University of California
Santa Barbara

Some Contributions to Smoothing Spline Density Estimation and Inference

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Jian Shi

Committee in charge:

Professor Yuedong Wang, Chair
Professor Wendy Meiring
Professor Andrew Carter

December 2017

The Dissertation of Jian Shi is approved.

Professor Wendy Meiring

Professor Andrew Carter

Professor Yuedong Wang, Committee Chair

December 2017

Some Contributions to Smoothing Spline Density Estimation and Inference

by

Jian Shi

To my family

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Yuedong Wang, for introducing me to the area of smoothing splines and for his encouragement, invaluable guidance, patient advise and enthusiastic help in all phases of the preparation of this dissertation. Working with him, I learned a great deal through participation in collaborative research and became motivated in methodological research. I would also like to thank Professor Anna Liu, from Department of Mathematics and Statistics, University of Massachusetts, Amherst, for her remote advise and many valuable discussions during the whole process of this research.

Thanks also go to my committee: Professor Wendy Meiring and Professor Andrew Carter, for serving as members of my committees, their reading of this dissertation and many helpful comments and suggestions. Thank Qiyang Qiu, Changwei Xu, Runfei Luo, Ling Zhu, Yuqi Chen, Danqing Xu, Jingyi and Ruimeng Hu for their accompany for my five years life in Santa Barbara, not only on academic aspect, but also in mental and spirit.

Finally I would like to thank my mother and my husband whose support made the endeavor possible.

Curriculum Vitæ

Jian Shi

Education

2017	Ph.D. in Statistics and Applied Probability (Expected), University of California, Santa Barbara.
2012	B.S. in Mathematics with an emphasis on Mathematical Statistics, Nankai University, Tianjin, China.

Experience

2012-2017	Teaching Assistant, Department of Statistics and Applied Probability, University of California, Santa Barbara.
2016	Research Assistant, Collective Health, Fresno.
2015	Data Scientist Summer Intern, WalmartLabs, San Bruno.

Presentation

JSM 2017 Meeting	Spline Density Estimation with Model-Based Penalties.

Abstract

Some Contributions to Smoothing Spline Density Estimation and Inference

by

Jian Shi

Density estimation plays a fundamental role in many areas including statistics and machine learning. The estimated density functions are useful for model building and diagnostics, inference, prediction, classification and clustering. The goal of our research is to develop new methods for density estimation and inference. This dissertation consists of three projects involving smoothing spline density estimation and inference.

In the first project, we apply smoothing spline density estimation method to test for the normality under both univariate (Chapter 2) and multivariate (Chapter 3) settings. Using the fact that the null hypothesis is equivalent to the logistic density function belonging to the null space of a quintic spline, we construct new test statistics based on quintic polynomial spline and thin-plate spline estimates of the density function. We compare these new tests with some existing normality tests using simulations.

In the second project, we propose model-based penalties for smoothing spline density estimation and inference. These model-based penalties incorporate indefinite prior knowledge that the density is close to, but not necessarily in a family of distributions. The Pearson and generalization of the generalized inverse Gaussian families are used to illustrate the derivation of penalties and reproducing kernels. We also propose new inference procedures to test the hypothesis that the density belongs to a specific family of distributions.

Maximum likelihood estimation within a parametric family and nonparametric estimation are two traditional approaches for density estimation. Often in practice it is

desirable to model some components of the density function parametrically while leaving other components unspecified. In the third project, we study a general semiparametric density model, which contains many existing semiparametric density models as special cases. We develop computational procedures for different cases, and study the theoretical properties including consistency and asymptotic distribution for the semiparametric linear case. Extensive simulations show that the proposed computational methods perform well and the semiparametric model can outperform many existing nonparametric and semiparametric density estimation methods. Real data applications are also provided.

Contents

Chapter 1

Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables from a density function f on a domain $\mathcal{X}$. Density estimation is a procedure to estimate the underlying probability density function f based on observed samples $X_1, \ldots, X_n$. This is a fundamental problem in statistics and machine learning as the density function is useful in many areas including model building and diagnostics, inference, prediction, clustering, and classification. There are three approaches to density estimation: parametric, nonparametric and semiparametric. The parametric approach assumes the density function is known except for a finite number of parameters, and the parameters are estimated based on the observed data. The parametric approach is usually simple but often the form of density is hard to specify. The parametric approach will not be explored in this dissertation. Details can be found in Pearson [1], Pearson [2], Kendall et al. [3], and Fisher [4]. The nonparametric approach does not assume any apriori form of the density model. The form of the density is entirely determined by the data. Silverman [5] and Lzenmman et al. [6] summarized many nonparametric methods. Two nonparametric methods, kernel and smoothing spline density estimation, will be reviewed in Sections 1.1 and 1.2 respectively. The semiparametric approach aggregates both the parametric and

the nonparametric components in one model. It combines flexibility of the nonparametric approach and interpretability of the parametric approach. Some existing semiparametric density estimation methods will be reviewed in Section 1.3. We will develop new estimation and inference procedures for density functions. Some relevant existing density inference procedures will be reviewed in Section 1.4.

1.1 Kernel Density Estimation

When $\mathcal{X}$ is a continuous interval on $\mathbb{R}$, the kernel estimator with kernel K is defined by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.1}$$

where h is the window width, the kernel function K is everywhere nonnegative and satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{1.2}$$

In other words, K is a known probability density function. The kernel estimator can be treated as a sum of "bumps" placed at the observations. The kernel function K determines the shape of the bumps while the window width h determines their widths. The limit as h tends to zero is (in a sense) a sum of Dirac delta functions that spike at the observations. As h becomes large, all details, spurious or otherwise, are obscured. With regard to the selection of an optimal value of the band width h, two methods are commonly used: least square cross validation and likelihood cross-validation. Details can be found in Rudemo ([7]), Bowman ([8]), Hall ([9]) and Stone ([10]).

Despite the vast amount of literature on kernel method, there are still contentious issues regarding the implementation and practical performance. For example, it lacks local adaptability if we choose an h too large, but an h too small often results in being sensitive

to outliers and the presence of spurious bumps. In addition, most kernel estimators suffer from boundary effects – a phenomenon due to the fact that most kernels do not take specific knowledge about the domain of the data into account.

1.2 Smoothing Spline Density Estimation

Smoothing spline density estimation is an adaptive method through penalized likelihood. Two intrinsic constraints that a probability density must satisfy are the non-negativity constraint that $f \geq 0$ and the unity constraint that $\int_{\mathcal{X}} f dx = 1$. In order to eliminate the two intrinsic constraints, several transformations such as $\sqrt{f(x)}$ or $\log(f(x))$ have been explored. Assuming that $f > 0$ on $\mathcal{X}$, Gu [11] considered a logistic density transform $f = e^{\eta} / \int_{\mathcal{X}} e^{\eta} dx$. To make the transform one-to-one, a side condition $A\eta = 0$ is imposed, where A is an averaging operator on $\mathcal{X}$ that averages out the argument x to return a constant function. The estimate of η can then be obtained by minimizing the negative penalized likelihood functional,

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i) + \log\int_{\chi} e^{\eta} dx + \frac{\lambda}{2}J(\eta), \tag{1.3}$$

in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, in which the roughness penalty $J(\eta)$ is a square (semi) norm and

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i) + \log\int_{\chi} e^{\eta} dx \tag{1.4}$$

is the minus log likelihood. One may calculate the minimizer η_{λ}^{*} of (1.3) in a (data-adaptive) finite-dimensional space

$$\mathcal{H}^{*} = \mathcal{H}_0 \oplus span\{R_J(Z_j, \cdot), j = 1, ..., q\}, \tag{1.5}$$

where $\{Z_j, j = 1, \ldots, q\}$ is a random subset of $\{X_i, i = 1, \ldots, n\}$, the null space $\mathcal{H}_0$=$\{\eta$: $A\eta = 0, J(\eta) = 0\}$ and R_J is the reproducing kernel of the orthogonal complement space of $\mathcal{H}_0$. It can be shown that the estimate of η in $\mathcal{H}$ and η^* share the same asymptotic convergence rates with $q \equiv n^{2/(pr+1)+\epsilon}$ for some $r > 1$, $p \in [1, 2]$, and any $\epsilon > 0$. Therefore, in the rest of the dissertation, we focus on η^*_λ, but drop the star for simplicity. The estimate can be represented as

$$\eta_\lambda(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_J(Z_j, x) = \boldsymbol{\phi}^T \boldsymbol{d} + \boldsymbol{\xi}^T \boldsymbol{c}, \tag{1.6}$$

where $\boldsymbol{\xi} = (R_J(Z_1, x), ..., R_J(Z_q, x))^T$, $\boldsymbol{\phi} = (\phi_1(x), ..., \phi_m(x))^T$ is a vector of basis functions of $\mathcal{H}_0$, $\mathbf{c} = (c_1, ..., c_q)^T$ and $\mathbf{d} = (d_1, ..., d_m)^T$. Then the calculation of η_λ reduces to the minimization of

$$-\frac{1}{n}\mathbf{1}^{\boldsymbol{T}}(S\boldsymbol{d} + R\boldsymbol{c}) + \log \int_\chi \exp(\boldsymbol{\phi}^T \boldsymbol{d} + \boldsymbol{\xi}^T \boldsymbol{c}) dx + \frac{\lambda}{2} \boldsymbol{c}^T Q \boldsymbol{c} \tag{1.7}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is a $n \times m$ matrix with the (i, ν)th entry $\phi_\nu(X_i)$, R is a $n \times q$ matrix with the (i, j)th entry $\xi_j(X_i) = R_J(Z_j, X_i)$, and Q is a $q \times q$ matrix with the (j, k)th entry $R_J(Z_j, Z_k)$. Newton method is applied to obtain $\mathbf{c}$ and $\mathbf{d}$ (Gu [11]). Write $\mu_f(g) = \int g e^f dx / \int e^f dx, V_f(g, h) = \mu_f(gh) - \mu_f(g)\mu_f(h)$, and $V_f(g) = V_f(g, g)$. Let $\tilde{\eta}_\lambda = \boldsymbol{\phi}^T \tilde{\boldsymbol{d}} + \boldsymbol{\xi}^T \tilde{\boldsymbol{c}} \in \mathcal{H}^*$ be the point in the previous step, where $\tilde{\boldsymbol{c}} = (\tilde{c}_1, .., \tilde{c}_q)^T$ and $\tilde{\boldsymbol{d}} = (\tilde{d}_1, ..., \tilde{d}_m)^T$. Taking derivatives at $\tilde{\eta}_\lambda$ with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$, the Newton updating equation, after rearranging terms, becomes

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{1}/n - \mu_\phi + V_{\phi,\tilde{\eta}} \\ R^T \mathbf{1}/n - \mu_\xi + V_{\xi,\tilde{\eta}} \end{pmatrix}, \tag{1.8}$$

where $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \eta) = (V_{\tilde{\eta}}(\phi_1, \eta), ..., V_{\tilde{\eta}}(\phi_m, \eta))^T$ and $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \eta) = (V_{\tilde{\eta}}(\xi_1, \eta), ..., V_{\tilde{\eta}}(\xi_q, \eta))^T$.

