PHOTO-z ESTIMATION: AN EXAMPLE OF NONPARAMETRIC CONDITIONAL DENSITY ESTIMATION UNDER SELECTION BIAS[1]

BY RAFAEL IZBICKI[*], ANN B. LEE[†] AND PETER E. FREEMAN[†]

Federal University of São Carlos[*] and Carnegie Mellon University[†]

Redshift is a key quantity for inferring cosmological model parameters. In photometric redshift estimation, cosmologists use the coarse data collected from the vast majority of galaxies to predict the redshift of individual galaxies. To properly quantify the uncertainty in the predictions, however, one needs to go beyond standard regression and instead estimate the *full conditional density* $f(z|\mathbf{x})$ of a galaxy's redshift z given its photometric covariates $\mathbf{x}$. The problem is further complicated by *selection bias*: usually only the rarest and brightest galaxies have known redshifts, and these galaxies have characteristics and measured covariates that do not necessarily match those of more numerous and dimmer galaxies of unknown redshift. Unfortunately, there is not much research on how to best estimate complex multivariate densities in such settings.

Here we describe a general framework for properly constructing and assessing nonparametric conditional density estimators under selection bias, and for combining two or more estimators for optimal performance. We propose new improved photo-z estimators and illustrate our methods on data from the Sloan Data Sky Survey and an application to galaxy–galaxy lensing. Although our main application is photo-z estimation, our methods are relevant to any high-dimensional regression setting with complicated asymmetric and multimodal distributions in the response variable.

1. Introduction. Technological advances over the last two decades have ushered in the era of "precision cosmology," with the construction of catalogs that contain data on upward of 10^8 galaxies [e.g., Aihara et al. (2011)]. Cosmologists use these data to place progressively tighter constraints on the parameters of the ΛCDM model, the leading model explaining the structure and evolution of the Universe [see, e.g., Springel, Frenk and White (2006)].

To estimate distances to astronomical sources, that is, to place them in time relative to the Big Bang, cosmologists need to determine a galaxy's *redshift*, the
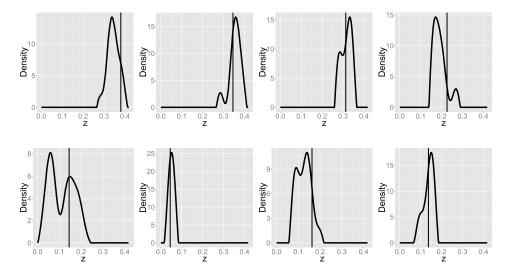
FIG. 1. *Photometric estimates of $f(z|\mathbf{x})$ for eight randomly chosen galaxies from the Sloan Digital Sky Survey (SDSS). Here we use the Series estimator from Section 5.3. Many of these densities are highly asymmetric and multimodal. The vertical lines indicate the true redshift, determined via high-resolution spectroscopy.*

increase in the wavelength of a traveling photon due to the expansion of the Universe. One can accurately estimate redshift via *spectroscopy*, but, because of cost and time considerations, more than 99 percent of today's galaxy observations are instead from *photometry*, a fast but low-resolution measuring technique where a few broad-band filters coarsely record the radiation from an astronomical object.

The goal of *photometric redshift estimation* (or photo-z estimation) is to estimate a galaxy's redshift z given its photometric covariates $\mathbf{x}$. Traditionally, redshift estimation has been viewed as a regression problem where one seeks the conditional mean $\mathbb{E}[z|\mathbf{x}]$ or the most probable redshift of a galaxy. Recent work, however, shows the importance of estimating the *full conditional density* $f(z|\mathbf{x})$ [Ball and Brunner (2010)]. In photometry, $f(z|\mathbf{x})$ is often asymmetric and multimodal, with errors that are heteroscedastic. Figure 1 shows density estimates for eight randomly chosen galaxies from the Sloan Digital Sky Survey (SDSS): these distributions are complicated non-Gaussian distributions which cannot easily be summarized by, for example, means and variances.

By working with a density estimate $\widehat{f}(z|\mathbf{x})$ instead of a point estimate of z, one can dramatically reduce systematic errors in downstream analysis, that is, when estimating functions $g(z)$ of an unknown redshift z; see, for example, Mandelbaum et al. (2008), Sheldon et al. (2012), Wittman (2009). Although optimizing the function g would be optimal for the problem of choice, often there is no clear g ahead of time or there exist many different functions g for the same application; for example, for the galaxy–galaxy lensing problem in Section 6, each of $\approx 500{,}000$ lenses

has a different calibration function g. Hence, a common practice in astronomy is to build catalogs of photo-z density estimates (for galaxies in a survey), which can then be used to address a range of different inference problems in astronomy and cosmology. Many of these estimates, however, include handpicked tuning parameters and, as we later show in Section 6 (Figure 10), estimates that yield good results for a particular application or function of z do not necessarily predict the redshift z per se well. This brings up the question of how to properly construct and assess conditional density estimators, in general, and photo-z estimators, in particular.

On the methodological side, photo-z estimation presents challenges at the boundaries of current statistical research. There exist a wide range of sophisticated techniques for high-dimensional inference, but few attempt to estimate full conditional densities or ratios of high-dimensional densities [see, e.g., Izbicki and Lee (2016), Izbicki, Lee and Schafer (2014) and references within]. Even less is known about how to estimate conditional densities *under selection bias*. Statistics and machine learning procedures typically assume that training and test data have similar distributions, but the two distributions can be very different for sky surveys that mix spectroscopy and photometry, and, for example, remote sensing applications where different sensors may malfunction at different rates or collect data at different rates [see Moreno-Torres et al. (2012) for data set shift in *classification*].

In the astronomy literature, existing photo-z algorithms roughly fall into two categories. In the first category, *template fitting* [e.g., Fernández-Soto, Lanzetta and Yahil (1998)] estimates $f(z|\mathbf{x})$ by directly comparing observed data with a suite of idealized photometric data sets for different types of galaxies at different redshifts. Our interest lies in the second category, *empirical redshift estimation*, in which one builds an estimator of $f(z|\mathbf{x})$ using a training set of galaxies with spectroscopically confirmed redshifts [see, e.g., Ball and Brunner (2010), Kind and Brunner (2013), Zheng and Zhang (2012)]. In photo-z estimation, however, high-resolution spectroscopy is extremely time-consuming. For example, one would already need $\sim$2.5 years of dedicated telescope time to estimate the spectroscopic redshifts for the 500,000-galaxy photometric SDSS sample in our paper; the problem only gets worse with deeper and larger surveys. As a result, usually only the rarest and brightest galaxies are spectroscopically observed, and these galaxies have characteristics and measured covariates that do not necessarily match those of more numerous dimmer galaxies [see, e.g., Ball and Brunner (2010), Oyaizu et al. (2008)]. The goal of this paper is to develop improved nonparametric methods for conditional density estimation that explicitly deal with photo-z estimation and selection bias:

(i) On the methods side, we present a general framework for supervised learning that accounts for differences in training and test data, and that, unlike standard nonparametric regression and classification, can handle highly asymmetric and multimodal distributions. We design appropriate loss functions for a multivariate setting with selection bias [equations (5) and (9)], and we show how to

estimate these functions using labeled and unlabeled data. Our setup provides a principled method for choosing tuning parameters, for selecting covariates, and for comparing and combining (Section 5.4) different conditional density estimators for optimal performance. The final density estimates lead to more accurate *predictive intervals* (see, e.g., Figure 10) for new observations. Estimating $f(z|\mathbf{x})$ is also a simple way of performing *nonparametric quantile regression* of many quantiles simultaneously [Koenker (2005)].

  (ii) In the context of photo-z prediction, we propose more accurate algorithms for estimating photo-z probability distributions in a setting where primarily nearby and bright galaxies have known redshifts. Using SDSS as well as simulated data, we analyze and compare different methods for estimating conditional densities and so-called importance weights (which are used to correct for selection bias). We introduce two new conditional density estimators, *Kernel nearest neighbors* (Section 5.2) and *Series* (Section 5.3), that both have better performance than the photo-z prediction method by Cunha et al. (2009), which represents the state of the art for empirical photo-z estimators under selection bias. We also present different diagnostic tests for evaluating the goodness of fit of estimated densities (Appendix A).

The organization of the paper is as follows: Section 2 describes our data. In Section 3, we introduce the statistical problem and the idea of importance weights. Section 4 compares different schemes for estimating these weights. Then, in Section 5, we shift our focus to the problem of estimating conditional densities under selection bias, and the issue of proper model selection and assessment. Finally, in Section 6, we offer some new insights on the galaxy lensing-lensing example from Sheldon et al. (2012). We conclude our work with Section 7.

2. Data.   There are two main types of astronomical data: (i) *spectroscopic* data, where both the covariates $\mathbf{x}$ and the redshift z (the label) can be measured with negligible error, and (ii) *photometric* data, where only the covariates $\mathbf{x}$ are known and there is no precise measurement of the redshift. In our study, we use photometric and spectroscopic data from the Sloan Digital Sky Survey [SDSS; York et al. (2000)], as well as SDSS-based simulated data with known levels of selection bias.

2.1. *SDSS photometric data.*   Since 1998, SDSS has collected data on over 200 million galaxies that are spread across one-quarter of the sky. The vast majority ($\gtrsim 99\%$) of these galaxies are only photometrically observed.

In photometry, different filters are sequentially placed into a telescope's light path. Each filter only allows passage of photons in a particular wavelength band. SDSS measures photon fluxes, or, equivalently, *magnitudes* (the logarithm of fluxes), in five bands, denoted u, g, r, i, and z, in the wavelength range 3.5×10^{-7} meters to 9×10^{-7} meters (i.e., from UV light through the optical regime to infrared light). Magnitude estimates are algorithm-dependent, in that they depend on

how one defines a boundary around a galaxy and how one sums the light within that boundary. SDSS catalogs include estimates from many different boundary-definition algorithms or *magnitude systems*; in this work, we follow Sheldon et al. (2012) and use `model` and `cmodel` magnitudes. We also work with *colors*, or differences of magnitudes in adjacent wavelength bands.

