High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation

Rafael Izbicki, Ann B. Lee, Chad M. Schafer
Department of Statistics – Carnegie Mellon University

Abstract

The ratio between two probability density functions is an important component of various tasks, including selection bias correction, novelty detection and classification. Recently, several estimators of this ratio have been proposed. Most of these methods fail if the sample space is high-dimensional, and hence require a dimension reduction step, the result of which can be a significant loss of information. Here we propose a simple-to-implement, fully nonparametric density ratio estimator that expands the ratio in terms of the eigenfunctions of a kernel-based operator; these functions reflect the underlying geometry of the data (e.g., submanifold structure), often leading to better estimates without an explicit dimension reduction step. We show how our general framework can be extended to address another important problem, the estimation of a likelihood function in situations where that function cannot be well-approximated by an analytical form. One is often faced with this situation when performing statistical inference with data from the sciences, due the complexity of the data and of the processes that generated those data. We emphasize applications where using existing likelihood-free methods of inference would be challenging due to the high dimensionality of the sample space, but where our spectral series method yields a reasonable estimate of the likelihood function. We provide theoretical guarantees and illustrate the effectiveness of our proposed method with numerical experiments.

1 INTRODUCTION

There has been growing interest in the problem of estimating the ratio of two probability densities, \( \beta(x) \equiv f(x)/g(x) \), given i.i.d. samples from unknown distributions \( F \) and \( G \). For example, these ratios play a key role in matching training and test data in so-called transfer learning or domain adaptation (Sugiyama et al., 2010a), where the goal is to predict an outcome \( y \) given test data \( x \) from a distribution \( G \) that is different from that of the training data \( F \). Estimated density ratios also appear in novelty detection (Hido et al., 2011), conditional density estimation (Sugiyama et al., 2010b), selection bias correction (Gretton et al., 2010), and classification (Nam et al., 2012).

Experiments have shown that it is suboptimal to estimate \( \beta(x) \) by first estimating the two component densities and then taking their ratio (Sugiyama et al., 2008). Hence, several alternative approaches have been proposed that directly estimate \( \beta(x) \); e.g., \( uLSIF \), an estimator obtained via least-squares minimization (Kanamori et al., 2009); \( KLIEP \), which is obtained via Kullback-Leibler divergence minimization (Sugiyama et al., 2008); \( KuLSIF \), a kernelized version of \( uLSIF \) (Kanamori et al., 2012); and \( kernel\ mean\ matching \), which is based on minimizing the mean discrepancy between transformations of the two samples in a Reproducing Kernel Hilbert Space (RKHS) (Gretton et al., 2010). For a review of techniques see Margolis (2011).

Existing methods are not effective when \( x \) is of high dimension, and hence authors recommend a dimension reduction prior to implementation (Sugiyama et al., 2011). As is the case with any data reduction, such a step can result in significant loss of information. Here we propose a novel series estimator of \( \beta(x) \) designed to take advantage of the intrinsic dimensionality of \( x \), but without an explicit dimension reduction step. Thus, this work addresses a critical need by constructing a nonparametric estimator for \( \beta(x) \) which performs well even when \( x \) is of high dimension.
Our method is fast and simple to implement. To approximate \( \beta(x) \), we expand \( \beta \) in terms of the eigenfunctions of a kernel-based operator. These eigenfunctions are orthogonal with respect to the underlying data distribution as opposed to the Lebesgue measure on the ambient space. In fact, the eigenfunctions form a Fourier-like basis adapted to the geometry of the data. This basis is well-suited for approximating smooth functions in the region around the spiral.

We extend our proposed methodology for estimating density ratios to the problem of estimating the likelihood function of observing data \( x \) given parameters \( \theta \). Estimation of the likelihood is necessary when the complexity of the data-generation process prevents direct estimating the ratio \( f(x|\theta)/g(x) \) for different \( \theta \) are similar, or when they have similar support in high dimensions. The organization of the paper is as follows: In Section 2 we present our density ratio estimator and apply it to a prediction problem in astronomy with covariate shift. In Section 3, we show how our method can be extended to estimating a likelihood function, and provide experiments that show its advantages over traditional methods. Finally, in Section 4, we provide theoretical guarantees and rates of convergence of the proposed estimators. Full proofs as well as details on the astronomy data are provided in Supplementary Materials.