The smoothing parameter λ is crucial to the performance of the estimation. Gu [11] has developed a data-driven selection method based on the Kullback-Leibler (KL) distance. To measure the discrepancy between the estimate $f_\lambda = e^{\eta_\lambda}/\int_\chi e^{\eta_\lambda} dx$ and the true density $f = e^{\eta}/\int_\chi e^{\eta} dx$, consider the KL distance

$$\mathrm{KL}(\eta, \eta_\lambda) = \mathrm{E}_f[\log(f/f_\lambda)] = \mu_\eta(\eta - \eta_\lambda) - \log \int_\chi e^{\eta} dx + \log \int_\chi e^{\eta_\lambda} dx. \tag{1.9}$$

Dropping terms in $\mathrm{KL}(\eta, \eta_\lambda)$ that do not involve η_λ, one has the relative Kullback-Leibler distance

$$\mathrm{RKL}(\eta, \eta_\lambda) = \log \int_\chi e^{\eta_\lambda} dx - \mu_\eta(\eta_\lambda). \tag{1.10}$$

The first term of (1.10) is readily computable, but the second term, $\mu_\eta(\eta_\lambda)$, involves the unknown density and will have to be estimated. Standard cross-validation suggests an estimate $\tilde{\mu}_\eta(\eta_\lambda) = n^{-1} \sum_{i=1}^n \eta_\lambda^{[i]}(X_i)$, where $\eta_\lambda^{[i]}$ minimizes the delete-one version of (1.3),

$$-\frac{1}{n-1} \sum_{j \neq i} \eta(X_j) + \log \int_\chi e^{\eta} dx + \frac{\lambda}{2} J(\eta). \tag{1.11}$$

Note that X_i does not contribute to $\eta_\lambda^{[i]}$. The delete-one estimates $\eta_\lambda^{[i]}$ are not analytically available, so it is too expensive to compute $\tilde{\mu}_\eta(\eta_\lambda)$ directly. For $g_1, g_2 \in \mathcal{H}$ and α real, define $L_{g_1,g_2}(\alpha) = \log \int_\chi e^{g_1 + \alpha g_2} dx$ as a function of α. Setting $g_1 = \tilde{\eta}, g_2 = \eta - \tilde{\eta}, \alpha = 1$, one has the Taylor expansion

$$\log \int_\chi e^{\eta} dx = L_{\tilde{\eta}, \eta - \tilde{\eta}}(1) \approx L_{\tilde{\eta}, \eta - \tilde{\eta}}(0) + \mu_{\tilde{\eta}}(\eta - \tilde{\eta}) + \frac{1}{2} V_{\tilde{\eta}}(\eta - \tilde{\eta}). \tag{1.12}$$

Substituting the right-hand side of (1.12) for the term $\log \int_\chi e^\eta dx$ in (1.3) and dropping terms that do not involve η, one obtains the quadratic approximation of (1.3) at $\tilde{\eta}$

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i)+\mu_{\tilde{\eta}}(\eta)-V_{\tilde{\eta}}(\tilde{\eta},\eta)+\frac{1}{2}V_{\tilde{\eta}}(\eta)+\frac{\lambda}{2}J(\eta). \tag{1.13}$$

The delete one version of (1.13) is

$$-\frac{1}{n-1}\sum_{j\neq i}\eta(X_j)+\mu_{\tilde{\eta}}(\eta)-V_{\tilde{\eta}}(\tilde{\eta},\eta)+\frac{1}{2}V_{\tilde{\eta}}(\eta)+\frac{\lambda}{2}J(\eta). \tag{1.14}$$

Set $\tilde{\eta}=\eta_\lambda$ and write $\breve{\boldsymbol{\xi}}=(\boldsymbol{\phi}^T,\boldsymbol{\xi}^T)^T$, $H=V_{\tilde{\eta}}(\breve{\boldsymbol{\xi}},\breve{\boldsymbol{\xi}}^T)+\mathrm{diag}(\mathbf{0},\lambda Q)$, $\breve{R}^T=(\breve{\boldsymbol{\xi}}(X_1),...,\breve{\boldsymbol{\xi}}(X_n))=(S,R)^T$. This leads to a cross-validation estimate of $\mu_\eta(\eta_\lambda)$,

$$\hat{\mu}_\eta(\eta_\lambda)=-\frac{1}{n}\sum_{i=1}^{n}\eta_\lambda(X_i)-\frac{tr(P_1^\perp\breve{R}H^{-1}\breve{R}^TP_1^\perp)}{n(n-1)}, \tag{1.15}$$

where $P_1^\perp=I-\mathbf{1}\mathbf{1}^T/n$, I is the identity matrix and $\mathbf{1}$ is a vector of all ones. Then the corresponding estimate of the relative Kullback-Leibler distance is

$$V(\lambda)=-\frac{1}{n}\sum_{i=1}^{n}\eta_\lambda(X_i)+\log\int_\chi e^{\eta_\lambda}dx+\alpha\frac{tr(P_1^\perp\breve{R}H^{-1}\breve{R}^TP_1^\perp)}{n(n-1)}, \tag{1.16}$$

where $\alpha=1$ is"unbiased" for the minimization of Kullback-Leibler loss but may yield severe undersmoothing, whereas a larger α yields smoother estimates. In the simulations throughout this ariticle, we use $\alpha=1.4$ as suggested by Gu [12].

1.3 Semiparametric Density Estimation

Often in practice, it is desirable to model some components of the density function parametrically while leaving other components unspecified. Several semiparamet-

ric density models have been proposed for different purposes. Olkin et al [13] proposed to fit a combination of a parametric and a nonparametric density functions, $g(x,\pi) = \pi f_1(x,\boldsymbol{\theta}) + (1-\pi)f_2(x)$, where $f_1(x,\boldsymbol{\theta})$ is known up to parameters $\boldsymbol{\theta}$, $f_2(x)$ is a nonparametric density function, and $\pi \in [0,1]$ is an unknown weight to be estimated from the data. They showed that the semiparametric density estimate provides a compromise between the parametric and nonparametric versions, and the semiparametric model converges to the true density at the same rate as the traditional maximum likelihood estimator when the parametric model holds, and at the same rate as kernel estimators when the nonparametric model does not hold. Hjort et al. [14] proposed an density estimation procedure by starting out with a parametric density estimate $f(x,\hat{\boldsymbol{\theta}})$, and then multiply with a nonparametric kernel type estimate of a correction function $r(x) = f(x)/f(x,\hat{\boldsymbol{\theta}})$, producing $\hat{f}(x) = f(x,\hat{\boldsymbol{\theta}})\hat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)f(x,\hat{\boldsymbol{\theta}})/f(X_i,\hat{\boldsymbol{\theta}})$. Hjort et al [14] showed that their semiparametric density model can perform better than a nonparametric fit when the true density is in the neighborhood of the initial parametric density. Efron et al.[15] proposed a specially designed exponential family for density estimation, $g_\beta(x) = g_0(x)\exp(\beta_0 + \boldsymbol{t}(x)\boldsymbol{\beta}_1)$, where $g_0(x)$ is a carrier density and estimated by kernel density estimation, $\boldsymbol{t}(x)$ is a p-dimensional vector of sufficient statistics, $\boldsymbol{\beta}_1$ is a p-dimensional vector of parameters and β_0 is a normalizing parameter making $g_\beta(x)$ integrate to 1 over $\mathcal{X}$. The proposed method matches the estimated expectation of $\boldsymbol{t}(x)$ with sample expectation of $\boldsymbol{t}(y)$. For example, setting $\boldsymbol{t}(x) = (x, x^2)$, the method matches the first two moments between the estimation and sample. They also use the exponential family model to investigate density differences in multisample situations, with shared carrier $g_0(x)$ estimated nonparametrically, but with possibly different values of the exponential parameters β_0 and $\boldsymbol{\beta}_1$. Lenk [16] proposed a flexible semiparametric model for Bayesian testing of $f(x|\boldsymbol{\beta},Z) = \exp[\boldsymbol{h}(x)'\boldsymbol{\beta} + Z(x)]/\int_{\mathcal{Y}}\exp[\boldsymbol{h}(x)'\boldsymbol{\beta} + Z(x)]dG(x)$, where $\boldsymbol{h}(x) = [h_1(x),...,h_m(x)]'$ is a vector of m nonconstant functions, Z is a zero mean,

second-order Gaussian process with bounded, continuous covariance function, and G is a known dominating measure on the support $\mathcal{X}$. The choice of $\boldsymbol{h}$ and G are based on theoretical or scientific considerations. The semiparametric model allows the predictive distribution to deviate from the parametric family. If the parametric family is inadequate, the semiparametric predictive density coherently adapts to the data. Yang [17] also used the logistic tranformation of density function as Lenk [16]. But he treated $Z(x)$ as an unknown smooth function defined on $\mathcal{X}$.

The semiparametric density models considered in Chapter 5 contain most existing semiparametric density models discussed above as special cases.

1.4 Inference on Density Functions

Denote F as the cumulative distribution function (CDF) of $X_1, ..., X_n$. We consider the null hypothesis $H_0 : F(x) = F_0(x)$ where $F_0(x)$ is a known CDF. In this section we review six existing tests: Kolmogorov-Smirnov (KS) test, Lilliefors (Lillie) test, Cramer-von Mises (CVM) test, Anderson-Darling (AD) test, Shapiro-Wilk (Shapiro) test, Shapiro-Francia (SF) test, and Pearson chi-square (Pearson) test.

The Kolmogorov-Smirnov test statistic (Kolmogorov [18]) is defined as

$$D_n = \sup_x |F_n(x) - F_0(x)|, \tag{1.17}$$

where $x \in \mathbb{R}$, F_n is the empirical distribution function (EDF), and $F_0(x)$ is the theoretical CDF under H_0. When $F_0(x)$ contains unknown parameters, Lilliefors [19] extended

Kolmogorov-Smirnov test,

$$D_n = \sup |F_n(x) - F_0(x, \hat{\theta})|, \tag{1.18}$$

where $\hat{\theta}$ is the MLE of θ.

Cramer-von Mises and Anderson-Darling test statistics are special cases of following quadratic form

$$Q = \int (F_n(x) - F_0(x))^2 \psi(x) dF_0(x), \tag{1.19}$$

where $\psi(x)$ is a suitable weight function. The CVM statistic uses the weight function $\psi(x) = 1$, while the AD statistic uses the weight function $\psi(x) = (F_0(x)(1 - F_0(x)))^{-1}$. Compared with the CVM test, the AD test places more weight on observations in the tails of the distribution.

Shapiro-Wilk and Shapiro-Francia tests are tests for normality where $F_0(x)$ is a normal distribution with unknown mean and variance. They have the same form, differing only in the definition of the coefficients. Let $X_{(1)}, ..., X_{(n)}$ be the order statistics of iid random variables sampled from the standard normal distribution, $\boldsymbol{m} = (m_1, ..., m_n)'$ be the expected values of the order statistics and V be the covariance matrix of these order statistics. The test statistic

$$W = \frac{(\sum a_i X_{(i)})^2}{\sum (X_i - \bar{X})^2}, \tag{1.20}$$

where $\bar{X} = \sum_{i=1}^n X_i/n$ is the sample mean, and the constants $\boldsymbol{a} = (a_1, ..., a_n)$ are given by $\boldsymbol{a} = (\boldsymbol{m}'V^{-1})/(\boldsymbol{m}'V^{-1}V^{-1}\boldsymbol{m})^{1/2}$ for the Shapiro-Wilk test and $a_i = m_i/(\sum_{j=1}^n m_j^2)^{1/2}$ for the Shapiro-Francia test. For normality test, sample mean and sample variance will

be used as estimates of parameters. Tables for the null distributions have been computed by Monte Carlo methods.

Pearson test divides the data into k bins and the statistic is defined as (Pearson [20])

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i, \tag{1.21}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i under H_0. A bin can be a combination of levels when X is a categorical variable, or an interval when X is a continuous variable. When X is a continuous variable and the ith bin is an interval $(a, b]$, then

$$E_i = n(F_0(b) - F_0(a)). \tag{1.22}$$

The test statistic approximately follows a chi-square distribution with $k - c$ degrees of freedom where k is the number of non-empty cells and c is the number of unknown parameters to be estimated if there are any. An attractive feature of the Pearson chi-square test is that it can be applied to each of continuous, categorical and binned data. The disadvantage is that it requires a sufficient sample size in order for the chi-square approximation to be reasonably accurate. In addition, the test is sensitive to the choice of bins.

