Supplementary Material for "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

Section 1 provides details on the photometric data set used to illustrate the spectral series density ratio estimator. Section 2 provides additional details on the galaxy data set used to illustrate the spectral series likelihood estimator. This section also includes additional graphics with the estimated likelihood functions, omitted from the main document for the sake of space. In Section 3 we prove the bounds in the paper for the Spectral Series Density Ratio estimator. Section 4 shows analogous bounds to the Spectral Series Likelihood estimator, and presents an outline of the proofs.

1 Details on Photometric Redshift Problem and Sloan Digital Sky Survey Data

In spectroscopy, the flux of a galaxy, i.e., the number of photons emitted per unit area per unit time, is measured as a function of wavelength. By using these measurements it is possible to determine the redshift of a galaxy with great precision. On the other hand, in photometry–an extremely low-resolution spectroscopy– the photons are collected into a few (≈ 5) wavelength bins (also named bands). In each of these bins, the magnitudes - which are logarithmic measurements of photon flux - are measured. Typical instruments measure in five bands, denoted by u, g, r, i, and z. The differences between contiguous magnitudes (also named *colors*; e.g., $g-r$) are useful predictors for the redshift of the galaxy. Multiple estimators of the magnitudes exist, here we work with two of them: `model` and `cmodel` (Sheldon et al., 2012). Our covariates are the 4 colors in each magnitude system, plus the raw value of r-band magnitude in both system. Hence there are $4 \times 2 = 10$ covariates $\mathbf{x}$.

Because it is difficult to acquire the spectroscopic redshift of faint galaxies, these data suffer from selection bias. We take this into account by making the *covariate shift* assumption: although $f_L(\mathbf{x})$ can be different from $f_U(\mathbf{x})$, we assume the conditional distribution $f(z|\mathbf{x})$ is the same in both populations (Shimodaira, 2000). This correction will reweight labeled data to account for the difference in their distribution as compared to the unlabeled data (Gretton et al., 2010).

The data we use are similar to that used by Sheldon et al. (2012). The labeled data set contains spectroscopic information about 435,875 galaxies from the Sloan Digital Sky Survey (SDSS). It also contains `model` and `cmodel` magnitudes in the bands u, g, r, i and z. These samples were chosen by applying cuts to both main sample galaxies and luminous red galaxies of SDSS. Only galaxies in which the redshift information has confidence level at least 0.9 were selected. It was also required that the galaxies were not too faint. The unlabeled data set contains a subset of 538,974 galaxies of SDSS data. The only variables that are observed are the photometric magnitudes. These samples were chosen by applying cuts to the original unlabeled SDSS data imposing that the galaxies are not too faint and have reasonable colors, see Sheldon et al. (2012) for more details.

2 Additional Figures for Galaxy Likelihood Estimation Example

The first column of Figure 1 shows examples of galaxies with different parameter values, generated by GalSim toolkit. To make the situation more realistic, we assume we cannot observe the images of the uncontaminated galaxies in the first row, but instead only the 20 × 20 images from the last row. These are low-resolution images degraded by observational effects, background noise, and pixelization; see Bridle et al. (2009) for more details.

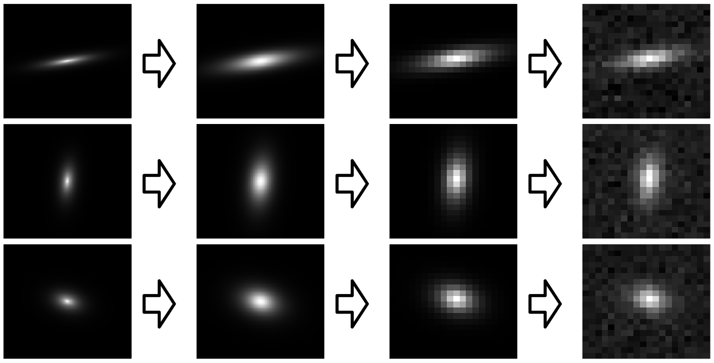

Figure 1: Examples of galaxies with different orientations and axis ratios. From left to right: High-resolution, uncontaminated galaxy image; effect of PSF caused by atmosphere and telescope; pixelated image; and observed image containing additional Poisson noise. We only observe images on the right.