## 2 SPECTRAL SERIES ESTIMATOR OF A DENSITY RATIO

In this section we will present the mathematical details behind our spectral series estimator of a density ratio. To begin, let \( x \) denote a \( d \)-dimensional random vector, assumed to lie in the subspace \( X \). We observe an \( i.i.d. \) sample \( x_1^F, \ldots, x_n^F \) from an unknown distribution \( F \), as well as an \( i.i.d. \) sample \( x_1^G, \ldots, x_{n^G}^G \) from an unknown distribution \( G \). The goal is to estimate

\[
\beta(x) \equiv f(x)/g(x).
\]

We assume that \( F \ll G \) so that this ratio is well-defined.

Let \( K_x(z, y) \) be a bounded, symmetric, and positive definite kernel\(^2\), and let \( \{\psi_j\}_{j \in \mathbb{N}} \) be the eigenfunctions

\(^1\)A trivial example: If \( x \) is independent of \( \theta \), \( f(x|\theta)/g(x) = 1 \) is a constant function, whereas \( f(x|\theta) = f(x) \) may be a harder to estimate (nonsmooth) function.

\(^2\)In our applications, we use the Gaussian kernel
of the operator $\mathbf{K}_x : L^2(\mathcal{X}, G) \rightarrow L^2(\mathcal{X}, G)$ (Rosasco et al., 2010):

$$\mathbf{K}_x(h)(z) = \int_{\mathcal{X}} K_x(z, y) h(y) dG(y).$$

(1)

Our spectral series estimator relies on the fact that \( \{\psi_j\}_{j \in \mathbb{N}} \) is an orthonormal basis of \( L^2(\mathcal{X}, G) \), i.e., the eigenfunctions are orthonormal with respect to the data distribution \( G \) rather than the Lebesgue measure:

$$\int_{\mathcal{X}} \psi_i(z) \psi_j(z) dG(z) = \mathbb{I}(i = j).$$

Hence, for \( \beta(x) \in L^2(\mathcal{X}, G) \), we can write

$$\beta(x) = \sum_{j \in \mathbb{N}} \beta_j \psi_j(x),$$

(2)

where \( \beta_j = \int \beta(x) \psi_j(x) dG(x) = E_F[\psi_j(X)] \).

Since \( G \) is unknown, the \( \psi_j \)'s must be estimated. First, we compute the \( J \leq n_G \) eigenvectors \( \tilde{\psi}_1, \ldots, \tilde{\psi}_J \) (with largest eigenvalues) of the Gram matrix based on the sample from \( G \),

$$[K_x(x^G_i, x^G_j)]_{i,j=1}^{n_G}.$$

These functions are then extended to all \( x \in \mathcal{X} \) via the Nyström Extension (Drineas and Mahoney, 2005)

$$\hat{\psi}_j(x) = \frac{\sqrt{n_G}}{\hat{\ell}_j} \sum_{k=1}^{n_G} \tilde{\psi}_j(x^G_k) K_x(x, x^G_k),$$

where \( \hat{\ell}_j \) is the eigenvalue associated to the eigenvector \( \tilde{\psi}_j \). Next we estimate the \( \beta_j \)'s in Eq. (2) using the sample from \( F \):

$$\hat{\beta}_j = \frac{1}{n_F} \sum_{k=1}^{n_F} \hat{\psi}_j(x^F_k).$$

Our spectral series estimator is finally given by

$$\hat{\beta}(x) = \left( \sum_{j=1}^{J} \hat{\beta}_j \hat{\psi}_j(x) \right) +=$$

(3)

This approach can be motivated as follows. The basis functions \( \hat{\psi}_j \) are consistent estimators of the eigenfunctions \( \psi_j \) of the corresponding integral operator (Bengio et al., 2004). In our nonparametric model, \( J \) is a tuning parameter that controls the bias/variance tradeoff: Decreasing \( J \) decreases the variance, but increases the bias of the estimator. We choose \( J \) (and the other tuning parameters) in a principled way described below.