Other parametric (Dey et al. [21]), nonparametric (Ying el al. [22], Fan [23], Rubio et al. [24], Cai et al. [25]) and semiparametric (Li et al. [26], De Wet [27]) methods have been developed to test density function for independent and correlated data.

1.5 Thesis Summary

The goal of our research is to develop new methods for density estimation and inference. This dissertation consists of three projects involving smoothing spline density estimation and inference. In Chapters 2 and 3, we will review related literatures on the univariate and multivariate normality tests. Using the fact that the null hypothesis is equivalent to the logistic density function belonging to the null space of a quintic spline, we develop new tests based on smoothing spline density estimation, and compare them with some existing normality tests using simulations. Chapter 4 proposes model-based penalties for smoothing spline density estimation and inference. These model-based penalties incorporate indefinite prior knowledge that the density is close to, but not necessarily in a family of distributions. A general semiparametric density model, which contains many existing semiparametric density models as special cases, is studied in Chapter 5. The computational procedures and theoretical properties are proposed and discussed. The Derivations of reproducing kernels for some L-splines are given in the Appendix.

Chapter 2

Univariate Normality Test Using Smoothing Spline Density Estimation

2.1 Introduction

Normal distribution is the most commonly used probability distribution. Many distributions in nature can be well approximated by a normal distribution. The well-known Central Limit Theorem says that if a random variable X is the sum of a large number of iid random variables, then X will be approximately normally distributed. It explains why the normal random variable appears in so many diverse applications. Normal distribution has many good properties. For example, normal distribution is closed under convolution and linear transformations. The conjugate prior of the mean of a normal distribution is another normal distribution. Many other broadly used distributions, such as binomial, Poisson, chi-squared, Student t, Rayleigh, Logistic, Log-normal, Hyper-geometric, are related to the normal distribution.

Normality tests are used to determine if a data set can be well-modeled by a normal distribution. Specifically we are interested in the hypothesis $H_0 : X_i$ for $i = 1, \ldots, n$ are from a normal distribution. This is a special case of the general hypothesis discussed in Section 1.4 with F_0 being a normal CDF with unknown mean and variance. In Section 1.4 we have reviewed six existing tests: Kolmogorov-Smirnov (KS), Lilliefors (Lillie), Cramer-von Mises (CVM), Anderson-Darling (AD), Shapiro-Wilk (Shapiro), Shapiro-Francia (SF), and Pearson chi-square (Pearson). Six new test statistics based on smoothing spline density estimation will be developed in this chapter.

Two spline models are used to estimate the density function, the polynomial spline and thin plate spline (TPS). We provide a brief review of the polynomial spline and TPS models for density estimation in Section 2.2. Six new test statistics are introduced in Section 2.3. Section 2.4 presents the simulations results.

2.2 Polynomial Spline and Thin Plate Splines

2.2.1 Polynomial Spline

The polynomial spline is the minimizer of

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i) + \log\int_a^b e^{\eta}dx + \frac{\lambda}{2}\int_a^b (\eta^{(m)})^2 dx, \tag{2.1}$$

in the Sobolev space $W_2^m[a, b]$ where

$$W_2^m[a, b] = \{\eta : \eta, \eta', ..., \eta^{(m)} \text{ are absolutely continuous}, \int_a^b (\eta^{(m)})^2 dx < \infty\}.$$

For simplicity, one can set $a = 0$ and $b = 1$. Equipped with appropriate inner products, the Sobolev space is a reproducing kernel Hilbert space. Wang [28] presents two constructions. The first construction defines the inner product

$$\langle \eta, \tilde{\eta} \rangle = \sum_{\nu=0}^{m-1} \eta^{(\nu)}(0)\tilde{\eta}^{(\nu)}(0) + \int_0^1 \eta^{(m)}\tilde{\eta}^{(m)}dx, \text{ for any } \eta, \tilde{\eta} \in W_2^m[0,1]. \tag{2.2}$$

Then $W_2^m[0,1]$ can be decomposed into two orthogonal RKHS's

$$\mathcal{H}_0 = \text{span}\{1, x, ..., x^{m-1}/(m-1)!\}, \tag{2.3}$$

$$\mathcal{H}_1 = \{\eta : \eta^{(\nu)}(0) = 0, \nu = 0, ..., m-1, \int_0^1 (\eta^{(m)})^2 < \infty\}, \tag{2.4}$$

with corresponding RKs

$$R_0(x,z) = \sum_{\nu=1}^{m} \frac{x^{\nu-1}}{(\nu-1)!}\frac{z^{\nu-1}}{(\nu-1)!}, \tag{2.5}$$

$$R_1(x,z) = \int_0^1 \frac{(x-u)_+^{m-1}}{(\nu-1)!}\frac{(z-u)_+^{m-1}}{(m-1)!}du, \tag{2.6}$$

where function $(x)_+ = \max\{x, 0\}$. It is clear that the null space $\mathcal{H}_0$ contains the polynomial of order m, the functions which are not penalized. For identifiability we set $Af \triangleq f(0) = 0$ and, with some abuse of notation, keep using the notation $W_m^2[0,1]$ to represent the Sobolev space under this constraint. When $m = 1, 2$ respectively, the null spaces $\mathcal{H}_0$ are the constant and linear functions which correspond to the uniform and exponential distributions as suggested in Silverman [29]. For the normal estimation, quintic spline ($m = 3$) is a better choice since the logistic density transformation falls in the null space of quintic spline. The estimate of η can then be obtained by minimizing the penalized likelihood functional in (2.1) in the space $W_2^3[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_J$, where

$J(\eta) = \int_0^1 (\eta^{(3)})^2 dx$, $\mathcal{H}_0 = \text{span}\{x, x^2\}$ and

$$\mathcal{H}_J = \{\eta : \eta(0) = \eta'(0) = \eta''(0) = 0, \int_0^1 (\eta^{(3)})^2 dx < \infty\}.$$

The second construction for $W_2^m[0,1]$ defines the inner product

$$\langle \eta, \tilde{\eta} \rangle = \sum_{\nu=0}^{m-1} \left(\int_0^1 \eta^{(\nu)} dx \right) \left(\int_0^1 \tilde{\eta}^{(\nu)} dx \right) + \left(\int_0^1 \eta^{(m)} \tilde{\eta}^{(m)} dx \right), \text{ for any } \eta, \tilde{\eta} \in W_2^m[0,1].$$

With identifiability condition $Af \triangleq \int_0^1 \eta dx = 0$, we have $W_2^m[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_J$, where

$$\begin{aligned} \mathcal{H}_0 &= \text{span}\{k_0(x), k_1(x), \ldots, k_{m-1}(x)\}, \\ \mathcal{H}_J &= \{\eta : \int_0^1 \eta^\nu dx = 0, \nu = 0, \ldots, m-1, \int_0^1 (\eta^{(m)})^2 dx < \infty\} \end{aligned}$$

are RKHS's with corresponding reproducing kernel (RK) functions

$$\begin{aligned} R_0(x,z) &= \sum_{\nu=0}^{m-1} k_\nu(x) k_\nu(z), \\ R_1(x,z) &= k_m(x) k_m(z) + (-1)^{m-1} k_{2m}(|x-z|), \end{aligned}$$

where $k_r(x) = B_r(x)/r!$ are scaled Bernoulli polynomials, and B_r are defined recursively by $B_0(x) = 1, B_r'(x) = rB_{r-1}(x)$ and $\int_0^1 B_r(x)dx = 0$ for $r = 1, 2, \ldots$. With the same data, these two constructions can be expected to yield similar density estimates. In the simulations, we utilize the second construction.

2.2.2 Thin Plate Spline

The Thin Plate Spline (TPS) density estimate is the minimizer of penalized functional

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i)+\log\int_{\chi}e^{\eta}dx+\frac{\lambda}{2}J_m^d(\eta) \tag{2.7}$$

on the d-dimensional domain $\mathcal{X}=\mathbb{R}^d$, where the penalty functional is

$$J_m^d(\eta)=\sum_{\alpha_1+\ldots+\alpha_d=m}\frac{m!}{\alpha_1!\ldots\alpha_d!}\int\ldots\int\left(\frac{\partial^m\eta}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}\right)^2 dx_1\ldots dx_d. \tag{2.8}$$

The corresponding null space is the space spanned by polynomials in d variables of total degree from first up to $m-1$. Denote E_m as the Green function for the m-iterated Laplacian $E_m(x,z)=E(\|x-z\|)$, where $\|x-z\|$ is the Euclidean distance and

$$E(u)=\begin{cases}(-1)^{d/2+1+m}|u|^{2m-d}\log u, & d \text{ even},\\ |u|^{2m-d}, & d \text{ odd}.\end{cases}$$

Although E_m is not the RK of $W_2^m(\mathbb{R}^d)$ since it is not nonnegative definite, it is conditionally nonnegative definite in the sense that $T^T\mathbf{c}=0$ implies that $\mathbf{c}^TK\mathbf{c}\geq 0$, where T is the matrix of null basis functions evaluated at observations, and $K=\{E_m(x_i,x_j\}_{i,j=1}^n$. Referred to as a semi-kernel, the function E_m is sufficient for the purpose of estimation. Note that TPS is defined on the whole Euclidean space while the polynomial spline is defined on a compact interval. Therefore, TPS can cover the whole domain of the normal distribution while a transformation and truncation is needed for the polynomial spline. The details will be discussed later. In this section we are interested in testing the univariate normal distribution, therefore we set $d=1$ and $m=3$.

In Section 2.3, the modified test statistics using the polynomial spline and TPS estimates

are introduced. The simulation results will be discussed in Section 2.4.

2.3 New Test Statistics Based on Smoothing Spline Estimates

2.3.1 Modified Anderson-Darling, Cramer-von Mises and Kolmogorov-Smirnov tests

Denote F_p as the CDF based on a polynomial spline density estimate and F_t as the CDF based on a TPS density estimate. Replacing the EDF in KS, AD and CVM test statistics by smoothing spline estimate, we have the following modified test statistics:

$$\text{KS-P} = \sup_x |F_p(x) - F_0(x)|, \tag{2.9}$$

$$\text{CVM-P} = \int (F_p(x) - F_0(x))^2 dF_0(x), \tag{2.10}$$

$$\text{AD-P} = \int (F_p(x) - F_0(x))^2 (F_0(x)(1 - F_0(x)))^{-1} dF_0(x), \tag{2.11}$$

$$\text{KS-T} = \sup_x |F_t(x) - F_0(x)|, \tag{2.12}$$

$$\text{CVM-T} = \int (F_t(x) - F_0(x))^2 dF_0(x), \tag{2.13}$$

$$\text{AD-T} = \int (F_t(x) - F_0(x))^2 (F_0(x)(1 - F_0(x)))^{-1} dF_0(x). \tag{2.14}$$

The null distributions of all the three test statistics will be approximated by the bootstrap method. Details will be given in Section 2.4.

2.3.2 Likelihood Ratio Test

The likelihood ratio statistic is

$$\text{LRT} = -2\ln\left(\frac{\text{likelihood for the null model}}{\text{likelihood for the alternative model}}\right). \tag{2.15}$$

For parametric models, LRT approximately follows a chi-squared distribution under the null hypothesis. In our case, the density function is estimated nonparametrically with an infinite dimensional model space. Consequently the null distribution is unknown in theory.

The maximized negative log likelihood under H_0 is

$$l_0 = -\frac{n}{2} - \frac{n}{2}\left[\log(2\pi) + \log\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\right]. \tag{2.16}$$

Let $\hat{f}_s$ be the smoothing spline estimate under H_1, which could be the polynomial spline or TPS. The negative log likelihood under H_1 is

$$l_1 = -\sum_{i=1}^{n}\log\hat{f}_s(X_i).$$

The likelihood ratio test is therefore defined as

$$\text{LRT} = 2\left(l_1 - l_0\right).$$

In the simulations, we denote LRT by LRT-P when f_s is the polynomial spline estimate, and by LRT-T when f_s is TPS estimate. Since the null distribution of the LRT statistic is unknown, we will use the bootstrap method to approximate the distribution. The details will be introduced in the simulation section.