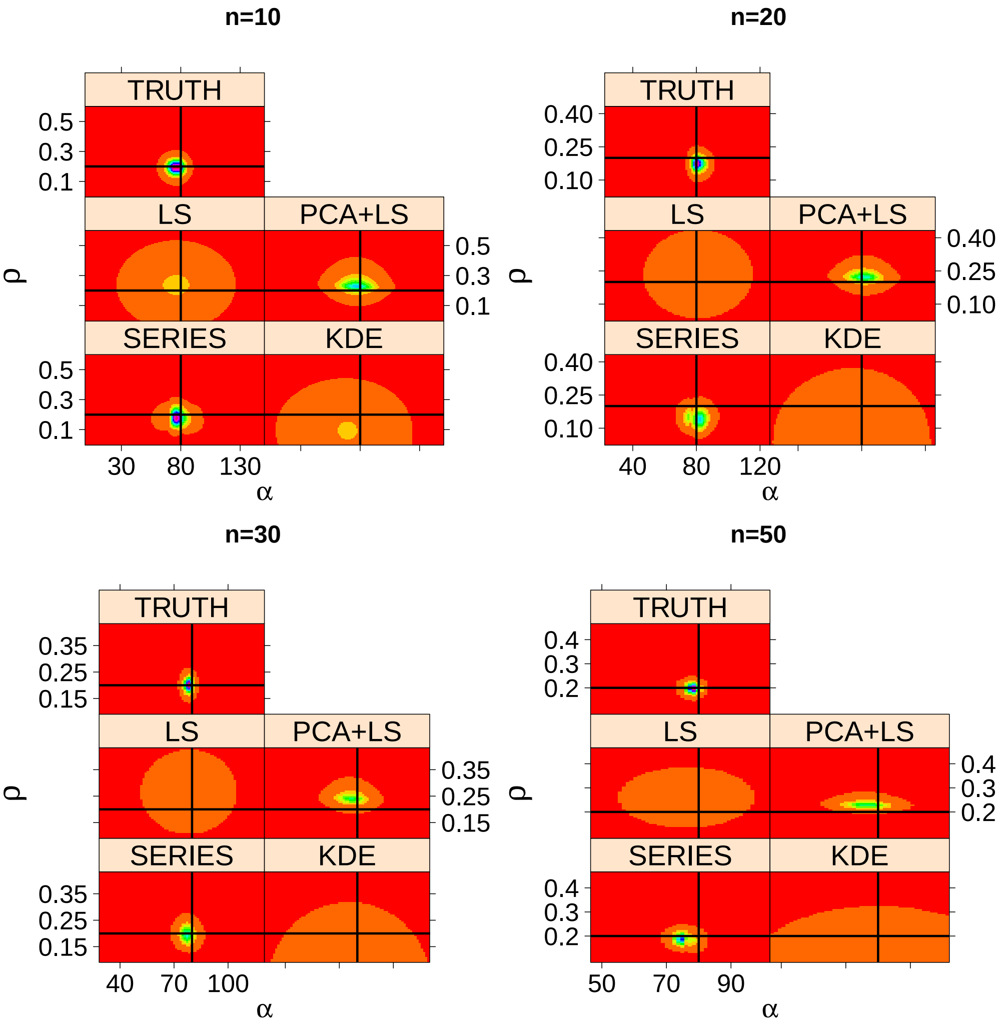

Figure 2: Comparison of level sets of estimated likelihood function $\mathcal{L}(\mathbf{x}; (\alpha, \rho))$ for the galaxy example for 4 samples sizes. Horizontal and vertical lines are the true values of the parameters. In all cases, the spectral series estimator gets closer to the real distribution, which is uncomputable in practice.

3 Spectral Series Density Ratio Estimator

Here we prove the bounds from the paper. We recall the assumption we make:

Assumption 1. $\int \beta^2(\mathbf{x}) dG(\mathbf{x}) < \infty$.

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

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

Define the following quantities:

$$\beta_J(\mathbf{x}) = \sum_{j=1}^{J} \beta_j \psi_j(\mathbf{x}), \quad \beta_j = \int \psi_j(\mathbf{x}) dF(\mathbf{x})$$
$$\widehat{\beta}_J(\mathbf{x}) = \sum_{j=1}^{J} \widehat{\beta}_i \widehat{\psi}_j(\mathbf{x}), \quad \widehat{\beta}_j = \frac{1}{n_F} \sum_{k=1}^{n} \widehat{\psi}_j(\mathbf{x}_k^F)$$