**Model Selection.** To evaluate the performance of an estimator \( \hat{\beta}(x) \), we use the loss function

$$L(\hat{\beta}, \beta) = \int \left( \hat{\beta}(x) - \beta(x) \right)^2 dG(x)$$

$$= \int \hat{\beta}(x)^2 dG(x) - 2 \int \hat{\beta}(x) dF(x) + K$$

where \( K \) does not depend on \( \hat{\beta} \). We estimate this quantity (up to \( K \)) using

$$\hat{L}(\hat{\beta}, \beta) = \frac{1}{n_G} \sum_{k=1}^{n_G} \hat{\beta}^2(x^G_k) - 2 \frac{1}{n_F} \sum_{k=1}^{n_F} \hat{\beta}(x^F_k),$$

(4)

where \( x^G_1, \ldots, x^G_{n_G} \) is a validation sample from \( G \), and \( x^F_1, \ldots, x^F_{n_F} \) is a validation sample from \( F \). Tuning parameters are chosen to minimize \( \hat{L}(\hat{\beta}, \beta) \). Note that because of the orthogonality of the \( \hat{\psi}_j \), it is not necessary to recompute the estimated coefficients \( \hat{\beta}_j \)'s for each value of \( J \) in Eq. (3), unlike most estimation procedures, where estimated coefficients have to be recomputed for each configuration of the tuning parameters. In other words, only the tuning parameters associated with the kernel (in our case, the kernel bandwidth \( \epsilon \)) affect the computation time.

2.1 Application: Correction to Covariate Shift in Photometric Redshift Prediction

Assume we observe a sample of unlabeled data, as well as a sample of labeled data, where the Z’s represent the labels and x’s are the covariates. One is often interested in estimating the regression function \( E[Z|x] \) under selection bias, i.e., in situations where the distributions of labeled and unlabeled samples \( (f_L(x) \) and \( f_U(x) \), respectively) are different. If the estimate \( \hat{E}[Z|x] \) is constructed using the labeled data with the goal of predicting Z from x on the unlabeled data, corrections have to be made. A key quantity for making this correction under the covariate shift assumption (Shimodaira, 2000) is the density ratio \( f_U(x)/f_L(x) \), the so-called importance weights (Gretton et al., 2010). We now compare various estimators of importance weights for a key problem in astronomy, namely that of redshift estimation (Sheldon et al., 2012).

We use a subset of the Sloan Digital Sky Survey (Ai-hara et al., 2011). The ultimate goal is to build a predictor of galaxy redshift \( Z \) based on photometric data \( x \); see Supplementary Materials for details. We are given a training set with covariates \( x \) of galaxies and their redshifts, as well as unlabeled target
data. Because it is difficult to acquire the true redshift of faint galaxies, these data suffer from selection bias. We compare our method of estimating the importance weights (Series) to uLSIF, KLINK, and KuLSIF, described in Section 1. We also compute LHSS, which uses uLSIF after applying a dimension reduction technique specifically designed for estimating a density ratio, see Sugiyama et al. (2011). Moreover, we include a comparison with a k-nearest neighbors estimator (Lima et al. 2008), which is not based on ratios. We do not show results of ratio-based estimators because the estimates of $f_L(x)$ are close to zero for many x’s, inducing estimates of $\beta$ that are infinity.

Figure 2 shows the estimated losses of the different methods of estimating the ratio $f_U(x)/f_L(x)$ when using 5,000 labeled and 5,000 unlabeled samples, and 10 photometric covariates $x$. We use 60% of the data for training, 20% for validation and 20% for testing. Even though this example has a covariate space with as few as 10 dimensions, we can already see the benefits of the spectral series estimator.

3 EXTENSION: SPECTRAL SERIES ESTIMATOR OF A LIKELIHOOD FUNCTION

Our framework for estimating density ratios can be extended to the problem of estimating a high-dimensional likelihood function. To the derivation described above, we add $\theta \in \Theta$, the $p$-dimensional parameter. In this context, $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is a random vector representing a single sample observation.