Our final SDSS photometric data set contains 10 covariates (four colors and the associated r-band magnitude from each algorithm) for 538,974 galaxies in an ≈ 72 square-degree sky patch.[2] These galaxies are extracted from SDSS Data Release 8 [DR8; Aihara et al. (2011)] and filtered according to Sheldon et al. (2012) so that the photometric sample lies in the same range as the spectroscopic sample.

2.2. *SDSS spectroscopic data.* Of the over 200 million galaxies in SDSS, some one million have been the subject of follow-up *spectroscopic* observations. In spectroscopy, a light-dispersing grating or prism is placed into a telescope's light path. The photon changes its path with an angle that is proportional to its wavelength. Thanks to high-resolution mapping of the dispersed light, one can finely resolve narrow spectral features (or lines) that are smeared out in photometry. These lines, which are caused by transitions of electrons between atomic energy levels, occur at known wavelengths λ, but are observed at wavelengths $(1+z)\lambda$, where z is the galaxy's redshift.[3] One can use the wavelength ratios of two or more observed lines to infer which transitions they represent; once that information is known, redshift estimation is trivial. The precision in the estimates is typically $\Delta z/z \sim 10^{-6}$, that is, for spectroscopic redshifts, we can safely ignore the measurement error.

In this paper, we use the same data (i.e., colors and r-band magnitudes, collectively denoted $\mathbf{x}$, and redshifts z) as in Sheldon et al. (2012). This data set includes 435,875 galaxies; the vast majority are taken from SDSS DR8, but some fainter galaxies have been added so that the spectroscopic covariates cover the same space as the photometric covariates, albeit with a different distribution (E. Sheldon, private communication).

Figure 2 shows the distributions of the spectroscopic and photometric samples. The left panel shows that there is a clear selection bias in the study where the galaxies in the spectroscopic sample tend to be brighter (i.e., they tend to have a lower r-band magnitude) than the galaxies in the photometric sample.

2.3. *Simulated samples with known levels of selection bias.* In addition to the SDSS data, we construct photometric samples with *known* levels of selection bias relative to the SDSS spectroscopic data. We use the following rejection sampling

---

[2]Celestial longitude, or right ascension (RA) $\in [168°, 192°]$ and celestial latitude, or declination $(\delta) \in [-1.5°, 1.5°]$.

[3]In astronomy, the notation z is used to denote both redshift and a particular photometric band. For the remainder of this work, z will always represent redshift.
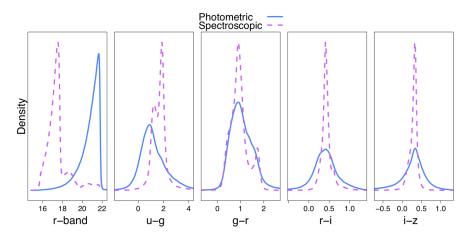
FIG. 2.   *Distribution of r-band* model *magnitude and colors (i.e., differences of* model *magnitude values in adjacent photometric bins) for photometric versus spectroscopic SDSS data. Many galaxies in the spectroscopic sample are brighter (i.e., they have a lower r-band magnitude as indicated in the left panel) than the galaxies in the photometric sample.*

algorithm: Let $\mathbf{x}$ be a data point in the spectroscopic sample with r model magnitude $x^{(r)}$ (scaled to be between 0 and 1). A large r model magnitude corresponds to a faint galaxy. Let S be a binary selection variable that decides whether a galaxy in the spectroscopic sample is included in the photometric sample ($S = 1$) or not ($S = 0$). We assume that the probability $P(S = 1|\mathbf{x})$ depends on $\mathbf{x}$ through $x^{(r)}$ only; that is, that $P(S = 1|\mathbf{x}) = P(S = 1|x^{(r)})$. To create levels of selection bias realistic for different astronomical surveys, we resample the SDSS spectroscopic sample according to the following:

Scheme 1: $P(S = 1|x^{(r)}) = f_{B(1,1)}(x^{(r)}) \equiv 1$,
Scheme 2: $P(S = 1|x^{(r)}) = f_{B(13,4)}(x^{(r)})/\max_{x^{(r)}} f_{B(13,4)}(x^{(r)})$,
Scheme 3: $P(S = 1|x^{(r)}) = f_{B(18,4)}(x^{(r)})/\max_{x^{(r)}} f_{B(18,4)}(x^{(r)})$,

where $f_{B(i,j)}$ denotes the density of a beta random variable with parameters (i, j).

Figure 3 shows the resulting r-band distributions. Scheme 1 involves no selection bias: the spectroscopic and photometric data have r-band magnitudes with the same distribution. At the other extreme is Scheme 3 with strong selection bias: most photometrically observed galaxies are significantly fainter (shifted toward large r-band magnitude) than the galaxies in the spectroscopic data set.

3. Selection bias, covariate shift, and importance weights.   A standard assumption in statistics and machine learning is that labeled and unlabeled data have similar distributions, but, as Figure 2 (left panel) shows, the two distributions can be very different for sky surveys that mix spectroscopy and photometry: Brighter galaxies (or galaxies with a lower r-band magnitude) are more likely to be selected
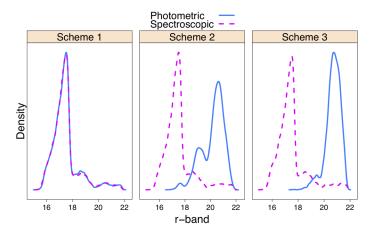
FIG. 3. *Distribution of the r-band* `model` *magnitudes for three different schemes of resampling the SDSS spectroscopic sample. The simulated data sets have* (*known*) *levels of selection bias realistic for different astronomical surveys.*

for follow-up spectroscopic observation. This fact motivates the methods and algorithms presented in this paper. Below, we fix our notation and describe the main ideas behind importance weights as a way of correcting for sample selection bias.

3.1. *Problem statement and notation.* Our data are covariates $\mathbf{x} \in \mathbb{R}^d$ (photometric colors and magnitudes) and labels $z \in \mathbb{R}$ (the redshift). Without loss of generality, we rescale the original redshift values so that the response $z \in [0, 1]$.

Suppose we have access to an i.i.d unlabeled sample $\mathbf{x}_1^U, \ldots, \mathbf{x}_{n_U}^U$ with only photometric data, and an i.i.d. labeled sample $(\mathbf{x}_1^L, z_1^L), \ldots, (\mathbf{x}_{n_L}^L, z_{n_L}^L)$ from a potentially different distribution, where the labels are from follow-up spectroscopic studies. Because of the high cost of spectroscopy in sky surveys, in terms of telescopic resources, $n_L \ll n_U$, and the distributions of the labeled and unlabeled samples will *not* be the same. Our goal in this paper is to construct a photo-z density estimator $\widehat{f}(z|\mathbf{x})$ that performs well on the *unlabeled* photometric data, which roughly dominate 99% of today's galaxy observations.

To fix our notation, let $\mathbb{P}_L$ and $\mathbb{P}_U$ denote the distributions on the labeled and unlabeled samples, respectively; that is, let $(\mathbf{x}^L, z^L) \sim \mathbb{P}_L$ and $(\mathbf{x}^U, z^U) \sim \mathbb{P}_U$, where z^U are missing data labels. We say that there is a *data set shift* if $\mathbb{P}_L \neq \mathbb{P}_U$. To understand how this affects learning algorithms, one needs to make additional assumptions about the relationship between $\mathbb{P}_L$ and $\mathbb{P}_U$ [e.g., Quionero-Candela et al. (2009), Gretton et al. (2010), Moreno-Torres et al. (2012)]. For our application, we assume that the probability that a galaxy is labeled with a spectroscopic redshift is independent of the response variable z if we condition on the covariates $\mathbf{x}$ [Lima et al. (2008), Sheldon et al. (2012)]; that is,

$$(1) \qquad \mathbb{P}(L = 1|\mathbf{x}, z) = \mathbb{P}(L = 1|\mathbf{x}),$$
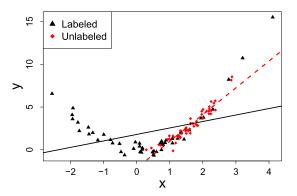
FIG. 4. *Toy example for covariate shift in linear regression*: *The solid line is reasonable for predicting new data from the labeled sample*, *but it is far from optimal for the unlabeled sample*.

where the random variable L equals 1 if a datum is labeled and 0 otherwise. This type of sample selection bias where equation (1) holds is sometimes referred to as a *missing at random* [MAR; Rubin (1976), Moreno-Torres et al. (2012)] bias.[4] MAR bias implies *covariate shift*, defined as

$$(2) \qquad f_L(\mathbf{x}) \neq f_U(\mathbf{x}), \qquad f_L(z|\mathbf{x}) = f_U(z|\mathbf{x}).$$

Under certain conditions on the support of $f_L(\mathbf{x})$ and $f_U(\mathbf{x})$, covariate shift also implies MAR bias [Moreno-Torres et al. (2012)]. Below we use the terms interchangeably to refer to the assumption in equation (1).