and note that

$$\begin{aligned}
\int \left(\widehat{\beta}_J(\mathbf{x}) - \beta(\mathbf{x})\right)^2 dG(\mathbf{x}) &\leq \int \left(\widehat{\beta}_J(\mathbf{x}) - \beta_J(\mathbf{x}) + \beta_J(\mathbf{x}) - \beta(\mathbf{x})\right)^2 dG(\mathbf{x}) \\
&\leq 2\left(\text{VAR}(\widehat{\beta}_J(\mathbf{x}), \beta_J(\mathbf{x})) + B(\beta_J(\mathbf{x}), \beta(\mathbf{x}))\right).
\end{aligned}$$

where

$$B(\beta(\mathbf{x})_J, \beta(\mathbf{x})) \equiv \iint \left(\beta_J(\mathbf{x}) - \beta(\mathbf{x})\right)^2 dG(\mathbf{x})$$

can be interpreted as a bias term (or approximation error) and

$$\text{VAR}(\widehat{\beta}_J(\mathbf{x}), \beta_J(\mathbf{x})) \equiv \iint \left(\widehat{\beta}_J(\mathbf{x}) - \beta_J(\mathbf{x})\right)^2 dG(\mathbf{x}) dz$$

can be interpreted as a variance term. First we bound the variance.

Lemma 1. *There exists $C > 0$ such that $|\beta(\mathbf{x})| < C$ for all $\mathbf{x} \in \mathcal{X}$.*

Proof. Using Assumption 3 and the fact the kernel is bounded, it follows from the reproducing property and Cauchy-Schwartz inequality that

$$\beta(\mathbf{x}) = \langle \beta(.), K(\mathbf{x}, .) \rangle_{\mathcal{H}_{K_\mathbf{x}}} \leq ||\beta(.)||_{\mathcal{H}_{K_\mathbf{x}}} \sqrt{K_\mathbf{x}(\mathbf{x}, \mathbf{x})} < C$$

for some $C > 0$. □

Lemma 2. *For all $1 \leq j \leq J$,*

$$\int \left(\widehat{\psi_j}(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x}) = O_P\left(\frac{1}{\lambda_j \delta_j^2 n_G}\right),$$

where $\delta_j = \lambda_j - \lambda_{j+1}$.

For a proof of Lemma 2 see, e.g., Sinha and Belkin (2009).

Lemma 3. *For all $1 \leq j \leq J$,*

$$\int \left(\widehat{\psi_j}(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dF(\mathbf{x}) = O_P\left(\frac{1}{\lambda_j \delta_j^2 n_G}\right),$$

where $\delta_j = \lambda_j - \lambda_{j+1}$.

Proof. It follows from Lemmas 1 and 2 that

$$\int \left(\widehat{\psi_j}(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dF(\mathbf{x}) = \int \left(\widehat{\psi_j}(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 \beta(\mathbf{x}) dG(\mathbf{x}) \leq$$
$$C \int \left(\widehat{\psi_j}(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x}) = O_P\left(\frac{1}{\lambda_j \delta_j^2 n_G}\right)$$

□

Lemma 4. *For all* $1 \leq j \leq J$*, there exists* $C < \infty$ *that does not depend on* n_G *such that*

$$E\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right)^2\right] < C.$$

Proof. Let $\delta \in (0,1)$. From Sinha and Belkin (2009), it follows that

$$\mathbb{P}\left(\int \left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x}) > \frac{16\log\left(\frac{2}{\delta}\right)}{\delta_j^2 n_G}\right) < \delta,$$

and therefore for all $\epsilon > 0$,

$$\mathbb{P}\left(\int \left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x}) > \epsilon\right) < 2e^{-\frac{\delta_j^2 n_G \epsilon}{16}}.$$

Hence , using Lemma 1,

$$\mathbb{E}\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right)^2\right] = \mathbb{E}\left[\int \left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dF(\mathbf{x})\right] \leq C\mathbb{E}\left[\int \left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x})\right]$$
$$\int_0^\infty \mathbb{P}\left(\int \left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dG(\mathbf{x}) > \epsilon\right) d\epsilon \leq \int 2e^{-\frac{\delta_j^2 n_G \epsilon}{16}} d\epsilon < \int 2e^{-\frac{\delta_j^2 \epsilon}{16}} d\epsilon < \infty$$

□

Lemma 5. *For all* $1 \leq j \leq J$*, there exists* $C < \infty$ *that does not depend on* m *such that*

$$\mathbb{E}\left[\mathbb{V}\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right) \middle| \mathbf{X}_1^G, \ldots, \mathbf{X}_{n_G}^G\right]\right] < C$$

