subsample is called *chromosome* and is denoted as $\mathbf{D}_i^{(1)} = \{\mathbf{X}_{i,1}^{(1)}, \mathbf{X}_{i,2}^{(1)}, ..., \mathbf{X}_{i,b}^{(1)}\}$, $i = 1, 2, ..., B$, where $\mathbf{X}_{i,j}^{(1)}$, $j = 1, 2, ..., b$, is the j-th data point of the i-th subsample called *gene*. The *population* in generation 1 is written as $\mathbf{D}^{(1)} = \{\mathbf{D}_1^{(1)}, \mathbf{D}_2^{(1)}, ..., \mathbf{D}_B^{(1)}\}$.

Step 2: The generations $g = 2, 3, ..., G$:

1. Inherit population $\mathbf{D}^{(g-1)} = \{\mathbf{D}_1^{(g-1)}, \mathbf{D}_2^{(g-1)}, ..., \mathbf{D}_B^{(g-1)}\}$ from the previous generation $g - 1$.

2. For each subsample $\mathbf{D}_i^{(g-1)}$, $i = 1, 2, ..., B$, calculate the fitness value $V(\mathbf{D}_i^{(g-1)}, \mathbf{H}_i^{(g-1)})$ along with the optimal bandwidth matrix $\mathbf{H}_i^{(g-1)}$. Then, sort the elements in $\mathbf{D}^{(g-1)} = \{\mathbf{D}_1^{(g-1)}, \mathbf{D}_2^{(g-1)}, ..., \mathbf{D}_B^{(g-1)}\}$ in descending order according to their fitness values $V(\mathbf{D}_i^{(g-1)}, \mathbf{H}_i^{(g-1)})$, $i = 1, 2, ..., B$, and rename the resulting sequence as $\mathbf{D}^{(g)} = \{\mathbf{D}_1^{(g)}, \mathbf{D}_2^{(g)}, ..., \mathbf{D}_B^{(g)}\}$.

3. Make the replica $\mathbf{D}^{+(g)} \equiv \mathbf{D}^{(g)}$.

4. Breed two new subsamples using the pair of subsamples $\mathbf{D}_{2k-1}^{(g)}$ and $\mathbf{D}_{2k}^{(g)}$, $k = 1, 2, ..., B/2$; each pair of data points $\mathbf{X}_{2k-1,j}^{(g)}$ and $\mathbf{X}_{2k,j}^{(g)}$, $j = 1, 2, ..., b$, faces either of the following with a certain probability.

   (i) *Mutation* : With mutation probability p_m, $\mathbf{X}_{2k-1,j}^{(g)}$ and $\mathbf{X}_{2k,j}^{(g)}$ are replaced with the two data points randomly chosen from $\mathbf{D}_1^{+(g)} = \{\mathbf{X}_{1,1}^{+(g)}, \mathbf{X}_{1,2}^{+(g)}, ..., \mathbf{X}_{1,b}^{+(g)}\}$.

   (ii) *Uniform crossover* : $\mathbf{X}_{2k-1,j}^{(g)}$ is swapped for $\mathbf{X}_{2k,j}^{(g)}$ with crossover probability p_u.

   (iii) *Reproduction* : $\mathbf{X}_{2k-1,j}^{(g)}$ and $\mathbf{X}_{2k,j}^{(g)}$ remain unchanged with probability $1 - p_u - p_m$.

5. For each renewed subsample $\mathbf{D}_i^{(g)}$, for $i = 1, 2, ..., B$, calculate the fitness value $V(\mathbf{D}_i^{(g)}, \mathbf{H}_i^{(g)})$ along with the optimal bandwidth matrix $\mathbf{H}_i^{(g)}$. Then, sort the renewed subsamples in descending order by their renewed fitness values, and rename the resulting sequence as $\mathbf{D}^{*(g)} = \{\mathbf{D}_1^{*(g)}, \mathbf{D}_2^{*(g)}, ..., \mathbf{D}_B^{*(g)}\}$.