We will adopt a Bayesian perspective, and let $F_\theta$ be the marginal distribution for $\theta$, i.e., the prior, and let $G$ denote the marginal distribution for $x$. Then, let $(x_1^p, \theta_1), \ldots, (x_{n_F}^p, \theta_{n_F})$ be an i.i.d. sample from the joint distribution of $x$ and $\theta$. Further, let $x_1^G, \ldots, x_{n_G}^G$ be an i.i.d. sample from $G$.

Our objective is to estimate the ratio

$$\mathcal{L}(x; \theta) \equiv \frac{f(x|\theta)}{g(x)},$$

where $f(x|\theta)$ is the conditional density of $x$ given $\theta$, and $g(x)$ is the marginal density for $x$. This is, up to a multiplicative factor that is not a function of $\theta$, the standard definition of the likelihood function.

To estimate $\mathcal{L}(x; \theta)$, we use a spectral series approach as before, but because the likelihood is a function of both $x$ and $\theta$, we consider the tensor product of a basis for $x$ and a basis for $\theta$, $\{\Psi_{i,j}\}_{i,j \in \mathbb{N}}$, where

$$\Psi_{i,j}(x, \theta) = \psi_j(x) \phi_i(\theta), \ i, j \in \mathbb{N}.$$
where \( \tilde{\theta} \) joint distribution of \( (\theta_1, \ldots, \theta_B) \). By choosing an appropriate kernel, the Remark.

evaluating function of a given estimator, we use the loss function

\[
L \left( \hat{\mathcal{L}}, \mathcal{L} \right) = \int \left( \hat{\mathcal{L}}(\mathbf{x}; \theta) - \mathcal{L}(\mathbf{x}; \theta) \right)^2 dG(\mathbf{x})dF(\theta) = \\
= \int \hat{\mathcal{L}}(\mathbf{x}; \theta)^2 dG(\mathbf{x})dF(\theta) - \\
- 2 \int \hat{\mathcal{L}}(\mathbf{x}; \theta)dF(\theta, \mathbf{x}) + K, \tag{7}
\]

where \( K \) does not depend on \( \hat{\mathcal{L}} \). We can estimate this quantity (up to \( K \)) by\[
\hat{\mathcal{L}} \left( \hat{\mathcal{L}}, \mathcal{L} \right) = \frac{1}{B} \sum_{b=1}^{B} \left[ \frac{1}{n} \sum_{k=1}^{n} \left( \hat{\mathcal{L}} \left( \tilde{x}^{(l)}_k \theta^{(l)}_k \right) \right)^2 \right] \\
- 2 \frac{1}{n} \sum_{k=1}^{n} \hat{\mathcal{L}}(\tilde{x}^{(l)}_k \theta^{(l)}_k),
\]

where \( \tilde{x}^{(l)}_1, \ldots, \tilde{x}^{(l)}_B \) is a validation sample from \( G; (\tilde{\theta}_1, \tilde{x}^{(1)}_1), \ldots, (\tilde{\theta}_B, \tilde{x}^{(B)}_B) \) is a validation sample from the joint distribution of \( \mathbf{x} \) and \( \theta \); \( \tilde{\theta}^{(l)}_1, \ldots, \tilde{\theta}^{(l)}_B \) for \( l = 1, \ldots, B \) are random permutations of the original sample \( \tilde{\theta}_1, \ldots, \tilde{\theta}_B \); and \( B \) is a number limited only by computational considerations. We choose tuning parameters so as to minimize \( \hat{\mathcal{L}} \).

Remarks. By choosing an appropriate kernel, the spectral series approach can be extended to discrete data. For example, in Lee et al. (2010), p. 185, the authors suggest a distance kernel that takes into account the discrete nature of genetic SNP data. By minimizing the loss in Eq. (7), one can also select the best kernel from a set of reasonable candidate kernels. Finally, our procedure can be scaled up to large sets of simulated data by speeding up the eigendecomposition of the Gram matrix. Possible methods include the Nyström extension (Drineas and Mahoney, 2005) and procedures described in, e.g., Belabbas and Wolfe (2009) and Halko et al. (2011). In particular, some of these approaches can be parallelized for even higher computational efficiency.