Proof. We have that

$$\mathbb{E}\left[\mathbb{V}\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right) \middle| \mathbf{X}_1^G, \ldots, \mathbf{X}_{n_G}^G\right]\right]$$
$$\leq \mathbb{V}\left[\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right] \leq \mathbb{E}\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right)^2\right]$$

The result follows from Lemma 4. □

Lemma 6. *For all* $1 \leq j \leq J$,

$$\left[\frac{1}{n}\sum_{k=1}^{n}\left(\widehat{\psi}_j(\mathbf{X}_k^F) - \psi_j(\mathbf{X}_k^F)\right) - \int\left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)dF(\mathbf{x})\right]^2 = O_P\left(\frac{1}{n}\right)$$

Proof. By Chebyshev's inequality it holds that for all $M > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\left(\widehat{\psi}_j(\mathbf{X}_k^F) - \psi_j(\mathbf{X}_k^F)\right) - \int\left(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)dF(\mathbf{x})\right|^2 > M\middle|\mathbf{X}_1^G, \ldots, \mathbf{X}_{n_G}^G\right) \leq$$
$$\frac{1}{n_F M}\mathbb{V}\left[\left(\widehat{\psi}_j(\mathbf{X}^F) - \psi_j(\mathbf{X}^F)\right)\middle|\mathbf{X}_1^G, \ldots, \mathbf{X}_{n_G}^G\right].$$

The conclusion follows from taking an expectations with respect to sample from G on both sides of the equation and using Lemma 5.

□

Note that $\widehat{\psi}'$s are random functions, and therefore the proof of Lemma 6 relies on the fact that these functions are estimated using a different sample than $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Lemma 7. *For all* $1 \leq j \leq J$,

$$\left(\widehat{\beta}_j - \beta_j\right)^2 = O_P\left(\frac{1}{n}\right) + O_P\left(\frac{1}{\lambda_j \delta_j^2 n_G}\right).$$

Proof. It holds that

$$\frac{1}{2}\left(\widehat{\beta}_j - \beta_j\right)^2 \leq \left(\frac{1}{n_F}\sum_{k=1}^{n_F}\psi_j(\mathbf{X}_k^F) - \beta_j\right)^2 + \left(\frac{1}{n_F}\sum_{k=1}^{n_F}(\widehat{\psi}_j(\mathbf{X}_k^F) - \psi_j(\mathbf{X}_k^F))\right)^2.$$

The first term is $O_P\left(\frac{1}{n_F}\right)$. The second term divided by two is bounded by

$$\frac{1}{2}\left(\frac{1}{n_F}\sum_{k=1}^{n_F}(\widehat{\psi}_j(\mathbf{X}_k^F) - \psi_j(\mathbf{X}_k^F)) - \int(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))dF(\mathbf{x}) + \int(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))dF(\mathbf{x})\right)^2$$
$$\leq \left(\frac{1}{n_F}\sum_{k=1}^{n_F}(\widehat{\psi}_j(\mathbf{X}_k^F) - \psi_j(\mathbf{X}_k^F)) - \int(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))dF(\mathbf{x})\right)^2 + \int(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))^2 dF(\mathbf{x})$$
$$= O_P\left(\frac{1}{n_F}\right) + O_P\left(\frac{1}{\lambda_j \delta_j^2 n_G}\right)$$

The result follows from Lemma 3.

□

Lemma 8. *[Sinha and Belkin 2009,* ***Corollary 1****] Under the stated assumptions,*

$$\int \widehat{\psi}_j^2(\mathbf{x})dG(\mathbf{x}) = O_P\left(\frac{1}{\lambda_j \Delta_J^2 n_G}\right) + 1$$

and

$$\int \widehat{\psi}_i(\mathbf{x})\widehat{\psi}_j(\mathbf{x})dG(\mathbf{x}) = O_P\left(\left(\frac{1}{\sqrt{\lambda_i}} + \frac{1}{\sqrt{\lambda_j}}\right)\frac{1}{\Delta_J\sqrt{n_G}}\right)$$

where $\Delta_J = \min_{1\leq j\leq J}\delta_j$*.*

Lemma 9. *Let* $h(\mathbf{x}) = \sum_{j=1}^J \beta_j \widehat{\psi}_j(\mathbf{x})$*. Then*

$$\int \left|\widehat{\beta}_J(\mathbf{x}) - h(\mathbf{x})\right|^2 dG(\mathbf{x}) = J\left(O_P\left(\frac{1}{n_F}\right) + O_P\left(\frac{1}{\lambda_J \Delta_J^2 n_G}\right)\right).$$