At first glance, it may seem that MAR bias would not pose a problem for density estimation: Because $f(z|\mathbf{x})$ is the same for both labeled and unlabeled samples [equation (2)], one might conclude that a good estimator of $f(z|\mathbf{x})$ based on labeled data would also have good performance for unlabeled data. This is in general not true. Often $f(z|\mathbf{x})$ is well estimated only in regions where there is plenty of labeled data, and these regions may not coincide with regions where there is plenty of unlabeled (target) data; see Figure 4 for a toy example of an equivalent regression problem with covariate shift. From a statistical perspective, this problem arises because the loss function used for estimating $f(z|\mathbf{x})$ (or, analogously, $\mathbb{E}[z|\mathbf{x}]$ in regression) depends on the *marginal* distribution of $\mathbf{x}$. Hence, *an estimator that performs well with respect to $f_L(\mathbf{x})$ may not perform well with respect to $f_U(\mathbf{x})$.*

One solution to this mismatch problem is to reweight the labeled data so that their distribution $\mathbb{P}_L$—after a reweighing with the so-called *importance weights* $\beta(\mathbf{x}) := f_U(\mathbf{x})/f_L(\mathbf{x})$—matches the distribution $\mathbb{P}_U$ of the unlabeled data. We can

---

[4]MAR is to be distinguished from data missing completely at random (MCAR), which occurs when the sampling method is completely independent of $\mathbf{x}$ and z, that is, $\mathbb{P}(L = 1|\mathbf{x}, z) = \mathbb{P}(L = 1)$, and hence there is no data set shift.

then compute expectations with respect to the target distribution $\mathbb{P}_U$ using the distribution $\mathbb{P}_L$ of the labeled training data. In particular, if $l(\widehat{f}, f)$ is the loss function for the estimated conditional density $\widehat{f}$, we use that

$$(3) \quad \mathbb{E}_{(\mathbf{X}, Z) \sim \mathbb{P}_U}[l(\widehat{f}(Z|\mathbf{X}), f(Z|\mathbf{X}))] = \mathbb{E}_{(\mathbf{X}, Z) \sim \mathbb{P}_L}[l(\widehat{f}(Z|\mathbf{X}), f(Z|\mathbf{X}))\beta(\mathbf{X})].$$

The last expression of equation (3) involves unknown weights $\beta(\mathbf{x})$. Open questions are as follows:

(i) how to best estimate the importance weights $\beta(\mathbf{x}) := f_U(\mathbf{x})/f_L(\mathbf{x})$, where $\mathbf{x}$ is our multivariate data, and

(ii) how to design an estimator of the full conditional density $f(z|\mathbf{x})$ that performs well on the target data, that is, that has a small risk according to equation (3).

In this work, we propose a greedy two-step approach to conditional density estimation under selection bias where one first selects the best model for estimating $\beta(\mathbf{x})$ (see Section 4), and then uses these estimates to search for the best model for estimating $f(z|\mathbf{x})$ under covariate shift (see Section 5). Note that a necessary condition for computing importance weights is that $\mathbb{P}_L$ *dominates* $\mathbb{P}_U$, that is, $\mathbb{P}_L(\mathbf{x}) \gg \mathbb{P}_U(\mathbf{x})$; hence, for this study, we have selected photometric data whose covariates lie within the domain of the spectroscopic covariates (see Section 2.1).

4. Estimating importance weights. A naive method for computing $\beta(\mathbf{x})$ is to separately estimate f_U and f_L and to then take their ratio, but this approach can enhance errors in the individual density estimates, particularly in regions where f_L is nearly zero [Sugiyama et al. (2008)]. There are also several other approaches for estimating density ratios (see Appendix B).

The goal in this section is to find out how these estimators perform in practice on our data. To assess the estimators, we use the weighted loss function by Izbicki, Lee and Schafer (2014) defined as

$$L(\widehat{\beta}, \beta) := \int (\widehat{\beta}(\mathbf{x}) - \beta(\mathbf{x}))^2 \, dP_L(\mathbf{x})$$

(4)

$$= \int \widehat{\beta}^2(\mathbf{x}) \, dP_L(\mathbf{x}) - 2 \int \widehat{\beta}(\mathbf{x}) \, dP_U(\mathbf{x}) + K,$$

where K is a constant that does not depend on $\widehat{\beta}(\mathbf{x})$. (Appendix B gives some intuition behind the choice of this loss function.) We divide our data into three parts: a training set used to fit the model, a validation set for model selection and tuning of parameters, and a test set for assessing the final model [Hastie, Tibshirani and Friedman (2009), page 222]. For model selection and model assessment, we estimate $L(\widehat{\beta}, \beta)$ according to Izbicki, Lee and Schafer (2014):

$$(5) \qquad \widehat{L}(\widehat{\beta}, \beta) = \frac{1}{\tilde{n}_L} \sum_{k=1}^{\tilde{n}_L} \widehat{\beta}^2(\tilde{\mathbf{x}}_k^L) - 2\frac{1}{\tilde{n}_U} \sum_{k=1}^{\tilde{n}_U} \widehat{\beta}(\tilde{\mathbf{x}}_k^U),$$

where $\tilde{\mathbf{x}}_1^L, \ldots, \tilde{\mathbf{x}}_{\tilde{n}_L}^L$ represent the labeled (validation or test) data, and $\tilde{\mathbf{x}}_1^U, \ldots, \tilde{\mathbf{x}}_{\tilde{n}_U}^U$ represent the unlabeled (validation or test) data.

Experiments.   We use this loss function to compare six different estimators of β:

- β-*NN*, the nearest neighbor estimator from Lima et al. (2008) [equation (20) in the Appendix B], but with the number of nearest neighbors M chosen so as to minimize our estimated loss [equation (5)] on the validation data [as in Kremer et al. (2015)];
- β-*NN1*, the nearest neighbor approach with $M = 1$ [Loog (2012)];
- β-*KLIEP* and β-*uLSIF*, importance weight estimators suggested by Sugiyama et al. (2008) and Kanamori, Hido and Sugiyama (2009), respectively, and implemented with the authors' MATLAB code;[5]
- β-*KuLSIF*, a kernelized version of β-*uLSIF* [Kanamori, Suzuki and Sugiyama (2012)]; and
- β-*Series*, the density ratio estimator from Izbicki, Lee and Schafer (2014).

Following Lima et al. (2008) and Cunha et al. (2009), our covariates are the four colors and the r-band magnitude in the model magnitude system. At the end of this section as well as at the end of Section 5.5 on CDE, we discuss variable selection on the full set of 10 covariates in Section 2.

We study the estimators under the simulated selection bias settings from Section 2.1 (Schemes 1–3), using labeled and unlabeled samples that are each of size 10,000. (For each sample, we randomly choose 2800 galaxies for training, 1200 for validation, and 6000 for testing.) We also apply the estimators to the SDSS data. These data have a large covariate shift (Figure 2), where $\widehat{\beta}(\mathbf{x}) = 0$ for $\approx 80\%$ of the labeled examples; that is, the majority of the spectroscopic sample lies in regions of covariate space where there are no unlabeled data. Having more labeled data may seem harmless, but it turns out that if one includes these labeled examples in the training sample, the *effective sample size* [defined in Shimodaira (2000) and Gretton et al. (2010); a quantity which measures how far the weights' distribution is from a uniform distribution] will be very small, ultimately resulting in poor conditional density estimates (see Section 5.5 and Figure 8). Essentially, many galaxies have no contribution (i.e., zero weight) in the conditional density estimation with $\widehat{\beta}(\mathbf{x}_k^L) = 0$ in, for example, equations (11) and (12). To mitigate this problem, we use a similar data cleaning and resampling step as in Lima et al. (2008): First, we construct a β-*NN* estimator using a preliminary sample of 10,000 spectroscopic and 10,000 photometric (randomly selected) galaxies.

Based on these estimates of $\widehat{\beta}(\mathbf{x})$, we then resample the spectroscopic survey data until we have a sample of size 15,000, which only includes galaxies from SDSS at points where $\widehat{\beta}(\mathbf{x}) \neq 0$. The SDSS results in this paper are based on the *new* spectroscopic sample together with 15,000 randomly selected photometric

---

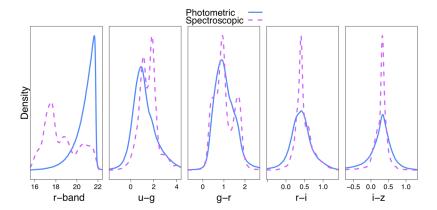[5]http://www.ms.k.u-tokyo.ac.jp/software.html.

FIG. 5. *Distribution of the r-band* `model` *magnitude and the four colors from* `model` *magnitude for the SDSS photometric and spectroscopic samples* after *we replace labeled examples for which the initial estimates of the importance weights are zero*; *compare with Figure* 2. *By restricting the labeled data to the regions of interest* (i.e., *to the regions where there are unlabeled data*), *we now discard examples that have no contribution* (i.e., *zero weight*) *in our conditional density estimators.*

observations; Figure 5 shows the distributions of the two samples after the pre-processing step. In all experiments on SDSS data, we use 3500 galaxies from each sample (labeled and unlabeled) for training, 1500 for validation, and the remainder for testing.

Figure 6 compares the different estimators of β. The nearest neighbor method is rarely mentioned in the literature on data set shift, but in our experiments we find that β-*NN*, where M is chosen by data splitting, consistently performs the same or better than other competing (and more complicated) methods; β-*Series* is a close second. This is in agreement with recent independent work by Kremer et al. (2015). We also note that by minimizing equation (5), we select $M = 8$ neighbors for the SDSS data set, a value similar to the value $M = 5$ chosen in an ad hoc manner by Cunha et al. (2009). Henceforth, we will use the β-*NN* method [defined in Appendix B, equation (20)] to estimate importance weights.

Variable selection. One can further improve on these results by variable selection on the full set of 10 covariates described in Section 2. Table 1 lists the selected covariates in a forward stepwise model search with β-*NN*. For Scheme 1, where there is no selection bias [and where $\beta(\mathbf{x})$, as a result, does not depend on $\mathbf{x}$], the final model includes only one covariate. In Schemes 2 and 3, the importance weights depend on the `model` r-band magnitude, and the selected model always includes this covariate. For the SDSS data, we achieve a loss of -2.41 ± 0.08 with variable selection, which is significantly smaller than the value -2.16 ± 0.05 when including all 10 covariates, and the value -1.97 ± 0.04 (see β-*NN* in Figure 6d) when using the five covariates from the `model` magnitude system only as in Sheldon et al. (2012).