3.1 Numerical Experiments

Estimation of a likelihood function is of particular value in cases where the complexity of the data and the data-generating process prevents construction of a sufficiently accurate analytical form, a situation typically present in high-dimensional scientific data. The general setup is as follows. We have data which are modeled as an \( i.i.d. \) sample \( \mathbf{x}_1, \ldots, \mathbf{x}_m \) from the distribution \( f(\mathbf{x} \mid \theta) \). Our goal is to infer the value \( \theta \). Although we are able to simulate from \( f(\mathbf{x} \mid \theta) \) for fixed \( \theta \), we lack an analytical form for the likelihood function. Hence, we use the methodology of Section 3 to estimate \( \mathcal{L}(\mathbf{x}; \theta) \) from a simulated sample. Once we have an estimate \( \hat{\mathcal{L}}(\mathbf{x}; \theta) \), we can approximate the likelihood of an observed sample according to

\[
\hat{\mathcal{L}} \left( \mathbf{x}_1, \ldots, \mathbf{x}_m ; \theta \right) = \prod_{k=1}^{m} \hat{\mathcal{L}}(\mathbf{x}_k ; \theta).
\]

This approximation can then be used in likelihood-based inference by, for example, plugging the expression into Bayes Theorem or by finding the maximum likelihood estimate.

In what follows we present five numerical examples where the ambient dimensionality of \( \mathbf{x} \) is larger than its intrinsic dimensionality. In all experiments, we choose a uniform prior distribution on the parameter space.

Spiral. The data are \( i.i.d. \) observations of \( (X^{(1)}, X^{(2)}) \), where

\[
X^{(1)} = \theta \cos \theta + N(0, 1) \quad \text{and} \quad X^{(2)} = \theta \sin \theta + N(0, 1)
\]

for \( 0 < \theta < 15 \). Although the dimension of the sample space is 2, the data lie close to a one-dimensional spiral.

Klein Bottle. The data are \( i.i.d. \) observations of \( (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}) \), where

\[
\begin{cases}
X^{(1)} = 2(\cos \theta_2 + 1) \cos \theta_1 + N(0, 1) \\
X^{(2)} = 2(\cos \theta_2 + 1) \sin \theta_1 + N(0, 1) \\
X^{(3)} = 2 \sin \theta_2 \cos \theta_1/2 + N(0, 1) \\
X^{(4)} = 2 \sin \theta_2 \sin \theta_1/2 + N(0, 1)
\end{cases}
\]

for \( 0 < \theta_1, \theta_2 < 2\pi \). The dimension of the sample is 4, but the data lie close to a two-dimensional Klein Bottle embedded in \( \mathbb{R}^3 \).

Transformed Images. In this example, we rotate and translate an image of a tiger, see the top row of Figure 3. The model parameters are \( (\theta, \rho_x, \rho_y) \). The transformed images are centered at \( (\rho_x + N_T(0, 10), \rho_y + N_T(0, 10)^3 \) with rotation angle \( \theta + N(0, 10)) \). The final images are cropped to \( 20 \times 20 \) pixels, i.e., the sample space has dimension 400.

Edges. Here we generate \( 20 \times 20 \) images of binary edges from a model with two parameters, \( \alpha \) and \( \lambda \). The data are \( i.i.d. \) observations of an edge with rotation

\[
N_T \text{ is the truncated normal to guarantee that the parameters are in the range of the image.}
\]
angle $\alpha \sim N(0, \pi/4)$ and displacement $\lambda \sim N_T(0, 0.5)$ from the center, see Figure 3 for some examples.

Simulated Galaxy Images. The last example is a simplified version of a key estimation problem in astronomy, namely that of shear estimation (Bridle et al., 2009). We use the GalSim Toolkit\(^4\) to simulate realistic galaxy images. We sample two parameters: First, the orientation with respect to the $x$-axis of the image and, second, the axis ratio of the galaxies, which measures their ellipticity. To mimic a realistic situation, the observed data are low-resolution images of size $20 \times 20$. Figure 3, bottom, shows some examples. These images have been degraded by observational effects such as background noise, pixelization, and blurring due to the atmosphere and telescope; see the Supplementary Materials for details.