6. The renewed population $\mathbf{D}^{(g)} = \{\mathbf{D}_1^{+(g)}, \mathbf{D}_2^{+(g)}, ..., \mathbf{D}_{p_eB}^{+(g)}, \mathbf{D}_1^{*(g)}, \mathbf{D}_2^{*(g)}, ..., \mathbf{D}_{(1-p_e)B}^{*(g)}\}$ is taken over to generation $g+1$, where p_e is the ratio of the number of subsamples in B inherited by the next generation according to the *elite selection* rule.

Step 3: Completion of the algorithm at $g = G + 1$:

1. Accept the KDE exhibiting the best fitness value $V(\mathbf{D}^*, \mathbf{H}^*)$ at the generation G written as

$$\widehat{f}_{\mathbf{H}^*,\boldsymbol{\beta}}(\mathbf{x}|\mathbf{D}^*) = \sum_{i=1}^{b} \beta_i K_{\mathbf{H}^*}(\mathbf{x} - \mathbf{X}_i^*),$$

   along with the resulting subsample $\mathbf{D}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_b^*\}$ and bandwidth matrix $\mathbf{H}^*$.

Remark 1. We modify the LSCV to our fitness function. Let $I(\cdot)$ denote an indicator function. We consider ISE written as

$$\begin{aligned}
&\widehat{ISE}\big(\widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\cdot|\mathbf{D}_i^{(g)}), f(\cdot)\big) \\
&= \int_{\mathbb{R}^d} \left[\widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\mathbf{t}|\mathbf{D}_i^{(g)}) - f(\mathbf{t})\right]^2 d\mathbf{t} \\
&= \int_{\mathbb{R}^d} \widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\mathbf{t}|\mathbf{D}_i^{(g)})^2 d\mathbf{t} - 2\int_{\mathbb{R}^d} \widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\mathbf{t}|\mathbf{D}_i^{(g)}) f(\mathbf{t}) d\mathbf{t} + R(f). \qquad (3)
\end{aligned}$$

Replacing the second term in (3) with its empirical form and excluding the third term, we obtain the fitness function

$$\begin{aligned}
&-V(\mathbf{D}_i^{(g)}, \mathbf{H}) \\
&= \int_{\mathbb{R}^d} \widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\mathbf{t}|\mathbf{D}_i^{(g)})^2 d\mathbf{t} - \frac{2}{b(N-1)} \sum_{j=1}^{N} \sum_{i=1}^{b} I(\mathbf{X}_j \neq \mathbf{X}_i^{(g)}) K_{\mathbf{H}}(\mathbf{X}_j - \mathbf{X}_i^{(g)}), \quad (4)
\end{aligned}$$

where the second term in (4) is the sample mean of $\widehat{f}_{\mathbf{H},\boldsymbol{\beta}}(\mathbf{t}|\mathbf{D}_i^{(g)})$ over the original sample $\mathbf{t} \in \{\mathbf{X}_i\}_{i=1}^N$. Here, the kernel centers constitute a subset of

the original sample and the KDE constructed from these centers is evaluated over the original observations, with coinciding points excluded according to a leave-one-out rule. Kernel-center-based leave-one-out averaging collapses the effective sample size to $b-1$, causing variance inflation and degeneracy of the KDE when $b \ll N$, which is equivalent to estimating the density from a severely reduced sample. Averaging over the original sample decouples estimation from evaluation, preserving a large effective sample size for the LSCV criterion and thereby preventing the degeneracy induced by kernel-center-based averaging.