(a) Scheme 1   (b) Scheme 2

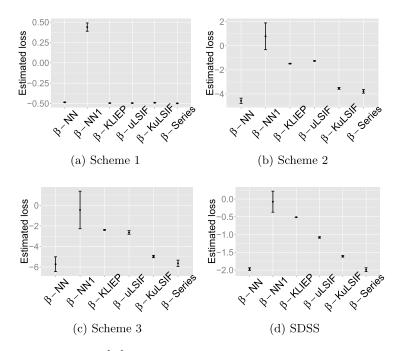(c) Scheme 3   (d) SDSS

FIG. 6.   *The estimated loss* $\widehat{L}(\widehat{\beta}, \beta)$ *of different estimators of* $\beta(\mathbf{x})$ *for Schemes* 1–3 *with, respectively,* (a) *no,* (b) *moderate, and* (c) *large covariate shift. Panel* (d) *shows* $\widehat{L}(\widehat{\beta}, \beta)$ *for observed SDSS data. The bars in each panel correspond to the mean estimated loss* ± 1 *standard error. These results indicate that* β-NN *consistently performs the same or better than other competing methods;* β-Series *is a close second.*

5. Conditional density estimation under covariate shift.   Conditional density estimators are typically designed to minimize the loss

$$(6) \qquad \iint \left(\widehat{f}(z|\mathbf{x}) - f(z|\mathbf{x})\right)^2 dP_L(\mathbf{x})\, dz$$

TABLE 1
Selected covariates for estimating importance weights with the β-NN *estimator*

Data set	**model**					**cmodel**				
	r	$u-g$	$g-r$	$r-i$	$i-z$	r	$u-g$	$g-r$	$r-i$	$i-z$
Scheme 1			X							
Scheme 2	X					X			X	
Scheme 3	X	X								
SDSS	X	X				X	X	X		

under the implicit assumption that $\mathbb{P}_L = \mathbb{P}_U$.[6] One can easily estimate this loss (up to a constant) using the labeled data:

$$(7) \qquad \frac{1}{\tilde{n}_L} \sum_{k=1}^{\tilde{n}_L} \int \hat{f}^2(z|\tilde{\mathbf{x}}_k^L)\,dz - 2\frac{1}{\tilde{n}_L} \sum_{k=1}^{\tilde{n}_L} \hat{f}(\tilde{z}_k^L|\tilde{\mathbf{x}}_k^L).$$

However, our goal is really to minimize

$$(8) \qquad L(\hat{f}, f) := \iint \big(\hat{f}(z|\mathbf{x}) - f(z|\mathbf{x})\big)^2\,d P_U(\mathbf{x})\,dz,$$

where $\mathbb{P}_U$ is the distribution of the *unlabeled* target data. The two losses can be very different if $\mathbb{P}_L \neq \mathbb{P}_U$. Hence, a density estimator that performs well on the labeled data may not be a good estimator for the data of interest; Figure 4 shows a similar problem in linear regression.

The challenge is to estimate equation (8) without knowing z at the unlabeled data points. Under the covariate shift assumption $f_U(z|\mathbf{x}) = f_L(z|\mathbf{x})$, one can use importance sampling. Up to a constant, equation (8) becomes

$$
\begin{aligned}
L(\hat{f}, f) &= \iint \hat{f}^2(z|\mathbf{x})\,d P_U(\mathbf{x})\,dz - 2 \iint \hat{f}(z|\mathbf{x}) f(z|\mathbf{x})\,d P_U(\mathbf{x})\,dz \\
&= \iint \hat{f}^2(z|\mathbf{x})\,d P_U(\mathbf{x})\,dz - 2 \iint \hat{f}(z|\mathbf{x})\beta(\mathbf{x})\,d P_L(z, \mathbf{x}),
\end{aligned}
$$

where for the second equality we use that $f_U(z|\mathbf{x})\,d P_U(\mathbf{x})\,dz = f_L(z|\mathbf{x}) \times \beta(x)\,d P_L(\mathbf{x})\,dz = \beta(x)\,d P_L(z, \mathbf{x})$. Hence, when $\mathbb{P}_L \neq \mathbb{P}_U$, we replace the standard empirical loss in equation (7) by

$$(9) \qquad \hat{L}(\hat{f}, f) = \frac{1}{\tilde{n}_U} \sum_{k=1}^{\tilde{n}_U} \int \hat{f}^2(z|\tilde{\mathbf{x}}_k^U)\,dz - 2\frac{1}{\tilde{n}_L} \sum_{k=1}^{\tilde{n}_L} \hat{f}(\tilde{z}_k^L|\tilde{\mathbf{x}}_k^L)\hat{\beta}(\tilde{\mathbf{x}}_k^L).$$

We can compute this loss using (labeled and unlabeled) validation data and $\hat{\beta}(\tilde{\mathbf{x}}_k^L)$, which are estimates of the importance weights at the labeled data points.

In what follows (Sections 5.1–5.4), we present four conditional density estimators specifically designed for multivariate data and a covariate shift (CS) setting: NN_{CS}, $ker\text{-}NN_{CS}$, $Series_{CS}$, and $Comb_{CS}$. The NN_{CS} nearest neighbor histogram estimator first appeared in Cunha et al. (2009); the other estimators represent novel approaches. For model selection and tuning of parameters, we minimize equation (9) with importance weights β estimated via equations (20) and (5). The same loss is also used for model assessment of the SDSS data. For model assessment of the *simulated* data (where we know the true labels $\tilde{z}^U$), we compute the more accurate

---

[6]Note that $\hat{f}(z|\mathbf{x})$ stands for the estimator of $f(z|\mathbf{x})$, the density at point $\mathbf{x}$. This estimator can be computed using a subset of the covariates in $\mathbf{x}$ as in Section 5.5 *Variable Selection*.

error estimate

$$(10) \qquad \widehat{L}(\widehat{f}, f) = \frac{1}{\widetilde{n}_U} \sum_{k=1}^{\widetilde{n}_U} \int \widehat{f}^2(z | \widetilde{\mathbf{x}}_k^U) \, dz - 2 \frac{1}{\widetilde{n}_U} \sum_{k=1}^{\widetilde{n}_U} \widehat{f}(\widetilde{z}_k^U | \widetilde{\mathbf{x}}_k^U),$$

which does not involve importance weights β.

REMARK 1 (Choice of loss function).   *One can replace the averaged L^2-loss in equation* (8) *with other measures of discrepancy, but many of the distance measures common in discrimination analysis (e.g., F-divergences and differences in log-densities) are profoundly sensitive to the tails of the distribution and not suitable for density estimation; see, e.g.,* Hall (1987) *and* Wasserman (2006). *Unlike the L^1 loss and related loss functions, we can (as shown in the paper) also estimate the L^2-loss from the data without knowledge of the true $f(z|x)$; this is not the case for L^1-related losses, which makes them less practical in applications.*

5.1. *Nearest neighbor histogram* (NN_{CS}).   In Cunha et al. (2009), the authors use a weighted nearest neighbor histogram to estimate the photo-z distribution of a galaxy with photometric covariates $\mathbf{x}$. Let $\mathcal{N}_N(\mathbf{x})$ denote the N nearest neighbors of $\mathbf{x}$ among the labeled data. Divide $[0, 1]$, the range of z, into B equal sized bins, and let $b(z)$ denote the bin that includes z for $z \in [0, 1]$. Then, the weighted histogram estimator is

$$(11) \qquad \widehat{f}(z|\mathbf{x}) \propto \sum_{k \in \mathcal{N}_N(\mathbf{x})} \widehat{\beta}(\mathbf{x}_k^L) \mathbb{I}(z_k^L \in b(z)),$$

where the importance weights $\widehat{\beta}(\mathbf{x}_k^L)$ reflect how representative each labeled galaxy $k \in \mathcal{N}_N(\mathbf{x})$ is of the target distribution. Cunha et al. choose the tuning parameters in their model by hand. Here we use equation (9) to find the optimal values of N and B for the conditional density estimator in equation (11), and we use equation (5) to select the best value of M in equation (20) for computing the weights $\widehat{\beta}(\mathbf{x})$.

5.2. *Kernel nearest neighbor estimator* (*ker-NN_{CS}*).   A simple way of improving upon NN_{CS} is to replace the indicator function in equation (11) with a kernel smoother:

$$(12) \qquad \widehat{f}(z|\mathbf{x}) \propto \sum_{k \in \mathcal{N}_N(\mathbf{x})} \widehat{\beta}(\mathbf{x}_k^L) K_\varepsilon(z - z_k^L),$$

where, for example, $K_\varepsilon(z - z_k) = e^{-(z-z_k)^2/4\varepsilon}$.[7] As before, we choose the tuning parameters [here, N and ε in equation (12), and M in equation (20)] that minimize

---

[7]For a traditional kernel nearest neighbors estimator *not* corrected for selection bias [Zhao and Liu (1985)], let $\widehat{\beta}(\mathbf{x}) \equiv 1$ for all $\mathbf{x}$; we denote the uncorrected estimator by *ker-NN*.

the estimated losses in equations (9) and (5). This estimator leads to much more accurate density estimates than its histogram equivalent when we have limited amounts of labeled data (and hence small values of N).

5.3. *Spectral series CDE under covariate shift* (*Series$_{CS}$*). Suppose that the covariates $\mathbf{x}$ lie in a lower-dimensional subspace $\mathcal{X}$ of $\mathbb{R}^d$, where the number of covariates d may be large. Izbicki and Lee (2016) propose a spectral series estimator that expands the conditional density $f(z|\mathbf{x})$ in a basis $\Psi_{i,j}(z, \mathbf{x}) = \phi_i(z)\psi_j(\mathbf{x})$ adapted to the *intrinsic* (lower-dimensional) geometry of a reference distribution $\mathbb{P}$ on $\mathcal{X}$. Here we generalize the series approach to a setting with covariate shift. We choose $\mathbb{P}_L$ as the reference distribution and tune the estimator so as to minimize the loss with respect to $\mathbb{P}_U$.