2.3.3 Kullback-Leibler Test

The KL distance between two density functions f_1 and f_2 is defined as

$$\mathrm{KL}(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \tag{2.17}$$

Let $\hat{f}_0$ be the normal distribution with the estimated parameters under the null hypothesis, and $\hat{f}_s$ be the smoothing spline estimate of the density function. We will then use the KL distance between $\hat{f}_0$ and $\hat{f}_s$, $\mathrm{KL}(\hat{f}_0, \hat{f}_s)$, as the KL test statistic. KLD-P and KLD-T represent the test statistics computed with $\hat{f}_s$ being the polynomial spline estimate and TPS respectively.

2.3.4 Projection Test

The logistic transformation of normal distribution is

$$\eta(x) = -\frac{x(x - 2\mu)}{2\sigma^2}. \tag{2.18}$$

With the third order of polynomial spline, under the second construction, $W_3^2[0, 1]$ can be decomposed into two orthogonal space

$$\mathcal{H}_0 = \mathrm{span}\{x - .5, .5(x - .5)^2 - 1/24\},$$
$$\mathcal{H}_J = \{\eta : \int_0^1 \eta dx = \int_0^1 \eta' dx = \int_0^1 \eta'' dx = 0, \int_0^1 (\eta^{(3)}(x))^2 dx < \infty\}$$

Under H_0, we have $J(\eta) = \int_0^1 (\eta^{(3)}(x))^2 dx = 0$. Thus, taking advantage of orthogonal property, we define

$$\text{DFN-P} = \boldsymbol{c}^T Q \boldsymbol{c} \tag{2.19}$$

as the test statistic to represent the departure from normality, where $\boldsymbol{c}$ and Q are defined in (1.7). For TPS, DFN-T can be defined similarly. Large DFN provides evidence against the null hypothesis. The null distribution of the test statistics will be derived by the bootstrap method.

2.4 Simulations

In the simulations, we will evaluate the proposed tests AD-P, CVM-P, KS-P, LRT-P, DFN-P, KLD-P, AD-T, CVM-T, KS-T, LRT-T, KLD-T, and DFN-T, and compare them with the existing tests including AD, KS, CVM, Lillie, Shapiro, SF, and Pearson tests. The function *ssden* in the R package *gss* is used to compute polynomial spline and TPS estimates of density functions (Gu [11]). For the TPS, we set *type=list(x=list("tp",m=3))*. For the quintic spline, we need to define the corresponding basis function *mkphi.quintic* and RKs *mkrk.quintic*, and set *type=list("custom", list(nphi=2, mkphi=mkphi.quintic, mkrk=mkrk.quintic, env=c(0,1)))*. The functions *shapiro.test, sf.test, cvm.test, ad.test, lillie.test*, and *pearson.test* in the R package *nortest* and *ks.test* in the R package *dgof* are used to perform Shapiro, SF, CVM, AD, Lillie, Pearson and KS tests. We use the bootstrap method to approximate null distributions for all newly proposed tests, where the number of bootstrap samples is set to be 2000. We generate 100 data replicates for each simulation setting. For each simulated data set, we estimate the unknown parameters with sample mean and sample variance, and generate bootstrap samples from the normal distribution with the estimated parameters.

In the simulations, the null hypothesis is $H_0 : X_1, ..., X_n \sim^{iid}$ Normal Distribution. And we will consider

I) four sample sizes: n=20, 50, 100, 200;

II) three families of distributions: generalized normal, mixed normal, and skewed normal:

(a) The generalized normal distribution (also known as the exponential power distribution) is defined as

$$f(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)}\, e^{-(|x-\mu|/\alpha)^{\beta}}.$$

This is a parametric family of symmetric distributions. It includes the normal ($\beta = 2$) and Laplace distributions ($\beta = 1$), and as limiting cases with $\beta \to \infty$, it includes the continuous uniform distributions on bounded intervals of the real line $(\mu - \alpha, \mu + \alpha)$. This family allows for tails that are either heavier than normal ($\beta < 2$) or lighter than normal ($\beta > 2$). It is a useful way to parametrize a continuum of symmetric densities spanning from the normal ($\beta = 2$) to the uniform density ($\beta = \infty$), and a continuum of symmetric densities spanning from the Laplace ($\beta = 1$) to the normal density ($\beta = 2$). In the simulations, we consider $\mu = 0, \alpha = 1$, four choices of $\beta = 1, 2, 4, 8$, where powers with $\beta = 2$ gives us the probability of type I error.

(b) The mixed normal is a mixture of $a\mathcal{N}(0, 1/4)+(1-a)\mathcal{N}(2, 1)$, where $a \in (0, 1)$. As a approaches to 0, the distribution tends to $\mathcal{N}(2, 1)$. And as a approaches to 1, it tends to $\mathcal{N}(0, 1/4)$. To see the alteration of test power, we consider several choices of $a = 1/2, 1/4, 1/8, 1/16$.

(c) The skewed normal distribution is defined as

$$f(x) = \frac{1}{\omega}\sqrt{\frac{2}{\pi}} e^{-\frac{(x-\xi)^2}{2\omega^2}} \int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)} e^{-\frac{t^2}{2}}\, dt,$$

where ξ is location (real) parameter, ω is scale (positive, real) parameter, and α is shape (real) parameter. As the absolute value of shape parameter α is far from 0, the distribution becomes more skewed from normal distribution. The distribution is right skewed if $\alpha > 0$ and is left skewed if $\alpha < 0$. Note, however, that the skewness of the distribution is limited to the interval $(-1, 1)$. In the simulations, we consider $\xi = 0, \omega = 1$, four choices of $\alpha = 1, 2, 4, 8$.

Since the domain of the polynomial spline estimate is a compact interval $[0, 1]$, while the domain of the normal distribution is the whole real line, we truncate and scale the simulated data by

$$\frac{x_i - a}{b - a}, \tag{2.20}$$

where $[a, b]$ is the empirical domain of the data set which is large enough to cover all observations, so they are fixed for all the simulations. The way we find a and b is to generate 1000 data sets from the distribution of interest, and obtain the minimum of the lower bounds as a and maximum of the upper bounds as b. This is reasonable, as most practical data is valued in a compact interval. The simulated results for three families of distributions are shown below.

Generalized Normal

In Table 2.1, $\beta = 2$ in generalized normal distribution corresponds the normal distribution, whose power provides the probability of type I error. And it indicates the type I error rate is within a reasonable interval around 0.05. However, tests based on TPS

tend to have probability of type I error larger than the nominal value. As the generalized normal distribution deviates farther from normal distribution and sample size increases, all tests have stronger powers. Except for the case when $\beta = 4$, in general the new methods do not perform better then existing ones in this case.

Sample Size	20	50	100	200	20	50	100	200
Distribution	$\beta = 1$				$\beta = 2$			
Shapiro	0.26	0.52	0.81	0.98	0.07	0.10	0.03	0.06
SF	**0.33**	**0.56**	0.84	0.99	0.08	0.09	0.05	0.08
CVM	0.29	0.54	**0.89**	0.98	0.05	0.06	0.08	0.06
AD	0.29	0.54	0.86	**1.00**	0.06	0.08	0.07	0.06
Lillie	0.21	0.40	0.76	0.94	0.03	0.05	0.06	0.06
Pearson	0.13	0.27	0.46	0.77	0.05	0.02	0.05	0.08
DFN-P	0.14	0.26	0.41	0.67	0.08	0.04	0.04	0.06
LRT-P	0.20	0.30	0.45	0.68	0.06	0.04	0.06	0.08
CVM-P	0.20	0.31	0.45	0.68	0.07	0.07	0.05	0.08
KS-P	0.17	0.30	0.45	0.68	0.05	0.05	0.05	0.07
KLD-P	0.16	0.26	0.45	0.68	0.06	0.05	0.06	0.08
DFN-T	0.21	0.48	0.58	0.73	0.02	0.07	0.08	0.05
LRT-T	0.23	0.43	0.58	0.73	0.04	0.08	0.05	0.05
AD-T	0.26	0.45	0.58	0.73	0.06	0.10	0.06	0.06
CVM-T	0.26	0.44	0.58	0.73	0.06	0.10	0.06	0.06
KS-T	0.26	0.45	0.58	0.73	0.06	0.10	0.06	0.06
KLD-T	0.26	0.49	0.68	0.85	0.03	0.06	0.08	0.04
Distribution	$\beta = 4$				$\beta = 8$			
Shapiro	0.05	0.10	0.34	**0.81**	0.14	0.46	0.87	0.99
SF	0.02	0.06	0.21	0.60	0.08	0.24	0.69	0.99
CVM	0.04	0.11	0.27	0.66	**0.16**	0.29	0.64	0.97
AD	0.04	0.12	0.28	0.74	0.15	0.39	0.76	0.99
Lillie	0.05	0.11	0.20	0.47	0.10	0.21	0.47	0.85
Pearson	0.04	0.09	0.15	0.27	0.09	0.11	0.29	0.78
DFN-P	**0.11**	**0.17**	**0.50**	0.70	**0.16**	0.33	0.44	0.78
LRT-P	0.07	0.14	0.44	0.73	0.13	0.35	0.58	0.98
CVM-P	0.07	0.14	0.42	0.73	**0.16**	0.33	0.48	0.91
KS-P	0.07	0.14	0.46	0.73	**0.16**	0.33	0.47	0.89
KLD-P	0.09	0.14	0.42	0.71	0.10	0.24	0.37	0.73
DFN-T	0.00	0.01	0.19	0.52	0.00	0.04	0.32	0.91
LRT-T	0.06	0.13	0.37	0.76	0.15	**0.60**	**0.96**	**1.00**
AD-T	0.03	0.06	0.22	0.63	0.09	0.37	0.73	0.99
CVM-T	0.03	0.06	0.22	0.62	0.11	0.39	0.71	0.99
KS-T	0.03	0.06	0.22	0.62	0.12	0.38	0.67	0.97
KLD-T	0.01	0.01	0.13	0.31	0.03	0.06	0.21	0.59

Table 2.1: Powers of different tests when data are generated from the generalized normal distribution under different sample sizes and different values of shape parameter. Powers under $\beta = 2$ are probability of type I error.

Mixed Normal

In Table 2.2, all tests have stronger power, as the mixed normal distribution deviates farther from normal distribution and sample size increases. Tests including the existing and the newly proposed all can detect the mixed normal distribution very well. When the sample size increases to 100, the powers can reach almost 1. The existing tests such as SF and AD have slightly stronger power in some cases, while our proposed tests have similar powers to other existing methods.