Proof.

$$\int \left|\widehat{\beta}_J(z|\mathbf{x}) - h(\mathbf{x})\right|^2 dG(\mathbf{x})$$

$$= \sum_{j=1}^J \left(\widehat{\beta}_j - \beta_j\right)^2 \int \widehat{\psi}_j^2(\mathbf{x})dG(\mathbf{x}) + \sum_{j=1}^J \sum_{l=1,l\neq j}^J \left(\widehat{\beta}_j - \beta_j\right)\left(\widehat{\beta}_l - \beta_l\right)\int \widehat{\psi}_j(\mathbf{x})\widehat{\psi}_l(\mathbf{x})dG(\mathbf{x})$$

$$\sum_{j=1}^J \left(\widehat{\beta}_j - \beta_j\right)^2 \int \widehat{\psi}_j^2(\mathbf{x})dG(\mathbf{x}) + \left[\sum_{j=1}^J \left(\widehat{\beta}_j - \beta_j\right)^2\right]\left[\sqrt{\sum_{j=1}^J \sum_{l=1,l\neq j}^J \left(\int \widehat{\psi}_j(\mathbf{x})\widehat{\psi}_l(\mathbf{x})dG(\mathbf{x})\right)^2}\right]$$

where the last inequality follows from using Cauchy-Schwartz repeatedly. The result follows from Lemmas 7 and 8.

□

Lemma 10. *Let* $h(\mathbf{x})$ *be as in Lemma 9. Then*

$$\int |h(\mathbf{x}) - \beta_J(\mathbf{x})|^2 dG(\mathbf{x}) = JO_P\left(\frac{1}{\lambda_J \Delta_J^2 n_G}\right).$$

Proof. Using Cauchy-Schwartz inequality,

$$\int |h(\mathbf{x}) - \beta_J(\mathbf{x})|^2 dG(\mathbf{x}) \leq \int \left|\sum_{j=1}^J \beta_j\left(\psi_j(\mathbf{x}) - \widehat{\psi_j}(\mathbf{x})\right)\right|^2 dG(\mathbf{x})$$

$$= \left\{\sum_{j=1}^J \beta_j^2\right\}\left\{\sum_{j=1}^J \int \left[\psi_j(\mathbf{x}) - \widehat{\psi_j}(\mathbf{x})\right]^2 dG(\mathbf{x})\right\}.$$

The conclusion follows from Lemma 2 and by noticing that $\sum_{j=1}^{J} \beta_j^2 \leq ||\beta(\mathbf{x})||^2 < \infty$.

□

It is now possible to bound the variance term:

Theorem 1. *Under the stated assumptions,*

$$VAR(\widehat{\beta}_J(\mathbf{x}), \beta_J(\mathbf{x})) = J\left(O_P\left(\frac{1}{n_F}\right) + O_P\left(\frac{1}{\lambda_J \Delta_J^2 n_G}\right)\right).$$

Proof. Let h be defined as in Lemma 9. We have

$$\frac{1}{2}\text{VAR}(\widehat{\beta}_J(\mathbf{x}), \beta_J(\mathbf{x})) = \frac{1}{2}\int \left|\widehat{\beta}_J(\mathbf{x}) - h(\mathbf{x}) + h(\mathbf{x}) - \beta_J(\mathbf{x})\right|^2 dG(\mathbf{x})$$
$$\leq \int \left|\widehat{\beta}_J(\mathbf{x}) - h(\mathbf{x})\right|^2 dG(\mathbf{x}) + \int |h(\mathbf{x}) - \beta_J(\mathbf{x})|^2 \, dG(\mathbf{x}).$$

The conclusion follows from Lemmas 9 and 10.

□

We now bound the bias term.

Lemma 11. $\sum_{j\geq J} \beta_j^2 = c_{K_\mathbf{x}} O(\lambda_J)$.

Proof. Note that $c_{K_\mathbf{x}} = ||\beta(\mathbf{x})||_{\mathcal{H}}^2 = \sum_{j\geq 1} \frac{\beta_j^2}{\lambda_j}$ (Minh, 2010). Using Assumption 3 and that the eigenvalues are decreasing it follows that

$$\sum_{j\geq J} \beta_j^2 = \sum_{j\geq J} \beta_j^2 \frac{\lambda_j}{\lambda_j} \leq \lambda_J ||\beta(\mathbf{x})||_{\mathcal{H}}^2,$$

and therefore $\sum_{j\geq J} \beta_j^2 \leq \lambda_J c_{K_\mathbf{x}} = c_{K_\mathbf{x}} O(\lambda_J)$.