More specifically, we assume the functions ϕ_i to be standard (one-dimensional) Fourier basis functions, whereas the functions ψ_j are the eigenfunctions of the operator $\mathbf{K} : L^2(\mathcal{X}, \mathbb{P}_L) \longrightarrow L^2(\mathcal{X}, \mathbb{P}_L)$,

$$(13) \qquad \mathbf{K}(h)(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{y})h(\mathbf{y}) \, dP_L(\mathbf{y}),$$

where $K(\mathbf{x}, \mathbf{y})$ is a bounded, symmetric, and positive definite kernel. By construction [Lee and Izbicki (2016), Minh, Niyogi and Yao (2006)],

$$(14) \qquad \int_{\mathcal{X}} \psi_i(\mathbf{x})\psi_j(\mathbf{x}) \, dP_L(\mathbf{x}) = \delta_{i,j} \overset{\text{def}}{=} \mathbb{I}(i = j).$$

As a result, the coefficients in the series expansion are simply expectations over the eigenfunctions:

$$(15) \qquad \alpha_{i,j} = \iint f(z|\mathbf{x})\Psi_{i,j}(z, \mathbf{x}) \, dP_L(\mathbf{x}) \, dz = \mathbb{E}_{(\mathbf{X},Z)\sim\mathbb{P}_L}[\Psi_{i,j}(Z, \mathbf{X})].$$

In practice, we need to estimate both the functions ψ_j and the coefficients $\alpha_{i,j}$ from data: Using the labeled training examples, we compute the first J eigenvectors $\widetilde{\psi}_1, \ldots, \widetilde{\psi}_J$ of the Gram matrix

$$\left[K\left(\mathbf{x}_i^L, \mathbf{x}_j^L\right)\right]_{i,j=1}^n,$$

where $K(\mathbf{x}, \mathbf{y}) = \exp(-d^2(\mathbf{x}, \mathbf{y})/4\varepsilon)$ is the Gaussian kernel. We then extend these vectors to out-of-sample points via the Nyström extension

$$(16) \qquad \widehat{\psi}_j(\mathbf{x}) = \frac{\sqrt{n_L}}{\widehat{l}_j} \sum_{k=1}^{n_L} \widetilde{\psi}_j(\mathbf{x}_k^L) K(\mathbf{x}, \mathbf{x}_k^L),$$

where $\widehat{l}_j$ is the eigenvalue associated to the eigenvector $(\widetilde{\psi}_j(\mathbf{x}_1^L), \ldots, \widetilde{\psi}_j(\mathbf{x}_{n_L}^L))$. Next we estimate the expansion coefficients in equation (15) according to

$$\widehat{\alpha}_{i,j} = \frac{1}{n_L} \sum_{k=1}^{n_L} \widehat{\Psi}_{i,j}(z_k^L, \mathbf{x}_k^L);$$

that is, we average the empirical basis functions $\widehat{\Psi}_{i,j}(z_k, \mathbf{x}_k) = \phi_i(z_k)\widehat{\psi}_j(\mathbf{x}_k)$ over the labeled data. We define the new series estimator Series$_{CS}$ by

$$(17) \qquad \widehat{f}(z|\mathbf{x}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \widehat{\alpha}_{i,j} \widehat{\Psi}_{i,j}(z, \mathbf{x}),$$

where the parameters I, J, and ε are chosen so as to minimize the loss in equation (9) relative to the *unlabeled* data P_U.

5.4. *Combined estimator* (*Comb$_{CS}$*). Finally, we present a procedure for combining, or *stacking*, multiple estimators in a principled way. Suppose that $\widehat{f}_1(z|\mathbf{x}), \ldots, \widehat{f}_p(z|\mathbf{x})$ are different estimators of $f(z|\mathbf{x})$; these estimators could, for example, be any of the cross-validated estimators described in Sections 5.1–5.3. Now ask the question: Can we average these models so as to reduce the prediction performance of individual estimators? If we restrict ourselves to weighted averages, then the answer is to compute

$$\widehat{f}^{\boldsymbol{\alpha}}(z|\mathbf{x}) = \sum_{k=1}^{p} \alpha_k \widehat{f}_k(z|\mathbf{x}),$$

where the weights minimize the empirical loss $\widehat{L}(\widehat{f}^{\boldsymbol{\alpha}}, f)$ in equation (9) under the constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{p} \alpha_i = 1$. The weights $\boldsymbol{\alpha} = [\alpha_i]_{i=1}^{p}$ can then be found by solving a standard quadratic programming problem:

$$(18) \qquad \underset{\boldsymbol{\alpha}:\alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1}{\arg\min} \quad \boldsymbol{\alpha}' \mathbb{B} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}' b,$$

where $\mathbb{B}$ is the $p \times p$ matrix $[\frac{1}{\widetilde{n}_U} \sum_{k=1}^{\widetilde{n}_U} \int \widehat{f}_i(z|\widetilde{\mathbf{x}}_k^U) \widehat{f}_j(z|\widetilde{\mathbf{x}}_k^U) \, dz]_{i,j=1}^{p}$ and b is the vector $[\frac{1}{\widetilde{n}_L} \sum_{k=1}^{\widetilde{n}_L} \widehat{f}_i(\widetilde{z}_k^L|\widetilde{\mathbf{x}}_k^L)\widehat{\beta}(\widetilde{\mathbf{x}}_k^L)]_{i=1}^{p}$.

In this work, *Comb$_{CS}$* denotes the estimator that combines the two estimators *ker-NN$_{CS}$* and Series$_{CS}$, although this procedure of combining models applies more generally to other conditional density estimators.

5.5. *Experiments*. Using the simulated and observed data (Section 2), we will now compare the performance of seven different estimators of $f(z|\mathbf{x})$. (As before, our covariates are the four colors and the r-band magnitude in the model magnitude system. We split our data into training, validation, and test sets as in Section 4.) The first three estimators *do not account for selection bias*. They are as follows:

- *NN*: the nearest neighbor estimator from Section 5.1 with $\widehat{\beta}(\mathbf{x}) \equiv 1$ for all $\mathbf{x}$;
- *ker-NN*: the kernel nearest neighbor estimator from Section 5.2 with $\widehat{\beta}(\mathbf{x}) \equiv 1$ for all $\mathbf{x}$; and
- *Series*: the spectral series estimator from Izbicki and Lee (2016).

The tuning parameters of these three estimators minimize the empirical loss in equation (7) on the (labeled) validation data. The second three estimators *correct for covariate shift by importance weights*. They are as follows:

- NN_{CS}: the nearest neighbor estimator from Section 5.1;
- $ker\text{-}NN_{CS}$: the kernel nearest neighbor estimator from Section 5.2; and
- $Series_{CS}$: the spectral series estimator from Section 5.3.

Finally, we have the following:

- $Comb_{CS}$: an estimator that combines $ker\text{-}NN_{CS}$ and $Series_{CS}$ according to Section 5.4.

We choose the tuning parameters of these last four estimators so as to minimize the reweighted empirical loss $\widehat{L}(\widehat{f}, f)$ in equation (9) on (labeled and unlabeled) validation data. By bootstrap, we estimate the standard error of $\widehat{L}(\widehat{f}, f)$ according to $\sqrt{\mathbb{V}[\widehat{L}(\widehat{f}, f)]} \approx \sqrt{B^{-1} \sum_{b=1}^{B} (\widehat{L}_b(\widehat{f}, f) - \overline{\widehat{L}(\widehat{f}, f)})^2}$, where $B = 500$ is the number of bootstrap samples of the test set, $\widehat{L}_b(\widehat{f}, f)$ is the estimated loss for the bth bootstrap sample, and $\overline{\widehat{L}(\widehat{f}, f)}$ is the mean of $\{\widehat{L}_b(\widehat{f}, f)\}_{b=1}^{B}$. Note that each bootstrap sample consists of a sample with a replacement from the labeled set and a sample with replacement from the unlabeled set.

Figure 7 shows the empirical loss [equation (10)] on the test set for the simulated samples with known covariate shifts. Figure 8(a) shows the estimated loss [equation (9)] for 15,000 SDSS samples with no preprocessing, and Figure 8(b) shows the loss after resampling training examples with estimated weights $\widehat{\beta}(\mathbf{x}^L) = 0$ so as to increase the effective sample size; note that the scales in these two plots differ.
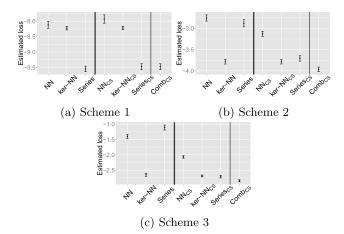

(a) Scheme 1   (b) Scheme 2

(c) Scheme 3

FIG. 7.   *Estimated loss of different density estimators for Schemes 1–3 with, respectively, (a) no, (b) moderate, and (c) large covariate shift. Bars correspond to mean plus and minus standard error. ker-NN is robust to selection bias and has a much smaller loss than previously proposed NN estimators (NN and NN$_{CS}$) even without importance weights; see text for discussion.*
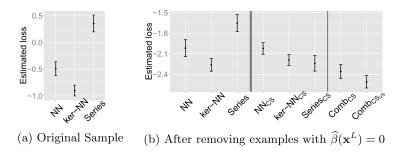
(a) Original Sample   (b) After removing examples with $\widehat{\beta}(\mathbf{x}^L) = 0$

FIG. 8.   *Estimated loss of conditional density estimators for SDSS data* (a) *in the original sample with no preprocessing, and* (b) *after resampling labeled training examples for which the initial weight estimates* $\widehat{\beta}(\mathbf{x}^L) = 0$ *(see Section 4 for details). Note that the scales in the two plots differ.*

Variable selection.   As in Section 4, one can further improve these results by choosing a subset of the ten covariates from the `model` and `cmodel` magnitude systems. Table 2 lists the covariates from a forward stepwise model search with the combined estimator, initialized by an estimate of the marginal distribution $f(z)$. For the SDSS data, the loss of the combined estimator with variable selection [$Comb_{CS_{VS}}$ in Figure 8(b)] is -2.51 ± 0.09, which is smaller than -2.36 ± 0.10, the loss achieved by the combined estimator ($Comb_{CS}$) based on a fixed set of five `model` covariates.