Sample Size	20	50	100	200	20	50	100	200
Distribution	$a = 1/2$				$a = 1/4$			
Shapiro	0.31	0.81	**1.00**	1.00	**0.75**	**1.00**	1.00	1
SF	0.19	0.72	0.97	1.00	0.74	0.99	1.00	1
CVM	0.36	0.86	0.99	1.00	0.73	**1.00**	1.00	1
AD	0.34	**0.88**	0.99	1.00	0.74	**1.00**	1.00	1
Lillie	0.20	0.79	0.97	1.00	0.58	0.98	1.00	1
Pearson	0.21	0.67	0.88	1.00	0.42	0.82	1.00	1
DFN-P	0.40	0.77	0.92	0.99	0.65	0.93	0.99	1
LRT-P	0.30	0.80	0.97	1.00	0.67	0.99	1.00	1
CVM-P	0.34	0.80	0.98	1.00	0.70	**1.00**	1.00	1
KS-P	**0.38**	0.81	0.98	1.00	0.68	1.00	1.00	1
KLD-P	0.25	0.77	0.96	1.00	0.60	0.99	1.00	1
DFN-T	0.02	0.45	0.80	1.00	0.11	0.88	1.00	1
LRT-T	0.26	0.73	0.96	1.00	0.57	0.99	1.00	1
AD-T	0.18	0.69	0.96	1.00	0.66	0.99	1.00	1
CVM-T	0.21	0.71	0.96	1.00	0.70	0.99	1.00	1
KS-T	0.21	0.72	0.96	1.00	0.71	0.99	1.00	1
KLD-T	0.09	0.46	0.75	1.00	0.37	0.90	1.00	1
Distribution	$a = 1/8$				$a = 1/16$			
Shapiro	0.60	**0.98**	1.00	1.00	0.43	0.79	**1.00**	1.00
SF	**0.62**	**0.98**	1.00	1.00	**0.46**	**0.82**	**1.00**	1.00
CVM	0.56	0.86	0.97	1.00	0.36	0.66	0.94	0.99
AD	0.56	0.93	1.00	1.00	0.40	0.71	0.95	1.00
Lillie	0.43	0.80	0.96	1.00	0.35	0.61	0.86	0.99
Pearson	0.32	0.58	0.90	1.00	0.17	0.37	0.62	0.89
DFN-P	0.56	0.81	0.95	0.99	0.38	0.63	0.93	1.00
LRT-P	0.55	0.94	1.00	1.00	0.39	0.75	0.98	1.00
CVM-P	0.58	0.94	1.00	1.00	0.38	0.75	0.98	1.00
KS-P	0.57	0.93	1.00	1.00	0.38	0.71	0.98	1.00
KLD-P	0.52	0.90	1.00	1.00	0.32	0.68	0.98	1.00
DFN-T	0.25	0.91	1.00	1.00	0.31	0.80	0.99	1.00
LRT-T	0.56	0.97	1.00	1.00	0.43	0.77	0.99	1.00
AD-T	**0.62**	**0.98**	1.00	1.00	0.45	0.77	0.99	1.00
CVM-T	0.61	**0.98**	1.00	1.00	0.44	0.77	0.99	1.00
KS-T	0.61	0.97	1.00	1.00	0.41	0.77	0.99	1.00
KLD-T	0.53	0.97	1.00	1.00	0.43	0.80	0.99	1.00

Table 2.2: Powers of different tests when data are generated from the mixed normal distribution under different sample sizes and different weights.

Skewed Normal

In Table 2.3, all tests have similar powers. All tests have little power when the distribution is not quite skewed while the powers are close to one when the distribution becomes more skewed.

Sample Size	20	50	100	200	20	50	100	200
Distribution	$\alpha = 1$				$\alpha = 2$			
Shapiro	0.06	0.03	0.06	0.11	**0.18**	0.20	0.34	0.59
SF	0.05	0.03	0.05	0.12	**0.19**	**0.21**	0.33	0.54
CVM	0.06	0.06	0.05	0.08	0.13	0.17	0.27	0.45
AD	0.07	0.05	0.05	0.08	0.12	0.18	0.27	0.49
Lillie	0.08	0.03	0.06	0.07	0.14	0.11	0.19	0.31
Pearson	0.04	0.04	0.03	0.07	0.05	0.10	0.14	0.21
DFN-P	0.06	**0.07**	0.06	0.12	0.09	0.19	0.27	0.47
LRT-P	0.08	**0.07**	0.07	0.16	0.07	0.17	0.32	0.55
CVM-P	0.08	0.04	0.06	**0.18**	0.09	0.19	0.34	0.56
KS-P	**0.09**	0.06	0.07	**0.18**	0.09	0.18	0.34	0.57
KLD-P	**0.09**	**0.07**	0.08	0.17	0.06	0.17	0.33	0.53
DFN-T	0.05	0.01	**0.11**	0.14	0.08	0.18	0.30	0.55
LRT-T	0.06	0.01	**0.11**	0.12	0.08	0.15	0.31	0.56
AD-T	0.06	0.01	0.09	0.14	0.11	**0.21**	0.34	**0.61**
CVM-T	0.06	0.02	0.09	0.14	0.11	**0.21**	0.34	**0.61**
KS-T	0.07	0.01	0.10	0.13	0.10	0.20	**0.35**	**0.61**
KLD-T	0.05	0.02	0.10	0.11	0.09	0.16	0.23	0.47
Distribution	$\alpha = 4$				$\alpha = 8$			
Shapiro	0.20	**0.54**	0.87	**1.00**	**0.38**	**0.82**	**0.99**	**1.00**
SF	**0.21**	0.50	0.84	**1.00**	0.32	0.77	0.97	**1.00**
CVM	0.19	0.40	0.78	0.99	0.28	0.62	0.91	0.99
AD	**0.21**	0.45	0.86	**1.00**	0.33	0.70	0.95	**1.00**
Lillie	0.16	0.32	0.67	0.90	0.22	0.48	0.84	**1.00**
Pearson	0.10	0.22	0.46	0.78	0.17	0.42	0.73	0.98
DFN-P	0.17	0.46	0.69	0.81	0.34	0.71	0.87	0.99
LRT-P	0.17	0.45	0.80	**1.00**	0.32	0.77	0.98	**1.00**
CVM-P	0.20	0.53	0.82	**1.00**	0.36	0.81	0.98	**1.00**
KS-P	**0.21**	0.50	0.81	**1.00**	0.36	0.80	0.98	**1.00**
KLD-P	0.13	0.42	0.75	0.98	0.25	0.67	0.97	**1.00**
DFN-T	0.09	0.32	0.70	0.97	0.08	0.33	0.81	0.99
LRT-T	0.17	0.51	0.86	**1.00**	0.32	0.78	0.97	**1.00**
AD-T	0.19	0.53	**0.89**	**1.00**	0.31	0.71	0.98	**1.00**
CVM-T	0.20	**0.54**	**0.89**	**1.00**	0.32	0.73	0.98	**1.00**
KS-T	0.20	**0.54**	0.88	**1.00**	0.33	0.73	0.98	**1.00**
KLD-T	0.12	0.32	0.64	0.85	0.16	0.27	0.68	0.98

Table 2.3: Powers of different tests when data are generated from the skewed normal distribution under different sample sizes and different values of shape parameter.

2.5 Conclusion

In this chapter, we apply the polynomial spline and TPS density estimation to construct several new normality test statistics. Extensive simulations show that our new proposed methods have similar power as the existing normal tests in most cases. Generally, the tests with TPS density estimation should be avoided due to large probability of type I error. Comparing all the existing and newly proposed tests, Shapiro test often behaves best in different sample sizes.

Chapter 3

Multivariate Normality Tests Using Smoothing Spline Density Estimation

3.1 Introduction

Many multivariate statistical methods, such as multivariate analysis of variance (MANOVA), linear discriminant analysis (LDA), principal component analysis (PCA), canonical correlation, and graphical modeling are based on multivariate normality (MVN) assumption. Assessing the MVN is crucial for the validity of these methods. Existing methods include Mardia's popular multivariate skewness and kurtosis statistics (Mardia & Kanti [30], Mardia & Kanti [31]), a consistent and invariant test proposed by Henze and Zirkler [32], and Royston's modified Shapiro-Wilk test ([33], [34], [35]). Recently new approaches to testing multivariate normality have been proposed by Székely and Rizzo [36] based on Euclidean distance between sample elements, and by Kellner and Celisse [37] based on Maximum Mean Discrepancy (MMD).

The rest of Section 3.1 reviews some existing MVN test statistics. Section 3.2 presents three new proposed tests based on smoothing spline density estimation. In Section 3.3, the simulation experiments compare the proposed methods with Mardia's MVN, Henze-Zirkler's MVN, Royston's MVN tests in the R package "MVN" (Korkmax et al [38]), and the method in Székely and Rizzo [36].

3.1.1 Existing Multivariate Normality Tests

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be iid samples from a multivariate density function $f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a p-dimensional vector. We are interested in testing the null hypothesis that

$$H_0 : f(\boldsymbol{x}) \text{ is MVN distribution.}$$

Mardia's MVN Test

Mardia [30] proposed a MVN test based on multivariate extensions of skewness ($\hat{\gamma}_{1,p}$) and kurtosis ($\hat{\gamma}_{2,p}$):

$$\hat{\gamma}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3, \quad \hat{\gamma}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} m_i^2,$$

where

$$m_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)' S^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j),$$
$$m_i = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})' S^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}).$$

The test statistic of skewness, $(n/6)\hat{\gamma}_{1,p}$, is approximately χ^2 distributed with $p(p+1)(p+2)/6$ degrees of freedom under H_0. Similarly, the test statistic of kurtosis, $\hat{\gamma}_{2,p}$, is approximately normally distributed with mean $p(p+2)$ and variance $8p(p+2)/n$ under H_0.

Mardia [31] introduced a correction term into the skewness statistic for small sample size $(nk/6)\hat{\gamma}_{1,p}$, where $k = (p+1)(n+1)(n+3)/(n(n+1)(p+1)-6)$. This statistic is also distributed as χ^2 with degrees of freedom $p(p+1)(p+2)/6$ under H_0.

Despite the widespread use of Mardia's statistics, Horswell [39] demonstrated that, generally speaking, MVN tests based on measures of skewness and kurtosis did not distinguish well between 'skewed' and 'non-skewed' distributions. To improve upon power, some authors have attempted to combine measures of skewness and kurtosis into a single 'omnibus' test statistic. Mardia and Foster [40] derived six statistics, including one that uses the Wilson-Hilferty approximation (Wilson and Foster[41]). However, Horswell and Looney [42] found that this statistic lacked power.

Royston MVN Test

Most reviews and comparative studies of tests for MVN refer to the Royston's extension [34] of the powerful Shapiro-Wilk goodness-of-fit test (Shapiro and Wilk [43]) for univariate normality. An algorithm for computing this extension is given in Royston [44] and Royston [45]. Specifically, let W_j be the Shapiro-Wilk/Shapiro-Francia test statistic for the jth variable $(j = 1, 2, ..., p)$ and R_j be the values obtained from the normality transformation

$$R_j = \left\{ \Phi^{-1} \left[\frac{1}{2} \Phi \left(-((1 - W_j)^{\lambda} - \mu) \right) / \sigma \right] \right\}^2,$$

where λ, μ and σ are calculated from polynomial approximations (Royston [33]) and Φ denotes the cumulative distribution function of the standard normal distribution. The Roystons test statistic for multivariate normality

$$\mathrm{H} = \frac{p/[1 + (p-1)\bar{c}] \sum_{j=1}^{p} R_j}{p} \overset{H_0}{\sim} \chi^2_e,$$

where e is the equivalent degrees of freedom, $\bar{c}$ is an estimate of the average correlation among the R_j's.

Henze-Zirkler's MVN Test

The Henze-Zirkler's test (Henze and Zirkler [32]) is based on a non-negative functional distance that measures the distance between two distribution functions. The Henze-Zirkler's multivariate normality test is defined as

$$\text{HZ} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}m_{ij}} - 2(1+\beta^2)^{-\frac{p}{2}}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}m_i} + (1+2\beta^2)^{-\frac{p}{2}},$$

where $\beta = \frac{1}{\sqrt{2}}\left(\frac{n(2p+1)}{4}\right)^{\frac{1}{p+4}}$, m_i gives the squared Mahalanobis distance of ith observation to the centroid $\bar{\boldsymbol{X}} = 1/n\sum_{i=1}^{n}\boldsymbol{X}_i$ and m_{ij} gives the Mahalanobis distance between the ith and the jth observations. Under null hypothesis, the test statistic is approximately log-normally distributed with mean μ and variance σ^2 given below:

$$\mu = 1 - \frac{a^{-\frac{p}{2}}\left(1 + p\beta^{\frac{2}{a}} + p(p+2)\beta^4\right)}{2a^2},$$

$$\sigma^2 = 2(1+4\beta^2)^{-\frac{p}{2}} + \frac{2a^{-p}(1+2p\beta^4)}{a^2} + \frac{3p(p+2)\beta^8}{4a^4} - 4\omega_\beta^{-\frac{p}{2}}\left(1 + \frac{3p\beta^4}{2\omega_\beta} + \frac{p(p+2)\beta^8}{2\omega_\beta^2}\right),$$

where $a = 1+2\beta^2$ and $\omega_\beta = (1+\beta^2)(1+3\beta^2)$. By using the log-normal distribution parameters, μ and σ, we can test the significance of multivariate normality. Among the class of invariant and consistent tests for MVN, Henze and Zirkler's proposal has a fame for its high power over a wide variety of alternative distributions.