Remark 2. Let us define $\gamma(i) = \sum_{j=1}^{b} I(\mathbf{X}_j^* = \mathbf{X}_i)\beta_j$, $i = 1, 2, ..., N$. The resulting KDE is then written as

$$\begin{aligned} \widehat{f}_{\mathbf{H}^*,\boldsymbol{\beta}}(\mathbf{x}|\mathbf{D}^*) &= \sum_{j=1}^{b} \beta_j K_{\mathbf{H}^*}(\mathbf{x} - \mathbf{X}_j^*) \\ &= \sum_{i=1}^{N} \gamma(i) K_{\mathbf{H}^*}(\mathbf{x} - \mathbf{X}_i), \end{aligned}$$

where the algorithm adjusts $\gamma(i)$, $i = 1, 2, ..., N$, to minimise the fitness value through data-point-specific weighting.

Remark 3. The PE weight $\boldsymbol{\alpha}_{\mathrm{PE}}$ yields the minimum possible variance of (1) for any new, unseen dataset. If we suppose $\max_i \alpha_i = O(N^{-\delta}), 0 < \delta < 1$, this follows directly from

$$\begin{aligned} \mathrm{Var}_{\mathbf{X}}[\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}}(\mathbf{x}|\mathbf{X})] &= \|\boldsymbol{\alpha}\|_2^2 \frac{f(\mathbf{x})}{|\mathbf{H}|} R(K) + O\Big(N^{-\delta}\Big) \\ &\geq \frac{f(\mathbf{x})}{N|\mathbf{H}|} R(K) + O\Big(\frac{1}{N}\Big) = \mathrm{Var}_{\mathbf{X}}[\widehat{f}_{\mathbf{H},\boldsymbol{\alpha}_{\mathrm{PE}}}(\mathbf{x}|\mathbf{X})], \end{aligned}$$

where the variance attains its lower bound precisely when $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{PE}}$. Therefore, tailoring the weights to the particular observed data does not lead to a variance-minimizing KDE for future data that have not yet been observed. In contrast, such weight adjustments do not influence the bias because the expectation of (1) with respect to $\mathbf{X}$ does not involve $\boldsymbol{\alpha}$.

§3.Numerical experiments

We demonstrate the proposed GA by Monte Carlo simulations under the same settings as in Nishida (2023), assuming a scalar bandwidth matrix

$\mathbf{H} = h^2\mathbf{I}_d$ and compare it with the Direct Plug-in method with full bandwidth matrix in Duong and Hazelton (2003) (henceforth, DPI) and RSDE. The set of GA parameters is $(B, G, p_u, p_m, p_e) = (50, 100, 0.475, 0.05, 0.1)$. As the true density, we use the Type J Trimodal II in Wand and Jones (1993), whose contour plot is shown in the upper-right panel of Figure 1. In the simulations, 10 KDEs are generated from a single sample using the proposed GA, and their average ISE is denoted by ISE^*. In calculating the MISE for DPI and RSDE, we generate 10 independent samples of size N, compute the ISE for each sample, and take their average as the MISE; one of these samples is identical to that used to compute the ISE* of our GA. In implementing RSDE, we employ multiplicative updating of the weighting coefficients.

The numerical results are given in Table 1 and are visually summarized in Figure 1. The upper-left panel in Figure 1 shows a representative contour plot of the proposed density estimator based only on the data points selected by our GA (asterisks), with the original data shown as gray dots. The bottom-left and bottom-right panels in the same figure show bar plots of the GA-selected data points (gray bars), together with a rug plot of the original data below the horizontal axis, for the x_1 and x_2 axes, respectively, where the vertical axes indicate the number of duplicated data points. We show the numerical results of data condensation ratio (DCR) in Table 2.

We observe that our GA achieves a smaller estimation error than its competitors, while simultaneously attaining a smaller DCR.

§4.Discussion

RSDE optimizes the ISE with respect to the coefficient parameters under a fixed bandwidth, and therefore does not necessarily yield a globally optimal pair $(\mathbf{H}, \boldsymbol{\alpha})$. In contrast, we design our GA to optimize the ISE with the aim of attaining a certain level of global optimality, albeit in a stochastic sense. As pointed out in Remark 3, the Parzen weights are variance-optimal for future samples, whereas the data-dependent weights produced by our GA are not. Accordingly, our aim is not prediction optimality, but ISE optimality conditional on the observed sample.