Summary.   Our main conclusions are as follows:

(i) A necessary condition for importance weighting is that $\mathbb{P}_L$ dominates $\mathbb{P}_U$. Our results (Figure 8) show that one should *also* restrict the labeled examples to regions where there is unlabeled data; a simple procedure is to search for labeled examples with $\widehat{\beta}(\mathbf{x}^L) = 0$ and replace these data with labeled examples with $\widehat{\beta}(\mathbf{x}^L) \neq 0$.

(ii) Our kernel-based estimators *ker-NN* and *ker-NN$_{CS}$* consistently perform better than their histogram counterparts *NN* and *NN$_{CS}$* introduced by Cunha et al. (2009).

TABLE 2
Selected covariates for conditional density estimation with $Comb_{CS_{VS}}$

Data set	model					cmodel				
	r	*u − g*	*g − r*	*r − i*	*i − z*	*r*	*u − g*	*g − r*	*r − i*	*i − z*
Scheme 1	X	X	X	X	X			X	X	X
Scheme 2	X		X	X		X				
Scheme 3	X		X	X		X				
SDSS	X		X	X		X				

(iii) Interestingly, *ker-NN* is robust to selection bias even without explicitly incorporating importance weights into the estimator. Here is an intuitive explanation: If the neighborhood $N(\mathbf{x}_0)$ around $\mathbf{x}_0$ is sufficiently small, then the covariate shift assumption [equation (2)] implies that $f_L[z|\mathbf{x} \in N(\mathbf{x}_0)] \approx f_U[z|\mathbf{x} \in N(\mathbf{x}_0)]$. As a result, *ker-NN* returns good density estimates even without correcting for selection bias. *NN* is less robust than *ker-NN* because smoothing via binning requires larger neighborhoods (over which the above approximation may not hold) than smoothing via kernels. For example, in Scheme 3, $N = 35$ for *NN*, versus 8 for *ker-NN*.

(iv) *Series* is sensitive to selection bias, but its covariate shift-corrected analogue $Series_{CS}$, which we introduce in this work, is one of the best estimators for conditional density estimation. When there is *no* selection bias [Figure 7(a)], spectral series (*Series* and $Series_{CS}$) perform significantly better than nearest neighbors methods (*ker-NN* and $ker\text{-}NN_{CS}$); this result is consistent with earlier work by Izbicki and Lee (2016) on *Series*. In settings *with* covariate shift [Figures 7(b)–(c) and Figure 8(b)], $Series_{CS}$ and $ker\text{-}NN_{CS}$ are comparable. We conjecture that $Series_{CS}$ may lose some of its competitive edge when $\mathbb{P}_L \neq \mathbb{P}_U$ because we compute the eigenvectors in the series using only labeled data, and then extrapolate to regions of the unlabeled data via the Nyström extension [equation (16)].

(v) However, by **combining** $ker\text{-}NN_{CS}$ and $Series_{CS}$ as in Section 5.4, we automatically get "the best of both worlds" under a variety of different settings. (The combined estimator $Comb_{CS}$ assigns weights $\alpha = 0.96, 0.50$, and 0.43 to $Series_{CS}$ in Schemes 1–3, and $\alpha = 0.53$ for the SDSS data.)

(vi) Finally, we can further improve our predictions by **variable selection** according to Sections 4 and 5.5 when estimating $\beta(\mathbf{x})$ and $f(z|\mathbf{x})$, respectively. Selection bias leads to smaller models with fewer variables (see Table 2), which is consistent with covariate shift decreasing the effective sample size [Gretton et al. (2010), Shimodaira (2000)], and thereby increasing the variance of the estimators.

Algorithm 1 summarizes the combined model with variable selection ($Comb_{CS_{VS}}$). This model includes two main steps, where we (i) first estimate the importance weights via β-NN with variable selection (lines 1–9), and (ii) then estimate the conditional density with the combined estimator with variable selection (lines 10–23) that combines $Series_{CS}$ (lines 11–14) and $ker\text{-}NN_{CS}$ (lines 16–20).

In Appendix A, we describe different diagnostic tests that can be used to more closely assess the quality of different models. Figure 9, for example, shows quantile-quantile (Q–Q) plots of the SDSS data for NN_{CS} and $Comb_{CS_{VS}}$. These plots tell us how well these density estimates actually fit the observed data.

6. Application to galaxy–galaxy lensing. By working with a probability distribution of the photometric redshift instead of a single best estimate, one can reduce systematic biases in cosmological analyses [Mandelbaum et al. (2008), Sheldon et al. (2012), Wittman (2009)]. In this section, we study the galaxy–galaxy weak-lensing application from Sheldon et al. (2012) and offer new insights on the

Algorithm 1 Redshift estimation under covariate shift. Returns the combined model with variable selection. To emphasize the dependence on tuning parameters and covariates, we use subindices for each estimator; for example, $\widehat{\beta}_{M,S}$ denotes the NN importance weight estimator with M nearest neighbors and covariates S

Input: Labeled training data $(\mathbf{x}_1^L, z_1^L), \ldots, (\mathbf{x}_{n_L}^L, z_{n_L}^L)$; labeled validation data $(\tilde{\mathbf{x}}_1^L, \tilde{z}_1^L), \ldots,$ $(\tilde{\mathbf{x}}_{\tilde{n}_L}^L, \tilde{z}_{\tilde{n}_L}^L)$; unlabeled training data $\mathbf{x}_1^U, \ldots, \mathbf{x}_{n_U}^U$; unlabeled validation data $\tilde{\mathbf{x}}_1^U, \ldots, \tilde{\mathbf{x}}_{\tilde{n}_U}^U$; grid over tuning parameters $\varepsilon_1, \varepsilon_2, I, J, M, N$.

1: **for all** M **do** $\triangleright$ Tune NN Importance Weights Estimator using all covariates
2:     Fit $\widehat{\beta}_M$ using all covariates $\triangleright$ Eq. (20)
3:     Estimate $L(\widehat{\beta}_M, \beta)$ $\triangleright$ Eq. (5)
4: **end for**
5: Define $M^* = \arg\min_M \widehat{L}(\widehat{\beta}_M, \beta)$ and $\widehat{\beta} := \widehat{\beta}_{M^*}$.
6: Replace labeled examples where $\widehat{\beta}(\mathbf{x}^L) = 0$ with additional spectroscopic data.

                          $\triangleright$ Tune NN Importance Weights Estimator with new labeled sample
7: **for all** $S \subset \{x_1, \ldots, x_d\}$ **do** $\triangleright$ Alternatively, one may use a stepwise approach
8:     **for all** M **do** $\triangleright$ Tune NN Importance Weights Estimator using covariates S
9:          Fit $\widehat{\beta}_{M,S}$ $\triangleright$ Eq. (20)
10:         Estimate $L(\widehat{\beta}_{M,S}, \beta)$ $\triangleright$ Eq. (5)
11:     **end for**
12:     Define $M^* = \arg\min_M \widehat{L}(\widehat{\beta}_{M,S}, \beta)$.
13:     Let $\widehat{\beta}_S := \widehat{\beta}_{M^*,S}$.
14: **end for**
15: Let $\widehat{\beta} := \widehat{\beta}_{S^*}$, where $S^* = \arg\min_S L(\widehat{\beta}_S, \beta)$

16: **for all** $S \subset \{x_1, \ldots, x_d\}$ **do** $\triangleright$ Alternatively, one may use a stepwise approach
17:     **for all** I, J, ε_1 **do** $\triangleright$ Tune Spectral Series Estimator using covariates S
18:         Fit $\widehat{f}_{I,J,\varepsilon_1,S}^{\text{series}}$ $\triangleright$ Eq. (17)
19:         Estimate $L(\widehat{f}_{I,J,\varepsilon_1,S}^{\text{series}}, f)$ $\triangleright$ Eq. (9)
20:     **end for**
21:     Let $(I^*, J^*, \varepsilon_1^*) = \arg\min_{I,J,\varepsilon_1} L(\widehat{f}_{I,J,\varepsilon_1,S}^{\text{series}}, f)$

22:     **for all** N, ε_2 **do** $\triangleright$ Tune Kernel Nearest-Neighbor Estimator using covariates S
23:         Fit $\widehat{f}_{N,\varepsilon_2,S}^{\text{ker-NN}}$ $\triangleright$ Eq. (12)
24:         Estimate $L(\widehat{f}_{N,\varepsilon_2,S}^{\text{ker-NN}}, f)$ $\triangleright$ Eq. (9)
25:     **end for**
26:     Let $(N^*, \varepsilon_2^*) = \arg\min_{N,\varepsilon_2} L(\widehat{f}_{N,\varepsilon_2,S}^{\text{ker-NN}}, f)$

27:     Find $\widehat{f}_S$ by combining $\widehat{f}_{I^*,J^*,\varepsilon_1^*,S}^{\text{series}}$ and $\widehat{f}_{N^*,\varepsilon_2^*,S}$ $\triangleright$ Section 5.4
28:     Estimate $L(\widehat{f}_S, f)$ $\triangleright$ Eq. (9)
29: **end for**

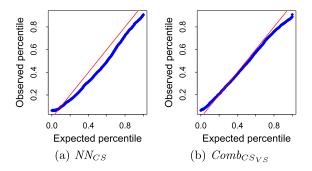30: Output $\widehat{f}_{S^*}$, where $S^* = \arg\min_S L(\widehat{f}_S, f)$

FIG. 9.   *Quantile-quantile plots for* (a) NN_{CS} *and* (b) $Comb_{CS_{VS}}$, *the combined model with selected covariates.*

proper use of redshift distributions in downstream analysis, that is, when estimating *functions* $g(z)$ of an unknown redshift z.