Energy MVN Test

The E-test (Energy-test) of multivariate normality was proposed and implemented by Székely and Rizzo [36]. The test statistic for p-variate normality is given by

$$\mathcal{E} = n\left(\frac{2}{n}\sum_{j=1}^{n} E||\boldsymbol{y}_j - Z|| - E||Z - Z'|| - \frac{1}{n^2}\sum_{j,k=1}^{n}||\boldsymbol{y}_j - \boldsymbol{y}_k||\right),$$

where $\boldsymbol{y}_i = \Sigma^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}), i = 1, \ldots, n$, are the standardized sample using the sample mean vector $\boldsymbol{\mu}$ and sample covariance matrix Σ. Z and Z' are iid standard MVN variables, and $||.||$ denotes Euclidean norm. The E-test rejects the null hypothesis for large values of $\mathcal{E}$.

3.1.2 Density Estimation Using Tensor Product of Polynomial Splines

In Section 2.2.1 we have reviewed the univariate polynomial spline. Now consider a multivariate function $\eta(\boldsymbol{x}) = \eta(x_1, \ldots, x_p)$ on a product domain $\mathcal{X} = \prod_{m=1}^{p}\mathcal{X}_m$, where $x_m \in \mathcal{X}_m$ denotes the mth coordinate of $\boldsymbol{x} \in \mathcal{X}$. Let A_m be an averaging operator on $\mathcal{X}_m$. An ANOVA decomposition of η can be defined as

$$\eta = \left\{\prod_{m=1}^{p}(I - A_m + A_m)\right\}\eta = \sum_{\mathcal{S}}\left\{\prod_{m\in\mathcal{S}}(I - A_m)\prod_{m\notin\mathcal{S}}A_m\right\}\eta = \sum_{\mathcal{S}}\eta_{\mathcal{S}}, \tag{3.1}$$

where $\mathcal{S} \subset \{1, ..., p\}$ lists the active terms in $\eta_{\mathcal{S}}$ and the summation is over all of the 2^p subsets of $\{1, ..., p\}$. The term $\eta_\emptyset = \prod A_m\eta$ is a constant, the term $\eta_d = \eta_{\{d\}} = (I - A_d)\prod_{\alpha\neq d}A_\alpha\eta$ is the main effect of x_d, the term $\eta_{d_1,d_2} = f_{\{d_1,d_2\}} = (I - A_{d_1})(I - A_{d_2})\prod_{\alpha\neq d_1,d_2}A_\alpha\eta$ is the interaction of x_{d_1} and x_{d_2}, and so forth. To fit such multivariate data, the tensor product reproducing kernel Hilbert space can be used to incorporate the

ANOVA decomposition. Paralleling with (3.1), the tensor product space $\mathcal{H} = \bigotimes_{m=1}^{p} \mathcal{H}_m$ has a tensor sum decomposition

$$\mathcal{H} = \bigotimes_{m=1}^{p} (\mathcal{H}_{0_m} \oplus \mathcal{H}_{1_m}) \bigoplus_{\mathcal{S}} \left\{ (\bigotimes_{m \in \mathcal{S}} \mathcal{H}_{1_m}) \otimes (\bigotimes_{m \notin \mathcal{S}} \mathcal{H}_{0_m}) \right\} = \bigoplus_{\mathcal{S}} \mathcal{H}_{\mathcal{S}}, \tag{3.2}$$

where $\mathcal{H}_{\mathcal{S}}$ is a RKHS with RK $R_{\mathcal{S}} \propto \prod_{m \in \mathcal{S}} R_{1_m}$, and the projection of $\eta \in \mathcal{H}$ in $\mathcal{H}_{\mathcal{S}}$ is the $\eta_{\mathcal{S}}$ in (3.1). The minimizer of $L(\eta) + \frac{\lambda}{2} J(\eta)$ in a tensor product RKHS is called a tensor product smoothing spline.

For multivariate density estimation, we have $f(\boldsymbol{x}) = e^{\eta(\boldsymbol{x})} / \int e^{\eta(\boldsymbol{x})}$ by logistic transformation. Consider the domain $\mathcal{X} = [0,1]^p$. Multiple-term models can be constructed using the tensor product splines of the above, with an ANOVA decomposition. For example, $p = 2$, then one can have

$$\eta = \eta_0 + \eta_1 + \eta_2 + \eta_{1,2},$$

where terms other than the constant η_0 satisfy certain side conditions. The constant shall be dropped for density estimation to maintain a one-to-one logistic density transform. The additive model implies the independence of the two coordinates. For each coordinate i, the RKHS are $\mathcal{H}_{00_i} \oplus \mathcal{H}_{01_i} \oplus \mathcal{H}_{1_i}$, where $\mathcal{H}_{00_i} = \{\eta : \eta \propto 1\}$, and corresponding RKs are $R_{00_i} = 1, R_{0_i}$ and R_{1_i}. Using this space for both marginal domains, one can construct a tensor product space with tensor sum terms. The subspace $\mathcal{H}_{00_1} \otimes \mathcal{H}_{00_2}$ spans the constant term which will be dropped in density estimation. The subspace $\mathcal{H}_{00_1} \otimes (\mathcal{H}_{01_2} \oplus \mathcal{H}_{1_2})$ and $(\mathcal{H}_{01_1} \oplus \mathcal{H}_{1_2}) \otimes \mathcal{H}_{00_2}$ span the main effects, and the subspace $(\mathcal{H}_{01_1} \oplus \mathcal{H}_{1_1}) \otimes (\mathcal{H}_{01_2} \oplus \mathcal{H}_{1_2})$ spans the interactions. And corresponding RKs are in table 3.1.

Subspace	Reproducing Kernel
$\mathcal{H}_{00_1} \otimes \mathcal{H}_{00_2}$	1
$\mathcal{H}_{00_1} \otimes (\mathcal{H}_{01_2} \oplus \mathcal{H}_{1_2})$	$R_{0_1} + R_{1_1}$
$(\mathcal{H}_{01_1} \oplus \mathcal{H}_{1_2}) \otimes \mathcal{H}_{00_2}$	$R_{0_2} + R_{1_2}$
$(\mathcal{H}_{01_1} \oplus \mathcal{H}_{1_1}) \otimes (\mathcal{H}_{01_2} \oplus \mathcal{H}_{1_2})$	$R_{0_1}R_{0_2} + R_{0_1}R_{1_2} + R_{1_1}R_{0_2} + R_{1_1}R_{1_2}$

Table 3.1: RKs in the tensor product space when $d = 2$.

3.1.3 Density Estimation Using Thin Plate Spline

The thin plate spline (TPS) density estimate is the minimizer of penalized functional

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i) + \log\int_{\chi} e^{\eta}dx + \frac{\lambda}{2}J(\eta) \tag{3.3}$$

on the p-dimensional domain $\mathcal{X} = \mathbb{R}^p$, where the penalty $J(\eta)$ has the form

$$J_m^p(\eta) = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_p!} \times \int\ldots\int \left(\frac{\partial^m \eta}{\partial x_{\langle 1\rangle}^{\alpha_1}\ldots\partial x_{\langle p\rangle}^{\alpha_p}}\right)^2 dx_{(1)}\ldots dx_{(p)}.$$

The null space of $J_m^p(\eta)$ consists of polynomials of first up to $(m-1)$ total order, which is of dimension $M = \binom{p+m-1}{p} - 1$. The quadratic functional $J_m^p(\eta)$ is invariant under a rotation of the coordinates. In the space $\mathcal{H} = \{\eta : J_m^p(\eta) < \infty\}$ with $J_m^p(\eta)$ as a square semi norm, it is necessary that $2m - p > 0$ for the evaluation functional to be continuous. In the MVN case, $m = 3$ and p is the corresponding dimension of MVN distribution, as long as $p < 6$.

In simulation experiments, we apply R package *gss* to implement the density estimation. Since the *"ssden"* function does not support multi-dimension TPS fit yet, tensor product of TPS is applied so that the transformation step in quintic spline estimation can be skipped.

3.2 Test Statistics

In this section, we will propose test statistics based on likelihood ratio and Kullback-Leibler divergence respectively. The hypothesis is

$$H_0 : f \text{ is a multivariate normal distribution,}$$

which is equivalent to $\eta \in \mathcal{H}_0$, where $\mathcal{H}_0 = \text{span}\{x - .5, .5(x - .5)^2 - 1/24, y - .5, .5(y - .5)^2 - 1/24, (x - .5)(y - .5)\}$ and $\boldsymbol{x} = (x, y)$ when $p = 2$.

3.2.1 LRT Test

Likelihood Ratio Tests (LRT) are a powerful, very general method of testing model assumptions. The general LRT is often about a parametrized family of probability density functions or probability mass functions $f(x|\theta)$. In these parameterized cases, as sample size n approaches ∞, the test statistic will be asymptotically χ^2 distributed. In our case, we apply smoothing spline to obtain the density function estimate $\hat{f}_s(x, y)$ under alternative hypothesis. Under null hypothesis, MLE $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ are used to estimate the unknown parameters ,

$$\begin{aligned} l_0(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) &= \prod_{i=1}^{n} \left\{ (2\pi)^{-1} |\hat{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \right] \right\} \\ &= \frac{1}{(2\pi)^n |\hat{\Sigma}|} e^{-n}, \end{aligned}$$

where $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = 1/n \sum_{i=1}^{n} \mathbf{x}_i$ and $\hat{\Sigma} = 1/n \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$. Then we can construct two likelihoods under null and alternative hypothesis, and further obtain the LRT test

statistics

$$\text{LRT} = -2\log(\frac{l_0}{l_1}),$$

where $l_1 = \prod \hat{f}_s(\mathbf{x}_i)$. Since the nonparametric LRT test statistic is no longer following a known distribution, we use traditional parametric bootstrap to calculate p-values. In the simulation, LRT-P denote the LRT when $\hat{f}_s$ is the polynomial spline estimate, and LRT-T denote the LRT based on TPS estimate.

3.2.2 KLD Test

We propose the KLD test statistic, applying KL divergence introduced in (2.17) to measure the difference between the null distribution $\hat{f}_0(x, y|\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ and the smoothing spline estimate $\hat{f}_s(x, y)$ from the observed data,

$$\text{KLD}(f_0, f_s) = \int\int \hat{f}_0(x, y|\hat{\boldsymbol{\mu}}, \hat{\Sigma}) \log \frac{\hat{f}_0(x, y|\hat{\boldsymbol{\mu}}, \hat{\Sigma})}{\hat{f}_s(x, y)} dxdy. \tag{3.4}$$

In the simulation, KLD-P represents KLD test with the polynomial spline estimate, and KLD-T represents the test based on TPS estimate.