□

Theorem 2. *Under the stated assumptions, the bias is bounded by*

$$B(\beta_J(\mathbf{x}), \beta(\mathbf{x})) = c_{K_\mathbf{x}} O(\lambda_J).$$

Proof. By using orthogonality, we have that

$$B(\beta_J(\mathbf{x}), \beta(\mathbf{x})) \overset{\text{def}}{=} \int (\beta(\mathbf{x}) - \beta_J(\mathbf{x}))^2 \, dG(\mathbf{x}) = \sum_{j>J} \beta_j^2.$$

The Theorem follows from Lemma 11. □

4 Spectral Series Likelihood Estimator

We now present similar bounds to those from shown for the Density Ratio estimator. To avoid confusions with the last section, from now on we denote by $\lambda_j^{\mathbf{x}}$ the eigenvalue λ_j relative to the eigenfunction ψ_j, and by $\delta_j^{\mathbf{x}}$ its eigengap previously denoted by δ_j.

In this section, we assume Assumption 1 and 2 from the last section, and, additionally:

Assumption 4. *For all $\theta \in \Theta$, let $g_\theta : \mathcal{X} \longrightarrow \Re$; $g_\theta(\mathbf{x}) = \mathcal{L}(\mathbf{x};\theta)$. $g_\theta \in \mathcal{H}_{K_{\mathbf{x}}}(c_\theta) \equiv \{g \in \mathcal{H}_{K_{\mathbf{x}}} : ||g||^2_{\mathcal{H}_{K_{\mathbf{x}}}} \leq c_\theta^2\}$ where c_θ's are such that $c_{K_{\mathbf{x}}} \equiv \int_\Theta c_\theta^2 dF(\theta) < \infty$.*

Assumption 4 requires that for every $\theta \in \Theta$ fixed, $\mathcal{L}(\mathbf{x};\theta)$ is a smooth function of $\mathbf{x}$, where we again measure smoothness in a RKHS through its norm, and is analogous to Assumption 3. The last assumption we need is analogous to 4, and requires that for every $\mathbf{x} \in \mathcal{X}$ fixed, $\mathcal{L}(\mathbf{x};\theta)$ is a smooth function of θ:

Assumption 5. *For all $\mathbf{x} \in \mathcal{X}$, let $h_{\mathbf{x}} : \Theta \longrightarrow \Re$; $h_{\mathbf{x}}(\theta) = \mathcal{L}(\mathbf{x};\theta)$. $h_{\mathbf{x}} \in \mathcal{H}_{K_\theta}(c_{\mathbf{x}}) \equiv \{h \in \mathcal{H}_{K_\theta} : ||h||^2_{\mathcal{H}_{K_\theta}} \leq c_{\mathbf{x}}^2\}$ where $c_{\mathbf{x}}$'s are such that $c_{K_\theta} \equiv \int_{\mathcal{X}} c_{\mathbf{x}}^2 dG(\mathbf{x}) < \infty$.*

First we note that an analogous decomposition of the loss

$$L\left(\widehat{\mathcal{L}},\mathcal{L}\right) = \int \left(\widehat{\mathcal{L}}(\mathbf{x};\theta) - \mathcal{L}(\mathbf{x};\theta)\right)^2 dG(\mathbf{x})dF(\theta)$$

in terms of bias and variance holds. The proof that the bound on the variance term is analogous to the one of the variance bound of the density ratio estimator. The main difference is in Lemmas 2 and 8. We state and prove the new version of these in Lemmas 12 and 13. Notice that analogous Lemmas to 2 and 8 hold for basis ϕ_i.

Lemma 12. *For all $1 \leq i \leq I$ and for all $1 \leq j \leq J$,*

$$\iint \left(\widehat{\Psi_{i,j}}(\theta,\mathbf{x}) - \Psi_{i,j}(\theta,\mathbf{x})\right)^2 dG(\mathbf{x})dF(\theta) = O_P\left(\max\left\{\frac{1}{\lambda_j^{\mathbf{x}}\delta^2_{\mathbf{x},j}n_G}, \frac{1}{\lambda_i^{\theta}\delta^2_{\theta,i}n_F}\right\}\right)$$