Weak gravitational lensing is the slight deflection of photons from distant astronomical sources that occurs when they pass near massive "lenses" (e.g., galaxies or galaxy clusters) lying closer to us. Lensing acts to magnify the sizes of the distant sources as well as to distort their shapes. Cosmologists can use it to directly probe the distribution of dark matter, a form of matter that does not interact with light (hence the moniker "dark") and which comprises $\approx 27\%$ of the mass-energy of the Universe. The *critical surface density* $\Sigma(z_l, z_s)$ determines the lensing strength of a given lens-source pair [Mandelbaum et al. (2008)]. The goal in this example is to estimate $g_l(z_s) = \Sigma^{-1}(z_l, z_s)$ for a source galaxy with *unknown* redshift z_s, assuming that the redshift z_l of lens l is known. A naive approach is to evaluate the function $g_l(\cdot)$ at a point estimate of the source galaxy redshift z_s, that is, to compute $g_l(\widehat{z}_s) = \Sigma^{-1}(z_l, \widehat{z}_s)$ where $\widehat{z}_s$ typically is an estimate of the regression $\mathbb{E}[Z|\mathbf{x}]$. However, because $\mathbb{E}[g_l(Z)|\mathbf{x}] \neq g_l(\mathbb{E}[Z|\mathbf{x}])$, the estimator

$$(19) \qquad \widehat{\Sigma^{-1}}(z_l, z_s) := \int \Sigma^{-1}(z_l, z) \widehat{f}(z|\mathbf{x}) \, dz$$

usually yields better results if $\widehat{f}(z|\mathbf{x})$ is a good estimate of $f(z|\mathbf{x})$ [Sheldon et al. (2012)]. Note that even though we ultimately are interested in $g_l(z)$, there are clear advantages in estimating the photo-z density $f(z|\mathbf{x})$ well—rather than just aiming for the best regression of $g_l(Z)$ on the photometric covariates $\mathbf{x}$. For example, the data set in this example includes $\approx 500,000$ lenses that each has a different function g_l, and future data sets will only increase in size. With a good density estimator $\widehat{f}(z|\mathbf{x})$, one can address several different inference problems with different g_l's simultaneously as well as construct reliable predictive intervals for each function $g_l(Z)$ for new observations of Z.

Following Sheldon et al. (2012), we use data from the DEEP2 EGS Region [Weiner et al. (2005)]. In addition to the conditional density estimators in Section 5, we implement a nearest neighbors density estimator with 7 neighbors [the

value handpicked by Sheldon et al. (2012) for this particular application], and *photoz*, which computes $\Sigma^{-1}(z_l, \widehat{z}_s)$ where $\widehat{z}_s$ is the nearest neighbor *regression* estimate of z_s. To assess the performance of these methods in galaxy–galaxy lensing, we use the two measures in Mandelbaum et al. (2008) called the lensing calibration bias and the variance ratio. Small values of bias and large values of variance ratio indicate good performance. We use 500 samples for training, 500 for validation, and 382 for testing.

Figure 10 shows the results. The nearest neighbor density estimator with 7 neighbors, *NN-7*, has a smaller calibration bias than *NN* with 27 neighbors, the value chosen via the technique described in the paper [see plot (a)]. However, 7 neighbors leads to poor density estimates [see plot (e)]. *Series* is the only estimator that returns both *accurate parameter estimates* for different values of z_l and lenses l [plots (a) and (b)], as well as *accurate photo-z density estimates* [plots (c) and (f)]. The other estimators do not have both of these properties simultaneously: depending on the chosen tuning parameter, either they have small bias but bad coverage (as *NN-7*) or good coverage but high bias (as *NN*) (see Appendix A for details on how the coverage is computed).

With conditional density estimates, we can also construct *predictive regions* for the redshift. The bottom panel in Figure 10 shows the 95% Highest Predictive Density regions (HPD; Appendix A) of the redshift for *Series* and *NN*. The former HPD region is typically more informative; the size of the HPD's from *Series* are on average $\approx 75\%$ the size of those from *NN*.

Finally, we note that estimating $f(z|\mathbf{x})$ by NN and then computing $\int g(z) \times \widehat{f}(z|\mathbf{x}) \, dz$ is essentially equivalent to estimating $\mathbb{E}[g(Z)|\mathbf{x}]$ directly by performing nearest neighbors *regression* of $g(Z)$ on $\mathbf{x}$. In other words, for this particular application, *NN-7* yields a good estimate of the regression of $g(Z)$ on $\mathbf{x}$, but not (as previously assumed) a reasonable estimate of photo-z.

7. Conclusions. Over the past 15 years, the number of applications for estimating photometric redshifts has grown rapidly, and today there exist a large number of techniques (or codes) for estimating redshifts; see, for example, Dahlen et al. (2013) and references therein. With next-generation surveys, we also expect to have access to additional data [e.g., surface brightness or sizes of galaxies, Lima et al. (2008), or other magnitudes such as grizYJHKs, Oyaizu et al. (2008)], which could potentially improve current photo-z estimates. The value of our work is that it provides a principled framework for properly tuning and assessing different estimators, as well as methods for selecting covariates and for combining two or more estimators for optimal performance. In this paper, we also compared some new and existing estimators of importance weights and photometric redshift under different settings. We found that the nearest neighbors estimator from Cunha et al. (2009) is very effective for estimating importance weights, even when compared to state-of-the-art approaches for density ratio estimation from the machine learning
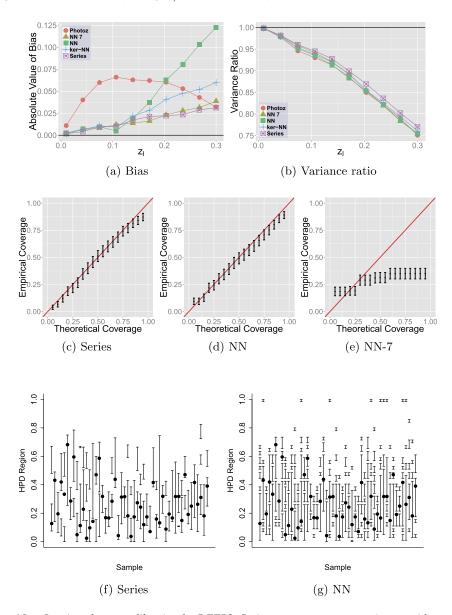
(a) Bias

(b) Variance ratio

(c) Series

(d) NN

(e) NN-7

(f) Series

(g) NN

FIG. 10. *Lensing photo-z calibration for DEEP2: Series returns parameter estimates with smaller biases and variances than the other approaches [see plots (a) and (b)]. Using 7 neighbors for nearest neighbors as in* Sheldon et al. *(2012) yields similar performance in terms of bias; however, the value 7 was handpicked for this specific task and results in poor density estimates [plot (e)]. Series also yields more informative predictive intervals for redshifts than NN [plots (f) and (g)].*

literature. We introduced two new nonparametric conditional density estimators, *kernel nearest neighbors* (Section 5.2) and *Series* (Section 5.3), that both have better performance than the photo-*z* prediction method by Cunha et al. (2009).

Finally, although the scope of this paper is conditional density estimation, one can directly apply the proposed methods to the regression of functions of the photometric redshift, as the regression $\mathbb{E}[g(Z)|\mathbf{x}] = \int g(z) f(z|\mathbf{x}) \, d\mathbf{x}$. The galaxy–galaxy lensing example in Section 6 clearly illustrates the importance of properly tuning and assessing photo-z density estimators in downstream cosmological analysis. In the example, our estimator yielded more accurate predictive regions for the redshift z, as well as better estimates of $g(z)$ for a range of different functions g, although we did not explicitly take g into account in the optimization. We believe our proposed techniques will be valuable for astronomers and cosmologists carrying out next-generation surveys, where accurate photo-z estimates will be needed within a range of different applications (and functions g).

APPENDIX A: DIAGNOSTIC TESTS FOR CONDITIONAL DENSITY ESTIMATION

The L^2 loss only conveys limited information on how well the final density estimates actually fit the observed data. Below we describe three diagnostic so-called *goodness-of-fit* tests that one can use to more closely assess the quality of different models; similar tests can be found in the time series literature [see, e.g., Corradi and Swanson (2006)]. Let $\widehat{F}_{z|\mathbf{x}_i}$ denote the estimated conditional cumulative distribution function for z given $\mathbf{x}_i$. Then:

(i) (*Q–Q Plot*) For every c in a grid of values on $[0, 1]$ and for every observation i in the test spectroscopic sample, compute $Q_i^c = \widehat{F}_{z|\mathbf{x}_i}^{-1}(c)$. Define $\widehat{c} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\beta}(\mathbf{x}_i^L) \mathbb{I}(z_i^L \leq Q_i^c)$. We plot the values of $\widehat{c}$ against the corresponding values of c. If the distributions $\widehat{F}_{z|\mathbf{x}}$ and $F_{z|\mathbf{x}}$ are similar, then the points in the Q–Q plot will approximately lie on the line $\widehat{c} = c$.

(ii) (*P-value*) For every test data point i, let $U_i = \widehat{F}_{z|\mathbf{x}_i}(Z_i)$. If the data are really distributed according to $\widehat{F}_{z|\mathbf{x}}$, then $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$. Hence, we compute the p-value for a Kolmogorov–Smirnoff test that compares the distributions of these statistics to the uniform distribution.

(iii) (*Coverage Plot and HPD Regions*) For every α in a grid of values in $[0, 1]$ and for every spectroscopic data point i in the test sample, let A_i be a set such that $\int_{A_i} \widehat{f}(z|\mathbf{x}_i^L) \, dz = \alpha$. Here we choose the set A_i with the smallest area: $A_i = \{z : f(z|\mathbf{x}_i^L) > t\}$ where t is such that $\int_{A_i} \widehat{f}(z|\mathbf{x}_i^L) \, dz = \alpha$ (i.e., A_i is a Highest Predictive Density region; HPD). Define $\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\beta}(x_i^L) \mathbb{I}(z_i^L \in A_i)$. If the distributions $\widehat{F}_{z|\mathbf{x}}$ and $F_{z|\mathbf{x}}$ are similar, then $\widehat{\alpha} \approx \alpha$. Hence, we plot a graph of $\widehat{\alpha}$'s versus α's and assess how close they are to the line $\widehat{\alpha} = \alpha$. For each α, we also include a 95% confidence interval based on a normal approximation to the binomial distribution.