3.2.3 Projection Test

For the multivariate normal density estimation, tensor product quintic spline is appropriate since the logistic transformation falls in the null space of the tensor product quintic spline. As discussed in Section 2.2.1, the quintic spline has the functional space $W_2^3[0, 1] = \mathcal{H}_0 \oplus \mathcal{H}_J$, where

$$\mathcal{H}_0 = \text{span}\{1, k_1(x), k_2(x)\}$$
$$\mathcal{H}_J = \{\eta : \int_0^1 \eta^{(\nu)} dx = 0, \nu = 0, \ldots, m-1, \int_0^1 (\eta^{(3)})^2 < \infty\},$$

with corresponding RKs

$$R_0(x,z) = \sum_{\nu=1}^{3} \frac{x^{\nu-1}}{(\nu-1)!}\frac{z^{\nu-1}}{(\nu-1)!}, \tag{3.5}$$

$$R_1(x,z) = \int_0^1 \frac{(x-u)_+^2}{(\nu-1)!}\frac{(z-u)_+^2}{(2)!}du, \tag{3.6}$$

where function $(x)_+ = \max\{x, 0\}$. Therefore, the two dimension tensor product quintic spline has the functional space $(\mathcal{H}_{0_1} \oplus \mathcal{H}_{1_1}) \otimes (\mathcal{H}_{0_2} \oplus \mathcal{H}_{1_2}) \triangleq \mathcal{H}_0 \oplus \mathcal{H}_1$, where

$$\begin{aligned}
\mathcal{H}_0 =& \mathcal{H}_{0_1} \otimes \mathcal{H}_{0_2} \\
=& \text{span}\{x-0.5, 0.5(x-0.5)^2 - 1/24, y-0.5, 0.5(y-0.5)^2 - 1/24, (x-0.5)(y-0.5), \\
& \{0.5(y-0.5)^2 - 1/24\}(x-0.5), \{0.5(x-0.5)^2 - 1/24\}(y-0.5)\}, \\
\mathcal{H}_1 =& (\mathcal{H}_{0_1} \otimes \mathcal{H}_{1_2}) \oplus (\mathcal{H}_{0_2} \otimes \mathcal{H}_{1_1}) \oplus (\mathcal{H}_{1_1} \oplus \mathcal{H}_{2_1}),
\end{aligned}$$

and the corresponding RKs of $\mathcal{H}_1$ are $R_{1,00}(x,y), R_{1,01}(x,y), R_{00,1}(x,y), R_{01,1}(x,y)$ and $R_{1,1}(x,y)$ in Appendix A.1. The tensor product quintic spline density estimation is to minimize the penalized likelihood by

$$\eta(\boldsymbol{x}) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(\boldsymbol{x}) + \sum_{j=1}^{q} c_j R_J(\boldsymbol{Z}_j, \boldsymbol{x}) = \boldsymbol{\phi}^T \boldsymbol{d} + \boldsymbol{\xi}^T \boldsymbol{c},$$

where $\phi_\nu(\boldsymbol{x})$ are the basis of $\mathcal{H}_0$, and

$$R_J(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \begin{pmatrix} R_{1,00}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ R_{1,01}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ R_{00,1}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ R_{01,1}(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ R_{1,1}(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{pmatrix}, \tag{3.7}$$

where the θ's are tunable smoothing parameters. Thus, the norm of second part in (3.2.3) provides the discrepancy of the observed distribution from the null space, i.e. the multivariate normal distribution. Let Q be the $q \times q$ matrix with the (j,k)th entry $R_J(\boldsymbol{Z}_j, \boldsymbol{Z}_k)$. The departure from the multivariate normal is

$$\begin{aligned} \text{DFMVN} &= \boldsymbol{c}^T Q \boldsymbol{c} \\ &= \sum\sum c_i R_J(\boldsymbol{x}_i, \boldsymbol{x}_j) c_j. \end{aligned}$$

3.3 Simulations

3.3.1 Type I Error

The first distribution simulated was the MVN distribution. In this case, the null hypothesis is true, so each test should reject at about the nominal rate of 5%. The performance of the new tests against the MVN distribution is found in Table 3.2. probability of type I error for all tests are close to the nominal value of .05.

Sample Size	LRT-T	KLD-T	LRT-P	KLD-P	DFMVN-P
50	0.06	0.09	0.01	0.03	0.01
100	0.04	0.04	0.08	0.08	0.04
Sample Size	Mardia's (Skew)	Mardia's (Kurtosis)	Henze-Zirkler	Royston	$\hat{\mathcal{E}}$
50	0.08	0.02	0.01	0.07	0.01
100	0.05	0.01	0.05	0.04	0.05

Table 3.2: probability of type I error for different tests.

3.3.2 Power Comparison

In order to assess the performance of the new test LRT-P, LRT-T, KLD-P, KLD-T and DFMVN, and compare them with the existing tests including Mardia's skewness, Mar-

dia's Kurtosis, Henze-Zirkler, Royston and Evergy($\hat{\mathcal{E}}$) tests, we performed a parametric bootstrap power study. The null distribution for all the proposed tests are approximated by 1000 bootstrap samples. We generate 100 data replicates for each simulation setting. For each generated data set, we estimate the mean and covariance by sample mean and sample covariance, and generate bootstrap samples from the normal distribution with the estimated parameters. In the simulations, we consider

1. two sample sizes: n=50, 100

2. higher order polynomial in logit transformation $\eta(x)$ and mixture bivariate normal distribution:

   i Higher orders:

   $$f(\mathbf{x}) \propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+((\mathbf{x}-\mathbf{0.5})'\mathbf{x})^2},$$

   where $\mathbf{x} = (x_1, x_2)^T, x_1, x_2 \in [0, 1]$, $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma}$ is a correlation matrix with $\rho = 0.5$ for all off-diagonal elements. Since the domain of this case is on a compact set, we will not conduct the TPS estimation.

   ii Mixture bivariate normal distributions

   (a) 0.5N((0,0),I)+0.5N((3,3),I)

   (b) 0.79N((0,0),I)+0.21N((3,3),I)

   (c) 0.5N((0,0),Σ_3)+0.5N((0,0),I)

   (d) 0.5N((0,0),Σ_1)+0.5N((3,3),Σ_2)

   (e) 0.9N((0,0),Σ_1)+0.1N((0,0),Σ_2)

   (f) 0.5N((0,0),Σ_1)+0.5N((3,3),I)

   (g) 0.5N((0,0),Σ_2)+0.5N((3,3), I)

   where Σ_1 is a correlation matrix with $\rho = 0.2$ for all off-diagonal elements, Σ_2 is a correlation matrix with $\rho = 0.5$ for all off-diagonal elements, and

Σ_3 denotes 1 on diagonal and 0.9 off diagonal. A 2-d mixed multivariate normal distribution is denoted as $\omega N_2(\mu_1, \sigma_1) + (1-\omega)N_2(\mu_2, \sigma_2)$, where the sampled populations is $N_2(\mu_1, \sigma_1)$ with probability ω, and $N_2(\mu_2, \sigma_2)$ with probability $1-\omega$. As the mixing parameter p and other parameters are varied, the multivariate normal mixtures have a wide variety of types of departures from normality. A 50% normal location mixture is symmetric with light tails, and a 90% normal location mixture is skewed with heavy tails. A normal location mixture with $p = 1 - \frac{1}{2}(1 - \frac{\sqrt{3}}{3}) \approx 0.79$, provides an example of a skewed distribution with normal kurtosis (Henze [46]). The scale mixtures in the comparison are symmetric with heavier tails than normal. Let ω be the mixing parameter, then

- $\star$ $\omega = 0.9$ indicates mild contamination and is skewed and leptokurtic,
- $\star$ $\omega = 0.79$ is moderately contaminated, skewed and mesokurtic, and
- $\star$ $\omega = 0.5$ is severely contaminated, symmetric, and platykurtic.

Simulation Results

The experimental results are summarized in Table 3.3 for tests of bivariate normality. These suggest that in 50% location mixtures (a), scale mixture (c) and location-scale mixtures (d,f,g), the KLD can outperform the existing methods. And among the three proposed methods, KLD is more powerful in most cases of the alternative distributions. The results provide some evidences that the geometric mean of likelihood can be more measurable for difference. In case (e) of mixture normal, all of the test statistics are not sensitive enough.

In Table 3.3 , the two test statistics LRT and KLD based on quintic and TPS estimation are also different. To illustrate the different fittings of the two spline models, Figure 3.1

shows the estimated densities using quintic and TPS. Both models can fit well in the center. Compared with TPS, quintic spline can fit better in light tails.

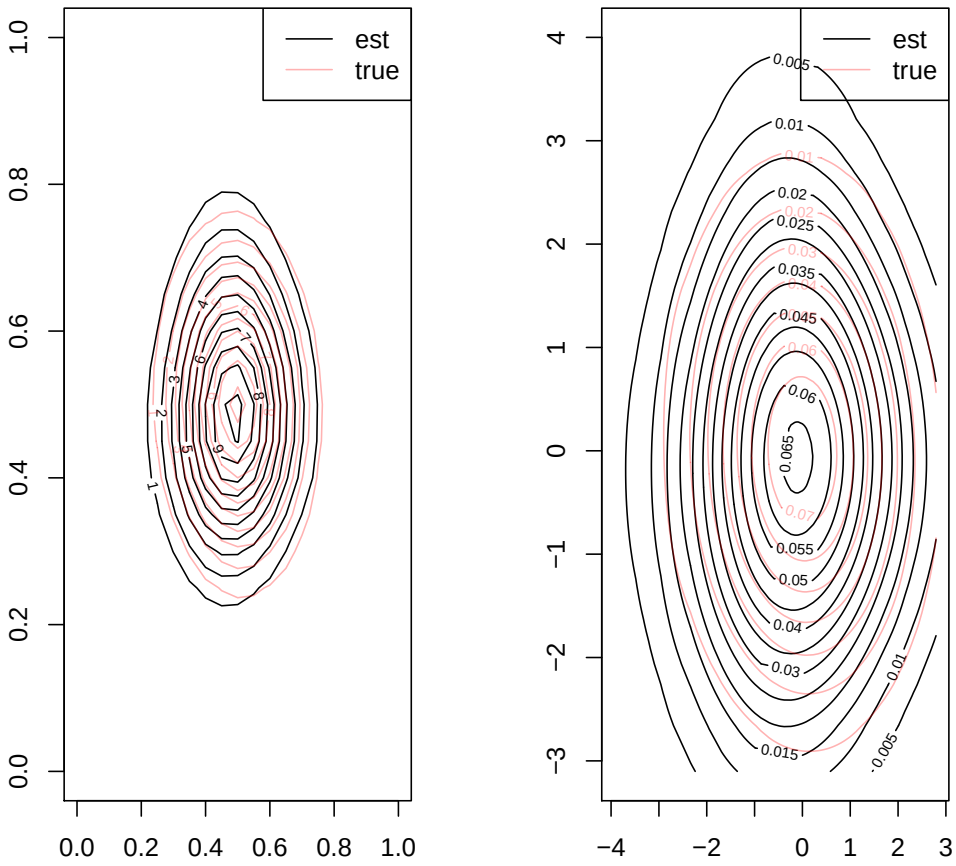

Figure 3.1: Quintic spline (left) density estimation and TPS (right) density estimation

Dist	Sample Size	LRT-T	KLD-T	LRT-P	KLD-P	DFMVN	Mardia's (Skew)	Mardia's (Kurtosis)	Henze-Zirkler	Royston	$\hat{\mathcal{E}}$
Higher Order	50			0.89	0.92	0.15	0.16	0.41	0.95	**1.00**	0.94
	100			**1.00**	**1.00**	0.18	0.41	0.74	**1.00**	**1.00**	**1.00**
(a)	50	0.51	0.55	0.65	**0.91**	0.19	0.00	0.09	0.85	0.49	0.72
	100	**1.00**	0.98	**1.00**	**1.00**	0.44	0.01	0.43	**1.00**	0.92	**1.00**
(b)	50	0.7	0.66	0.77	0.84	0.19	0.47	0.03	**0.91**	0.78	0.89
	100	1	0.99	**1.00**	**1.00**	0.45	0.86	0.02	**1.00**	0.98	**1.00**
(c)	50	0.98	0.96	**0.99**	0.97	0.87	0.92	0.11	**0.99**	0.40	**0.99**
	100	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.18	**1.00**	0.78	**1.00**
(d)	50	0.59	0.46	0.46	**0.69**	0.14	0.07	0.02	0.64	0.53	0.49
	100	0.96	0.89	0.95	**0.99**	0.39	0.11	0.22	0.96	0.94	0.90
(e)	50	0.08	0.08	0.06	0.05	0.06	0.02	0.01	0.05	0.06	0.05
	100	0.04	0.02	0.09	0.10	0.03	0.07	0.04	0.07	0.11	0.09
(f)	50	0.58	0.56	0.60	**0.86**	0.17	0.05	0.06	0.72	0.60	0.69
	100	**1.00**	0.96	0.99	**1.00**	0.39	0.07	0.31	0.98	0.89	0.97
(g)	50	0.61	0.53	0.67	**0.85**	0.15	0.10	0.05	0.75	0.57	0.63
	100	0.98	0.95	**1.00**	**1.00**	0.49	0.32	0.18	**1.00**	0.94	**1.00**

Table 3.3: Power Comparison based on Quintic Spline and TPS

3.4 Conclusion

We proposed three multivariate normality test statistics using smoothing spline density estimation. The three methods are based on likelihood ratio, KLD and RKHS norm respectively. It turns out that the KLD-statistic is more powerful in some 50% mixture multivariate cases (a,c,d,f,g) in the experiments, which indicates the geometric mean of likelihood can provide a more sensitive comparison of density estimation.