,

Proof. We have that

$$\begin{aligned}
&\frac{1}{2}\iint \left(\widehat{\Psi_{i,j}}(\theta,\mathbf{x}) - \Psi_{i,j}(\theta,\mathbf{x})\right)^2 dG(\mathbf{x})dF(\theta) \leq \\
&\iint \left((\widehat{\phi}_i(\theta) - \phi_i(\theta))\widehat{\psi}_j(\mathbf{x}) + \phi_i(\theta)(\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))\right)^2 dG(\mathbf{x})dF(\theta) \leq \\
&\int \widehat{\psi}_j^2(\mathbf{x})dG(\mathbf{x}) \int (\widehat{\phi}_i(\theta) - \phi_i(\theta))^2 dF(\theta) + \int \phi_i^2(\theta)dF(\theta) \int (\widehat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x}))^2 dG(\mathbf{x})
\end{aligned}$$

The results follows from orthonormality of ϕ_i wrt to $F(\theta)$, Lemmas 2 and 8. □

Lemma 13. *Under the stated assumptions,* $\iint \widehat{\Psi}_{i,j}(\theta,\mathbf{x})\widehat{\Psi}_{k,l}(\theta,\mathbf{x})dG(\mathbf{x})dF(\theta) =$

$$= \begin{cases} 1 + O_P\left(\max\left\{\frac{1}{\lambda_j^{\mathbf{x}}\delta_{\mathbf{x},j}^2 n_G}, \frac{1}{\lambda_i^{\theta}\delta_{\theta,i}^2 n_F}\right\}\right), & \textit{if } i = k \textit{ and } j = l \\ O_P\left(\left(\frac{1}{\sqrt{\lambda_l^{\mathbf{x}}}} + \frac{1}{\sqrt{\lambda_j^{\mathbf{x}}}}\right)\frac{1}{\Delta_J^{\mathbf{x}}\sqrt{n_G}}\right) & \textit{if } i = k \textit{ and } j \neq l \\ O_P\left(\left(\frac{1}{\sqrt{\lambda_i^{\theta}}} + \frac{1}{\sqrt{\lambda_k^{\theta}}}\right)\frac{1}{\Delta_I^{\theta}\sqrt{n_F}}\right) & \textit{if } i \neq k \textit{ and } j = l \\ O_P\left(\max\left\{\left(\frac{1}{\sqrt{\lambda_l^{\mathbf{x}}}} + \frac{1}{\sqrt{\lambda_j^{\mathbf{x}}}}\right)\frac{1}{\Delta_J^{\mathbf{x}}\sqrt{n_G}}, \left(\frac{1}{\sqrt{\lambda_i^{\theta}}} + \frac{1}{\sqrt{\lambda_k^{\theta}}}\right)\frac{1}{\Delta_I^{\theta}\sqrt{n_F}}\right\}\right) & \textit{if } i \neq k \textit{ and } j \neq l \end{cases}$$

Proof. The proof of these facts follow from noticing that $\iint \widehat{\Psi}_{i,j}(\theta,\mathbf{x})\widehat{\Psi}_{k,l}(\theta,\mathbf{x})dG(\mathbf{x})dF(\theta) = \int \widehat{\psi}_j(\mathbf{x})\widehat{\psi}_l(\mathbf{x})dG(\mathbf{x}) \int \phi_i(\theta)\phi_k(\theta)dF(\theta)$ and using Lemma 8 and its analogous for the basis ϕ_i. □

The bound on the bias presents some additional differences to the proof of the bias bound from the ratio estimator, we therefore show it in details in the sequence.

Lemma 14. *For each* $\theta \in \Theta$*, expand* $g_\theta(\mathbf{x})$ *into the basis* ψ : $g_\theta(\mathbf{x}) = \sum_{j\geq 1} \alpha_j^z \psi_j(\mathbf{x})$*, where* $\alpha_j^\theta = \int g_\theta(\mathbf{x})\psi_j(\mathbf{x})dG(\mathbf{x})$*. We have*

$$\alpha_j^\theta = \sum_{i\geq 1} \beta_{i,j}\phi_i(\theta) \textit{ and } \int \left(\alpha_j^\theta\right)^2 dF(\theta) = \sum_{i\geq 1} \beta_{i,j}^2.$$

Proof. It follows from projecting α_j^θ into the basis ϕ. □

Similarly, we have the following.