APPENDIX B: ESTIMATING IMPORTANCE WEIGHTS

Some common approaches to estimating density ratios include direct basis expansions of the form $\widehat{\beta}(\mathbf{x}) = \sum_{i=1}^{I} \widehat{\alpha}_i \lambda_i(\mathbf{x})$ [Izbicki, Lee and Schafer (2014), Kanamori, Hido and Sugiyama (2009), Sugiyama et al. (2008)], kernel mean matching [KMM; Gretton et al. (2010)], and various machine learning techniques [see, e.g., Bickel, Brückner and Scheffer (2009) and Margolis (2011) for a review]. Astronomers have also themselves explored nearest neighbor-based techniques for reweighting nonrepresentative training samples. In particular, Lima et al. (2008) and Cunha et al. (2009) have had success in photo-z estimation with the nearest neighbor estimator

$$(20) \qquad \widehat{\beta}(\mathbf{x}) = \frac{1}{M} \frac{n_L}{n_U} \sum_{k=1}^{n_U} \mathbb{I}(\mathbf{x}_k^U \in V_{\mathbf{x}}^M),$$

where

$$V_{\mathbf{x}}^M = \{\mathbf{y} \in \mathbb{R}^d : d(\mathbf{y}, \mathbf{x}) \le d(\mathbf{x}_{(M)}^L, \mathbf{x})\}$$

denotes the region of feature space with points that are closer to $\mathbf{x}$ than $\mathbf{x}_{(M)}^L$, the Mth nearest neighbor of $\mathbf{x}$ among labeled data, is to $\mathbf{x}$. This estimator has also been used in the machine learning literature [Loog (2012)], although only for the case $M = 1$.

Model selection and tuning of parameters.   To select the best method for estimating $\beta(\mathbf{x})$, we need to specify an appropriate loss function. Our ultimate goal is good photo-z prediction for new *unlabeled* data. If we use importance weighting according to equation (3), then we need good estimates of $\beta(\mathbf{x})$ *at the labeled points* or, more generally, in regions where the density of labeled points is large. Hence, we define the loss function according to equation (4), which is weighted with respect to $\mathbb{P}_L$. [These weights are later used in equation (9) to estimate the $\mathbb{P}_U$-weighted loss in equation (8) for CDE.]

SUPPLEMENTARY MATERIAL

Supplement to "Photo-z estimation: An example of nonparametric conditional density estimation under selection bias" (DOI: 10.1214/16-AOAS1013SUPP; .zip). We provide the data and code used in the paper as supplementary material.

REFERENCES

AIHARA, H. et al. (2011). The eighth data release of the Sloan Digital Sky Survey: First data from SDSS-III. *Astrophys. J.*, *Suppl. Ser.* **193** 29.

BALL, N. M. and BRUNNER, R. J. (2010). Data mining and machine learning in astronomy. *Internat. J. Modern Phys. D* **19** 1049–1106.

BICKEL, S., BRÜCKNER, M. and SCHEFFER, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.* **10** 2137–2155. MR2550104

CORRADI, V. and SWANSON, N. R. (2006). Predictive density evaluation. In *Handbook of Economic Forecasting*. North-Holland, Amsterdam.

CUNHA, C. E., LIMA, M., OYAIZU, H., FRIEMAN, J. and LIN, H. (2009). Estimating the redshift distribution of photometric galaxy samples—II. Applications and tests of a new method. *Mon. Not. R. Astron. Soc.* **396** 2379–2398.

DAHLEN, T., MOBASHER, B., FABER, S. M., FERGUSON, H. C., BARRO, G., FINKEL-STEIN, S. L., FINLATOR, K., FONTANA, A., GRUETZBAUCH, R., JOHNSON, S. et al. (2013). A critical assessment of photometric redshift methods: A CANDELS investigation. *Astrophys. J.* **775** 93.

FERNÁNDEZ-SOTO, A., LANZETTA, K. M. and YAHIL, A. (1998). A new catalog of photometric redshifts in the Hubble Deep Field. *Astrophys. J.* **513** 34–50.

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLKOPF, B. (2010). Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning* (J. Quionero-Candela, M. Sugiyama, A. Schwaighofer and N. D. Lawrence, eds.) Chapter 8. MIT Press, Cambridge, MA.

HALL, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.* **15** 1491–1519. MR0913570

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, Berlin.

IZBICKI, R. and LEE, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *J. Comput. Graph. Statist.* **25** 1297–1316. MR3572041

IZBICKI, R., LEE, A. B and FREEMAN, P. E (2017). Supplement to "Photo-z estimation: An example of nonparametric conditional density estimation under selection bias." DOI:10.1214/16-AOAS1013SUPP.

IZBICKI, R., LEE, A. B. and SCHAFER, C. M. (2014). High-dimensional density ratio estimation with extensions to approximate likelihood computation. *J. Mach. Learn. Res.* **33**.

KANAMORI, T., HIDO, S. and SUGIYAMA, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* **10** 1391–1445. MR2534866

KANAMORI, T., SUZUKI, T. and SUGIYAMA, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Mach. Learn.* **86** 335–367. MR2897527

KIND, M. C. and BRUNNER, R. J. (2013). Tpz: Photometric redshift pdfs and ancillary information by using prediction trees and random forests. *Mon. Not. R. Astron. Soc.* **432** 1483–1501.

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KREMER, J., GIESEKE, F., PEDERSEN, K. S. and IGEL, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astron. Comput.*

LEE, A. B. and IZBICKI, R. (2016). A spectral series approach to high-dimensional nonparametric regression. *Electron. J. Stat.* **10** 423–463. MR3466189

LIMA, M., CUNHA, C. E., OYAIZU, H., FRIEMAN, J., LIN, H. and SHELDON, E. (2008). Estimating the redshift distribution of photometric galaxy samples. *Mon. Not. R. Astron. Soc.* **390** 118–130.

LOOG, M. (2012). Nearest neighbor-based importance weighting. In *IEEE International Workshop on Machine Learning for Signal Processing*.

MANDELBAUM, R., SELJAK, U., HIRATA, C. M., BARDELLI, S., BOLZONELLA, M., BON-
    GIORNO, A., CAROLLO, M., CONTINI, T., CUNHA, C. E., GARILLI, B., IOVINO, A., KAM-
    PCZYK, P., KNEIB, J. P., KNOBEL, C., KOO, D. C., LAMAREILLE, F., LE FEVRE, O.,
    LEBORGNE, J. F., LILLY, S. J., MAIER, C., MAINIERI, V., MIGNOLI, M., NEWMAN, J. A.,
    OESCH, P. A., PEREZ-MONTERO, E., RICCIARDELLI, E., SCODEGGIO, M., SILVERMAN, J.
    and TASCA, L. (2008). Precision photometric redshift calibration for galaxy–galaxy weak lens-
    ing. *Mon. Not. R. Astron. Soc.* **386** 781–806.
MARGOLIS, A. (2011). A literature review of domain adaptation with unlabeled data. Available at
    http://ssli.ee.washington.edu/~amargoli/review_Mar23.pdf.
MINH, H. Q., NIYOGI, P. and YAO, Y. (2006). Mercer's theorem, feature maps, and smoothing. In
    Learning Theory, 19*th Annual Conference on Learning Theory*.
MORENO-TORRES, J. G., RAEDER, T., ALAÍZ-RODRÍGUEZ, R., CHAWLA, N. V. and HER-
    RERA, F. (2012). A unifying view on dataset shift in classification. *Pattern Recogn.* **45** 521–530.
OYAIZU, H., LIMA, M., CUNHA, C. E., LIN, H. and FRIEMAN, J. (2008). Photometric redshift
    error estimators. *Astrophys. J.* **689** 709–720.
QUIONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009).
    Dataset Shift in Machine Learning. MIT Press, Cambridge, MA.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.
SHELDON, E. S., CUNHA, C. E., MANDELBAUM, R., BRINKMANN, J. and WEAVER, B. A.
    (2012). Photometric redshift probability distributions for galaxies in the SDSS DR8. *Astrophys.
    J.*, *Suppl. Ser.* **201** 32.
SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the
    log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598
SPRINGEL, V., FRENK, C. S. and WHITE, S. D. M. (2006). The large-scale structure of the universe.
    Nature **440** 1137–1144.
SUGIYAMA, M., SUZUKI, T., NAKAJIMA, S., KASHIMA, H., VON BÜNAU, P. and KAWANABE, M.
    (2008). Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.* **60**
    699–746. MR2453568
WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York.
WEINER, B. J., PHILLIPS, A. C., FABER, S. M., WILLMER, C. N. A., VOGT, N. P. et al. (2005).
    The DEEP groth strip galaxy redshift survey. III. Redshift catalog and properties of galaxies.
    Astrophys. J. **620** 595.
WITTMAN, D. (2009). What lies beneath: Using $p(z)$ to reduce systematic photometric redshift
    errors. *Astrophys. J. Lett.* **700**.
YORK, D. G., ADELMAN, J., ANDERSON, J. E. JR., ANDERSON, S. F., ANNIS, J., BAH-
    CALL, N. A., BAKKEN, J. A., BARKHOUSER, R., BASTIAN, S., BERMAN, E. et al. (2000).
    The Sloan Digital Sky Survey: Technical summary. *Astrophys. J.* **120** 1579.
ZHAO, L. C. and LIU, Z. J. (1985). Strong consistency of the kernel estimators of conditional density
    function. *Acta Math. Appl. Sin. Engl. Ser.* **1** 314–318. MR0867902
ZHENG, H. and ZHANG, Y. (2012). Review of techniques for photometric redshift estimation. In
    Software and Cyberinfrastructure for Astronomy II **8451**.

R. IZBICKI
DEPARTAMENTO DE ESTATÍSTICA
FEDERAL UNIVERSITY OF SÃO CARLOS
RODOVIA WASHINGTON LUÍS, KM 235-SP-310
SÃO CARLOS, SÃO PAULO
BRAZIL
E-MAIL: rizbicki@ufscar.br

A. B. LEE
P. E. FREEMAN
DEPARTMENT OF STATISTICS
BAKER HALL
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: annlee@stat.cmu.edu
        pfreeman@cmu.edu