We assume that the orientation and axis ratio of galaxy $i$ are given by $a_i \sim \text{Laplace}(\alpha, 10)$ and $r_i \sim N_T(\rho, 0.1^2)$, respectively. We seek to infer $\theta = (\alpha, \rho)$ based on an observed i.i.d. sample of images $x$ contaminated by observational effects. Notice we do not observe $a_i$ and $r_i$, but only $x_i$, the 400-dimensional noisy image.

Methods. In the examples above, the likelihood function is estimated based on $n_F = n_G = 5,000$ observations from the simulation model; 60% of the data are used for training and 40% for validation. We compare Series, our spectral series estimator from Section 3, with two state-of-the-art estimators of $f(x|\theta)$. The first estimator is $KDE$ – a kernel density estimator based on taking the ratio of kernel estimates of $f(x, \theta)$ and $f(\theta)$. We use the implementation from the package “np” (Hayfield and Racine, 2008) for $R$. For the Spiral and Klein bottle examples, we select the bandwidths of KDE via cross-validation. However, the high dimension of the other examples (Transformed, Edges, and Galaxy) makes a cross-validation approach computationally intractable and the density estimates numerically unstable. For these examples, we instead use the default reference rule for the bandwidth, and we reduce the dimensionality of the data with PCA with number of components chosen by minimizing the estimated loss (Eq. 7). The second estimator in our comparisons is $LS$ – the direct least-squares conditional density estimator of Sugiyama et al. (2010b). This estimator is based on a direct expansion of the likelihood into a set of prespecified functions. This approach typically yields better results than estimators based on the ratio of random variables. Again, to avoid the problem of high dimensionality in the examples with $d > 4$, we implement $PCA+LS$, the direct least-squares conditional density estimator after dimension reduction via PCA, with number of components chosen so as to minimize the estimated loss. PCA has the additional goal of decorrelating adjacent pixels in the images examples.

Results. In Tables 1 and 2, we present the estimated $L^2$ loss (Eq. 7), as well as the estimated average likelihood $E_{(X, \theta)} L(X; \theta)$ based on a test set with $3,000$ observations\(^5\). Both measures indicate that, while traditional methods have better performance in low dimensions, our spectral series method yields substantial improvements when the ambient dimensionality of the sample space is large. Note that even after dimension reduction, $LS$ does not yield the same performance as Series. In fact, in some cases, a dimension reduction via PCA leads to less accurate estimates.

As a further illustration, Figure 4 shows the estimated likelihood function for a sample of size $m = 10$ drawn from the galaxy image model with parameters $\alpha = 80^\circ$ and $\rho = 0.2$. For comparison, we also include the true likelihood function (TRUTH), which is unavailable in practical applications\(^6\). It is apparent from the figure that the spectral series estimator comes closer to the truth than the other estimators, even without first reducing the dimensionality of the galaxy images. In the Supplementary Materials we present additional plots for other sample sizes. These results yield similar conclusions.

Furthermore, to quantify how the level sets of the like-\(^4\)https://github.com/GalSim-developers/GalSim

\(^5\)To make results comparable, we renormalize the estimated likelihood functions to integrate to $1$ in $\theta$.

\(^6\)Because the observed images are simulated, we can compute $a_i$ and $r_i$.
Table 1: Estimated $L^2$ loss (with standard errors) of the likelihood function estimators. Best-performing models with smallest loss are in bold fonts.

<table>
<thead>
<tr>
<th>DATA</th>
<th>DIM.</th>
<th>L$^2$ LOSS</th>
<th>SERIES</th>
<th>LS</th>
<th>PCA+LS</th>
<th>KDE</th>
<th>PCA+KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spiral</td>
<td>2</td>
<td>7.13 (0.14)</td>
<td>6.61 (0.12)</td>
<td>—</td>
<td>2.95 (0.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klein Bottle</td>
<td>4</td>
<td>1.45 (0.07)</td>
<td>2.02 (0.06)</td>
<td>—</td>
<td>1.60 (0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transf. Images</td>
<td>400</td>
<td>20.94 (0.03)</td>
<td>26.91 (0.04)</td>
<td>27.12 (0.03)</td>
<td>26.62 (0.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edges</td>
<td>400</td>
<td>0.70 (0.03)</td>
<td>1.77 (0.02)</td>
<td>1.55 (0.03)</td>
<td>—</td>
<td>1.60 (0.02)</td>
<td></td>
</tr>
<tr>
<td>Galaxy Images</td>
<td>400</td>
<td>40.94 (0.03)</td>
<td>42.57 (0.01)</td>
<td>42.53 (0.01)</td>
<td>—</td>
<td>43.99 (0.04)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Estimated average likelihood (with standard errors) of the likelihood function estimators. Best-performing models with largest average likelihood are in bold fonts.