The simulation results suggest that our method outperforms its competitors in terms of both estimation error and data condensation rate. When performing statistical analyses based on KDE estimates, the substantial reduction in data size can significantly lower the overall computational cost. Moreover, since the KDE is constructed using only a very small number of data points as kernel centers, the individual data points not used in the estimation can be discarded, thereby enhancing data anonymity.

g	1	50	75	100	DPI	RSDE	DPI*	RSDE*
$\underline{N = 200}$	—	—	—	—	1113 (227)	2062 (640)	913	1727
$b = 2$	3509 (472)	2891 (288)	2891 (288)	2891 (288)				
$b = 5$	2618 (415)	1529 (380)	1529 (380)	491 (80)				
$b = 25$	1588 (152)	529 (129)	522 (91)	491 (80)				
$b = 50$	1135 (182)	406 (152)	416 (106)	384 (91)				
$b = 100$	755 (118)	427 (110)	366 (109)	408 (119)				
$b = 150$	734 (80)	385 (106)	371 (85)	351 (80)				
$\underline{N = 400}$	—	—	—	—	713 (116)	1637 (510)	644	1169
$b = 2$	3780 (737)	2923 (397)	2923 (397)	2923 (397)				
$b = 5$	2942 (450)	1681 (235)	1681 (235)	1681 (235)				
$b = 25$	1683 (540)	984 (226)	933 (185)	898 (181)				
$b = 50$	1217 (241)	880 (288)	832 (260)	834 (216)				
$b = 100$	1223 (292)	799 (223)	773 (246)	786 (279)				
$b = 150$	1136 (3310)	884 (210)	919 (203)	982 (263)				
$\underline{N = 1000}$	—	—	—	—	396 (55)	910 (202)	378	966
$b = 2$	3509 (472)	2891 (288)	2891 (288)	2891 (288)				
$b = 5$	2618 (415)	1529 (380)	1529 (380)	491 (80)				
$b = 25$	1588 (152)	529 (129)	522 (91)	491 (80)				
$b = 50$	1135 (182)	406 (152)	416 (106)	384 (91)				
$b = 100$	755 (118)	427 (110)	366 (109)	408 (119)				
$b = 150$	735 (80)	385 (106)	371 (85)	351 (80)				

Table 1: [Results on estimation error: Type J] Results of estimation error $\mathrm{ISE}^*(g)$ $\times 10^5$ (S.D. $\times 10^5$). The numbers in the DPI and RSDE columns are MISE $\times 10^5$ (S.D. $\times 10^5$). The numbers in the columns of DPI* and RSDE* are ISE $\times 10^5$ calculated by the identical sample used in calculating ISE*. The minimum values of $\mathrm{ISE}^*(g)$ over the sizes of b are underlined.

	(I)	(II)	(II)/b	DCR.GA	DCR.RSDE
$\underline{N = 200}$	—	—	—	—	.2375 (.0330)
$b = 25$	14.4 (2.07)	3.80 (0.79)	0.15	.0720 (.0103)	—
$b = 50$	27.0 (2.21)	4.70 (1.49)	0.09	.1350 (.0111)	—
$b = 100$	39.9 (5.84)	8.90 (2.38)	0.09	.1995 (.0269)	—
$b = 150$	52.4 (5.32)	10.6 (1.96)	0.07	.2620 (.0266)	—
$\underline{N = 400}$	—	—	—	—	.2008 (.0150)
$b = 25$	15.5 (1.78)	3.60 (0.70)	0.14	.0388 (.0044)	—
$b = 50$	29.2 (4.44)	4.90 (0.88)	0.10	.0730 (.0111)	—
$b = 100$	49.0 (4.55)	7.10 (0.88)	0.07	.1225 (.0114)	—
$b = 150$	69.2 (5.81)	9.00 (2.62)	0.06	.1730 (.0145)	—
$\underline{N = 1000}$	—	—	—	—	.1405 (.0121)
$b = 25$	18.2 (2.30)	2.70 (0.67)	0.11	.0182 (.0023)	—
$b = 50$	31.7 (2.98)	4.10 (0.88)	0.08	.0317 (.0030)	—
$b = 100$	59.7 (5.91)	5.10 (0.74)	0.05	.0597 (.0059)	—
$b = 150$	78.3 (10.9)	6.50 (1.43)	0.04	.0783 (.0109)	—