Lemma 15. *For each* $\mathbf{x} \in \mathcal{X}$*, expand* $h_{\mathbf{x}}(\theta)$ *into the basis* ϕ : $h_{\mathbf{x}}(\theta) = \sum_{i\geq 1} \alpha_i^{\mathbf{x}}\phi_i(\theta)$*, where* $\alpha_i^{\mathbf{x}} = \int h_{\mathbf{x}}(\theta)\phi_i(\theta)dF(\theta)$*. We have*

$$\alpha_i^{\mathbf{x}} = \sum_{j\geq 1} \beta_{i,j}\psi_i(\mathbf{x}) \textit{ and } \int \left(\alpha_i^{\mathbf{x}}\right)^2 dG(\mathbf{x}) = \sum_{j\geq 1} \beta_{i,j}^2.$$

Lemma 16. *Using the same notation as Lemmas 14 and 15, we have*

$$\beta_{i,j} = \int \alpha_i^{\mathbf{x}}\psi_j(\mathbf{x})dG(\mathbf{x}) = \int \alpha_j^\theta \phi_i(\theta)dF(\theta).$$

Proof. Follows from plugging the definitions of $\alpha_i^{\mathbf{x}}$ and α_j^θ into the expressions above and recalling the definition of $\beta_{i,j}$. □

Lemma 17. *$\sum_{j\geq J}\int \left(\alpha_j^{\theta}\right)^2 dF(\theta) = c_{K_{\mathbf{x}}}O(\lambda_J^{\mathbf{x}})$ and $\sum_{i\geq I}\int \left(\alpha_i^{\mathbf{x}}\right)^2 dG(\mathbf{x}) = c_{K_{\theta}}O(\lambda_I^{\theta})$.*

Proof. Note that $||h_\theta(.)||^2_{\mathcal{H}_{K_{\mathbf{x}}}} = \sum_{j\geq 1}\frac{\left(\alpha_j^{\theta}\right)^2}{\lambda_j^{\mathbf{x}}}$. Using Assumption 4 and that the eigenvalues are decreasing it follows that

$$\sum_{j\geq J}\left(\alpha_j^{\theta}\right)^2 = \sum_{j\geq J}\left(\alpha_j^{\theta}\right)^2\frac{\lambda_j^{\mathbf{x}}}{\lambda_j^{\mathbf{x}}} \leq \lambda_J^{\mathbf{x}}||h_\theta(.)||^2_{\mathcal{H}_{K_{\mathbf{x}}}} \leq \lambda_J^{\mathbf{x}}c_\theta^2,$$

and therefore $\sum_{j\geq J}\int \left(\alpha_j^{\theta}\right)^2 dF(\theta) \leq \lambda_J^{\mathbf{x}}\int_z c_\theta^2 dF(\theta) = c_{K_{\mathbf{x}}}O(\lambda_J^{\mathbf{x}})$. The proof of the second statement is analogous to this.

□

Theorem 3. *Under the stated assumptions, the bias is bounded by*

$$B(\mathcal{L}_{I,J},\mathcal{L}) = c_{K_{\mathbf{x}}}O\left(\lambda_J^{\mathbf{x}}\right) + c_{K_\theta}O(\lambda_I^{\theta}).$$

Proof. By using orthogonality, we have that

$$B(\mathcal{L}_{I,J},\mathcal{L}) \overset{\text{def}}{=} \iint \left(\mathcal{L}(\mathbf{x};\theta) - \mathcal{L}_{I,J}(\mathbf{x},\theta)\right)^2 dG(\mathbf{x})dF(\theta) \leq \sum_{j>J}\sum_{i\geq 1}\beta_{i,j}^2 + \sum_{i>I}\sum_{j\geq 1}\beta_{i,j}^2$$
$$= \sum_{j\geq J}\int \left(\alpha_j^{\theta}\right)^2 dF(\theta) + \sum_{i\geq I}\int \left(\alpha_i^{\mathbf{x}}\right)^2 dG(\mathbf{x}),$$

where the last equality follows from Lemmas 14 and 15. The Theorem follows from Lemma 17 . □

References

S. Bridle et al. Handbook for the great08 challenge: An image analysis competition for cosmological lensing. *The Annals of Applied Statistics*, pages 6–37, 2009. 2

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 8. The MIT Press, 2010. 1

H. Q. Minh. Some properties of Gaussian Reproducing Kernel Hilbert Spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010. 8

E. S. Sheldon, C. E. Cunha, R. Mandelbaum, J. Brinkmann, and B. A. Weaver. Photometric redshift probability distributions for galaxies in the SDSS DR8. *The Astrophysical Journal Supplement Series*, 201(2):32, 2012. 1, 2

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. 1

K. Sinha and M. Belkin. Semi-supervised learning using sparse eigenfunction bases. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1687–1695. 2009. 4, 5, 7