<table>
<thead>
<tr>
<th>DATA</th>
<th>DIM.</th>
<th>AVERAGE LIKELIHOOD</th>
<th>SERIES</th>
<th>LS</th>
<th>PCA+LS</th>
<th>KDE</th>
<th>PCA+KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spiral</td>
<td>2</td>
<td>16.54 (0.16)</td>
<td>19.49 (0.14)</td>
<td>—</td>
<td>28.62 (0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klein Bottle</td>
<td>4</td>
<td>5.62 (0.08)</td>
<td>4.96 (0.08)</td>
<td>—</td>
<td>5.63 (0.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transf. Images</td>
<td>400</td>
<td>8.31 (0.03)</td>
<td>1.83 (0.03)</td>
<td>1.08 (0.02)</td>
<td>—</td>
<td>1.58 (0.06)</td>
<td></td>
</tr>
<tr>
<td>Edges</td>
<td>400</td>
<td>3.69 (0.04)</td>
<td>1.72 (0.02)</td>
<td>2.55 (0.03)</td>
<td>—</td>
<td>2.10 (0.02)</td>
<td></td>
</tr>
<tr>
<td>Galaxy Images</td>
<td>400</td>
<td>4.63 (0.04)</td>
<td>2.24 (0.01)</td>
<td>2.43 (0.02)</td>
<td>—</td>
<td>1.01 (0.04)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Estimated likelihood function for the galaxy example using different estimators. The contours represent the level sets; horizontal and vertical lines are the true values of the parameters. The spectral series estimator gets close to the true distribution, which is uncomputable in practice.

As the prior distribution is uniform, this quantity is $\mathbb{E}_{\theta^*} \left[ \int d(\theta^*, \theta) \hat{L}(x; \theta) d\theta \right]$, where the expectation is taken with respect to both $\theta^*$ and the observed data. Here we choose $d(\theta^*, \theta)$ to be the Euclidean distance between the vectors of parameters, standardized so that each component has minimum 0 and maximum 1. As a final comparison of methods, we study how the above likelihood metric changes as a function of the sample size $m$ of the observed data (the sample size of the simulated data used to estimate the likelihood is held constant); see Figure 5 for results. Because $\hat{L}(x; \theta)$ concentrates around the true parameter value $\theta^*$ for large sample sizes, we expect the average likelihood to decrease as $m$ increases — if the likelihood estimates are reasonable. Indeed, we observe this behavior for all methods in the comparison for the problems with low dimensionality. However, for the problems with high dimensionality, this is no longer the case. On the other hand, the results indicate that Series is able to overcome the curse of dimensionality and recover the true $\theta^*$ parameter as the number of observations increases.

4 THEORY

Next we provide theoretical guarantees of the performance of the estimator $\hat{\beta}$. The integral operator $K_x$ from Eq. 1 is self-adjoint, compact and has a countable number of eigenfunctions $\{\psi_i\}_i$ with respective eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J > 0$. These eigenfunctions therefore form an orthonormal basis of $L^2(\mathcal{X}, \mu)$ (Minh, 2010). We make the following assumptions:

**Assumption 1.** $\int \beta^2(x) d\mu(x) < \infty$.

**Assumption 2.** $\lambda_1 > \lambda_2 > \ldots > \lambda_J > 0$.

Assumption 1 states that the ratio is $L^2$ integrable. It implies that it is possible to expand $\beta$ into the basis...
ψ. Assumption 2 allows one to uniquely define each of the eigenfunctions (see, e.g., Ji et al. 2012 for similar assumptions, and Zwald and Blanchard 2005 on how to proceed if it does not hold). Let \( \mathcal{H}_{K_x} \) denote the Reproducing Kernel Hilbert Space (RKHS) associated to the kernel \( K_x \). We assume