Table 2: [Results on data condensation : Type J] (I): The average number of distinct data points used for density estimation (S.D.). (II): The maximum data-point multiplicity, averaged over replications. (S.D.).

As future work, we plan to investigate variable selection for KDE in high-dimensional settings using a GA. Here, we consider a setting in which a set of relevant variables is mixed with irrelevant ones, and aim to efficiently extract the relevant subset so as to identify variables that are meaningful for KDE. Following the framework for data condensation in Nishida (2023), one possible implementation of this idea is given as follows:

$$\begin{aligned} \widehat{f}_{\mathbf{H},\boldsymbol{\tau}}(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{d} K_{h_j}(x_j - X_{ij})^{\tau_j}, \\ \tau_j &\in \{0, 1, 2, \ldots, b(N)\}, \\ \sum_{j=1}^{d} \tau_j &= b(N), \end{aligned}$$

where $\mathbf{H} = \mathrm{diag}(h_1^2, h_2^2, ..., h_d^2)$, $\boldsymbol{\tau}^\top = (\tau_1, \tau_2, ..., \tau_d)$, $b(N) \ll d$, and the parameter vector $\boldsymbol{\tau}$ can be determined using a metaheuristic optimization algorithm.

Acknowledgements

This paper includes material omitted from Nishida (2023), along with additional findings developed after its publication. This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University, and by JSPS KAKENHI Grant Number 23K28043.

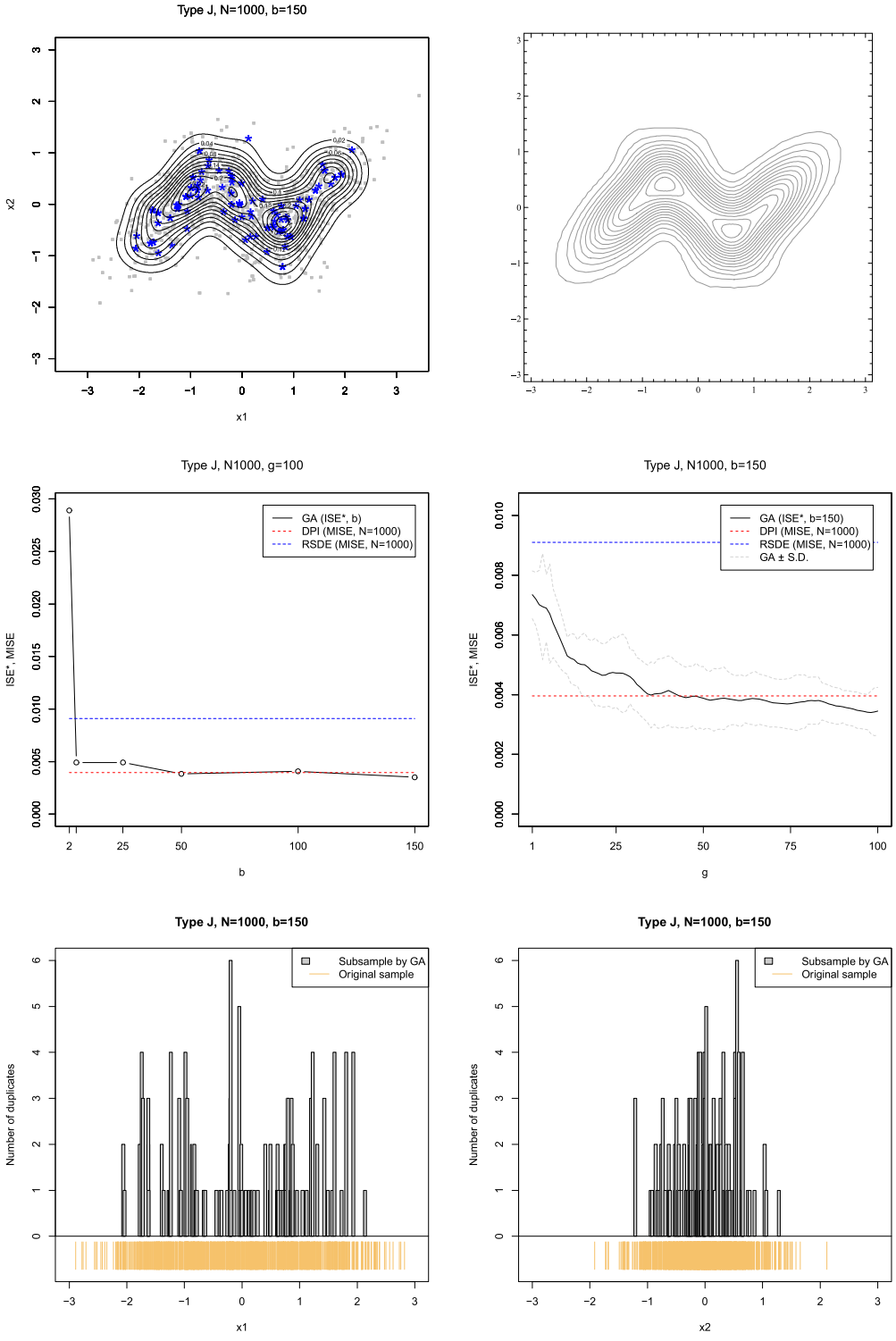

Figure 1: [Graphical results]: Type J. Trimodal II:

References

[1] Bowman, A. (1984). An Alternative Method of Cross-validation for the Smoothing of Density Estimates. *Biometrika*, **71**(2), pp.353-360.

[2] Duong, T., and Hazelton, M.L. (2003). Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation, *Journal of Nonparametric Statistics*, **15**(1), pp.17-30.

[3] Girolami, M., and He, C. (2003). Probability Density Estimation from Optimally Condensed Data Samples, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(10), pp.1253-1264.

[4] Grippo, L., & Sciandrone, M.(2000). On the Convergence of the Block Nonlinear Gauss Seidel Method under Convex Constraints, *Operations Research Letters*, **26**(3), pp.127-136.

[5] Haupt, R.L., and Haupt, S.E.(2004). Practical Genetic Algorithms, Second edition. *Wiley*.

[6] Nishida, K.(2023). Kernel Density Estimation by Genetic Algorithm, *Journal of Statistical Computation and Simulation*, **93**(8), pp.1263-1281.

[7] Parzen, E.(1962). On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics*, **33**(3), pp.1065-1076.

[8] Rudemo, M. (1982) Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, **9**, pp.65-78.

[9] Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization, *Journal of Optimization Theory and Applications*, **109**(3), pp.475-494.

[10] Wand, M.P., and Jones, M.C. (1993). Comparison of Smoothing Parametrizations in Bivariate Kernel Density Estimation, *Journal of the American Statistical Association*, **88**(422), pp.520-528.

Kiheiji NISHIDA (西田 喜平次)
Department of Business Administration
Kyoto Sangyo University
Kamigamo-Motoyama, Kita-Ku, Kyoto, 603-8047, JAPAN
E-mail address: kiheiji.nishida@gmail.com