**Assumption 3.** \( c_{K_x} \equiv \| \beta(x) \|_{\mathcal{H}_{K_x}}^2 < \infty. \)

Assumption 3 implies smoothness of \( \beta(x) \) as measured by the RKHS norm defined by \( K_x \). Smaller values of \( c_{K_x} \) imply smoother functions. In the Supplementary Materials we prove the following main result.

**Theorem 1.** Under Assumptions 1 – 3, the loss \( \int (\hat{\beta}_J(x) - \beta(x))^2 \, dG(x) \) is bounded by

\[
J \times \left[ O_P\left( \frac{1}{n_F} \right) + O_P\left( \frac{1}{\lambda_J^2 n_G} \right) \right] + c_{K_x} O(\lambda_J),
\]

where \( \Delta_J = \min_{1 \leq j \leq J} |\lambda_j - \lambda_{J+1}| \) and \( \hat{\beta}_J(x) \) is the spectral series ratio estimator truncated at \( J \).

The first term of the rate in Theorem 1 is the sample error. The second term is the approximation error. Smooth functions have a small value of \( c_{K_x} \), and therefore a smaller bias. These rates depend on the decay of the eigenvalues \( \lambda_j \) and the eigengaps \( \lambda_j - \lambda_{J+1} \).

As an illustration, assume a fixed kernel \( K_x \). Then, if \( n \equiv n_F = n_G, \lambda_J \asymp J^{-2\alpha} \) for some \( \alpha > \frac{1}{2} \), and \( \lambda_j - \lambda_{J+1} \asymp J^{-2\alpha-1} \) (see Ji et al. 2012 for an empirical motivation), then the optimal smoothing is given by \( J \asymp n^{1/(8\alpha+3)} \). With this choice of \( J \), the rate of convergence is

\[
O_P\left( n^{-2\alpha/(8\alpha+3)} \right).
\]

Note, however, that by changing the kernel (e.g., by using different bandwidths \( \epsilon \)), one can improve the performance of the estimator: different kernels lead to different eigenvalue decays, as well as different \( c_{K_x} \)'s (e.g., \( \alpha \) depends on \( \epsilon \)). Hence, choosing the tuning parameters properly is important. Note that \( J \times O_P\left( 1/n_F \right) \) is the traditional variance of orthogonal classical estimators in one dimension, in which one expands the target function with respect to a basis fixed beforehand (e.g., the Fourier basis) (Efronovich, 1999). The additional term \( J \times O_P\left( 1/(\lambda_J^2 n_G) \right) \) is the cost of estimating a basis (from data) that better captures the geometry of the data.

Under similar assumptions (see the Supplementary Materials for proofs and details), an analogous bound holds for the spectral series likelihood estimator truncated at \( I \) and \( J \):

\[
L\left( \hat{\theta}_{I,J}, \theta \right) = I J O_P\left( \max \left\{ \frac{1}{\lambda_J^2 \Delta_{J,J+1}^2 n_G}, \frac{1}{\lambda_J^2 \Delta_{J,J}^2 n_F} \right\} \right) + c_{K_x} O \left( \lambda_J^0 + c_{K_x} O \left( \lambda_J^2 \right) \right),
\]

where the superscript \( x \) and \( \theta \) denote quantities associated with the eigenfunctions \( \psi_j \)'s and \( \phi_j \)'s, respectively. Similar interpretation holds for this bound.

5 CONCLUSION

We have demonstrated the effectiveness of a new spectral series approach for estimating the ratio of two high-dimensional densities, with extensions to likelihood approximation in high dimensions. Traditional approaches typically fail for high-dimensional data (even with a dimension reduction by PCA) whereas the proposed method is designed to adapt to the underlying geometry of the data.

6 Acknowledgments

This work was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant 200959/2010-7) and the Estella Loomis McCandless Professorship.
References