Chapter 4

Spline Density Estimation with Model Based Penalties

In this chapter we propose model-based penalties for smoothing spline density estimation and inference. These model-based penalties incorporate indefinite prior knowledge that the density is close to, but not necessarily in a family of distributions. We will use the Pearson and generalization of the generalized inverse Gaussian families to illustrate the derivation of penalties and reproducing kernels. We also propose new inference procedures to test the hypothesis that the density belongs to a specific family of distributions. We conduct extensive simulations to show that the model-based penalties can substantially reduce both bias and variance in the decomposition of the Kullback-Leibler distance, and the new inference procedures are more powerful than some existing ones.

4.1 Introduction

Density estimation has been widely studied due to its principal role in statistics and machine learning. Often there is prior information suggesting that the density function

can be well approximated by a parametric family of densities. For example, it may be known that the density is close to, but not necessarily is a Gamma distribution. This kind of indefinite information has not been explored in the field of density estimation. In his classic book on density estimation, Silverman [29] alluded that different penalties may be considered for different situations in the context of penalized likelihood density estimation. In particular, he suggested penalties to the second and third derivatives of the logarithm of the density so that zero penalties correspond to the exponential and normal density functions respectively. To the best of our knowledge, no research has been done to incorporate indefinite prior information into the construction of the penalties.

We will consider different penalties through L-splines in this chapter. The L-spline has been developed to incorporate prior knowledge in nonparametric regression models. It is known that the L-spline can reduce bias in the estimation of a regression function (Wahba [47], Heckman & Ramsay [48], Wang [28], Gu[11]). The goal of this chapter is to develop novel density estimation methods that can incorporate indefinite prior knowledge and consequently lead to better estimation procedures. In particular, we will consider model-based penalties for the Pearson family and the generalization of the generalized inverse Gaussian (GGIG) family, and derive penalties and reproducing kernels for some special cases in these families of distributions. We will show that the model-based penalties can substantially reduce both bias and variance in the decomposition of the Kullback-Leibler (KL) distance of smoothing spline estimates of density functions. Many methods have been developed in the literature to test the hypothesis that the density belongs to a specific family of distributions (Anderson & Darling [49], Stephens [50], Stephens [51]). We will develop new inference procedures based on L-spline estimates. To the best of our knowledge, this chapter is the first to employ L-splines for density estimation and inference.

The remainder of the article is organized as follows. Section 2 reviews L-splines. Sections 3 and 4 present model constructions for the Pearson and GGIG families respectively. Section 5 introduces new inference procedures based on L-spline estimates. Section 6 presents simulation studies to compare the proposed L-spline based estimation and inference procedures with existing methods.

4.2 Model-based Penalty and L-splines

As discussed in Section 1.2, in the construction of a smoothing spline model, one needs to decide the penalty functional $J(\eta)$, or equivalently, the null space $\mathcal{H}_0$ consisting of functions which are not penalized. The most popular choice of the penalty is the roughness penalty with $J(\eta) = \int_a^b (\eta^{(m)})^2 dx$. When $m = 2$ and $m = 3$ respectively, the null spaces $\mathcal{H}_0$ are the linear and quadratic functions which correspond to the exponential and normal distributions suggested in Silverman [29].

Often there exists information suggesting that f can be well approximated by a parametric family of densities, and logistic transformation of density functions in this family satisfy the differential equation $L\eta = 0$ where

$$L = D^m + \sum_{j=1}^{m-1} \omega_j(x) D^j \tag{4.1}$$

is a linear differential operator with $m \geq 1$, D^j is the jth derivative operator, and ω_i are continuous real-valued functions. Two such families of distributions, Pearson and GGIG, will be discussed in Sections 4.3 and 4.4.

An L-spline density estimate is the solution to (3.3) with penalty $J(\eta) = \int_a^b (L\eta)^2 dx$. Instead of the standard roughness penalty, an L-spline uses a penalty constructed based on a parametric model. The null space $\mathcal{H}_0$ corresponds to the specified parametric family

of densities. Therefore, it allows us to incorporate the information that η is close to, but not necessarily in the null space $\mathcal{H}_0$. Ramsay [48] called $\mathcal{H}_0$ as the favored parametric model. We will show in Section 4.7 that the model-based penalty can lead to better estimates of density functions. We will construct test procedures for the hypothesis that the density belongs to the specific parametric family in Section 4.6.

Since $\eta \in W_{20}^m[a,b]$, $L\eta$ exists and is square integrable. There exists real-valued functions, $\phi_1, ..., \phi_m$, such that they form a basis of $\mathcal{H}_0 = \{\eta : L\eta = 0\}$. Let

$$W(x) = \begin{pmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_m(x) \\ \phi_1'(x) & \phi_2'(x) & \cdots & \phi_m'(x) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(m-1)}(x) & \phi_2^{(m-1)}(x) & \cdots & \phi_m^{(m-1)}(x) \end{pmatrix}$$

be the Wronskian matrix associated with $\phi_1, ..., \phi_m$, and

$$G(x,s) = \begin{cases} \boldsymbol{\phi}^T(x)\boldsymbol{\phi}^*(s), & s \le x, \\ 0, & s > x, \end{cases}$$

be the Green function associated with L where $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_m(x))^T$ and $\boldsymbol{\phi}^*(x) = (\phi_1^*(x), \ldots, \phi_m^*(x))^T$ is the last column of $W^{-1}(x)$. Then $W_{20}^m[a,b]$ is an RKHS under the inner product

$$(\eta, \tilde{\eta}) = \sum_{\nu=0}^{m-1} \eta^{(\nu)}(a)\tilde{\eta}^{(\nu)}(a) + \int_a^b (L\eta)(L\tilde{\eta})dx,$$

and $W_{20}^m[a,b] = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 =$ span $\{\phi_1, \ldots, \phi_m\}$ and $\mathcal{H}_1 = \{\eta \in W_{20}^m[a,b] :$

$\eta^{(\nu)}(a) = 0, \nu = 0, \ldots, m-1\}$ are RKHS's with corresponding RKs

$$R_0(x,z) = \boldsymbol{\phi}^T(x)\{W^T(a)W(a)\}^{-1}\boldsymbol{\phi}(z), \tag{4.2}$$

$$R_1(x,z) = \int_a^b G(x,s)G(z,s)ds. \tag{4.3}$$

See Wang [28] for details.

4.3 L-spline for Pearson Family of Distributions

The Pearson family is a continuous distribution system proposed by Karl Pearson [52]. A Pearson density function $f(x)$ is any valid solution to the Pearson differential equation

$$\frac{1}{f(x)}\frac{df(x)}{dx} + \frac{a_0 + (x-a_4)}{a_1(x-a_4)^2 + a_2(x-a_4) + a_3} = 0, \tag{4.4}$$

where $a_0 = a_2 = \sqrt{\mu_2\beta_1}(\beta_2+3)/(10\beta_2-12\beta_1-18)$, $a_1 = (2\beta_2-3\beta_1-6)/(10\beta_2-12\beta_1-18)$, $a_3 = \mu_2(4\beta_2-3\beta_1)/(10\beta_2-12\beta_1-18)$, β_1 is the skewness, β_2 is the kurtosis, and μ_2 is the second central moments. Pearson identified 12 types of distributions based on different values of parameters. The Pearson family includes most commonly used distributions such as the uniform, exponential, normal, Gamma, Beta, inverse Gamma, Student's t and Cauchy distributions.

It is not difficult to show that the logistic transformation of density function in the Pearson family satisfy the differential equation $L\eta = 0$ where

$$L = D^3 + \frac{2(2a_1(x-a_4)+a_2)}{a_1(x-a_4)^2 + a_2(x-a_4) + a_3}D^2 + \frac{2a_1}{a_1(x-a_4)^2 + a_2(x-a_4) + a_3}D. \tag{4.5}$$

Therefore we can construct model-based penalties using (4.5) for densities in the Pearson family. Explicit constructions can be derived for many special cases. We illustrate

two such cases in the following two subsections.

4.3.1 Gamma distribution

The Gamma distribution (denoted as Gamma(α, β)) has density function

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \tag{4.6}$$

where $\alpha > 0$ and $\beta > 0$ are the shape and rate parameters, and Γ is the Gamma function.. It is a special case of the Pearson family (type III) with $a_1 = a_3 = a_5 = 0$, $a_2 = 1/\beta$ and $a_0 = -a_2(\alpha - 1)$. The logistic transformation of the density $\eta(x) = -\beta x + (\alpha - 1)\log(x)$.

Now consider the L-spline with model space $\eta(x) \in W^3_{20}[a, b]$ and differential operator

$$L = D^3 + \frac{2}{x} D^2. \tag{4.7}$$

As the domain of the Gamma distribution is $(0, \infty)$, we set a to be a small value closed to 0 and b large enough to cover all observations. The same method will be used for other distributions in the rest of this chapter which are not defined on compact intervals. It can be shown that $\mathcal{H}_0 = \text{span}\{x, \log(x)\}$ and the RK of $\mathcal{H}_1$

$$\begin{aligned} R_1(x, z) \;=\; & [1 + \log(z) + \log(x) + \log(z)\log(x)] I_4(x \wedge z) - [z + x + z\log(x) + x\log(z)] I_3(x \wedge z) \\ & + xz I_2(x \wedge z) + I_{4,2}(x \wedge z) - [2 + \log(z) + \log(x)] I_{4,1}(x \wedge z) + (z + x) I_{3,1}(x \wedge z), \end{aligned}$$

where $x \wedge z = \min(x, z)$, $I_p(s) = \int_0^s x^p dx = s^{p+1}/(p+1)$, and $I_{p,k}(s) = \int_0^s x^p [\log(x)]^k = s^{p+1}[\log(s)]^k/(p+1) - k I_{p+1,k-1}(s)/(p+1)$. A brief derivation of the RK can be found in A.2.

4.3.2 Beta distribution

The Beta distribution has the density function

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \tag{4.8}$$

where $\alpha > 0$ and $\beta > 0$ are the shape parameters. It is a special case of the Pearson family (type I) with $a_5 = a_3 = 0$, $a_1 = -a_2$, $\alpha = a_0/a_1 + 1$ and $\beta = (a_0 - 1)/a_1 + 1$. The logistic transformation $\eta(x) = (\alpha - 1)\log(x) + (\beta - 1)\log(1 - x)$.

Now consider the L-spline with model space $\eta(x) \in W^3_{20}[a, b]$ and differential operator

$$L = D^3 + \frac{2(2x-1)}{x(x-1)}D^2 + \frac{2}{x(x-1)}D. \tag{4.9}$$

It can be shown that $\mathcal{H}_0 = \text{span}\{\log(x), \log(1-x)\}$, and the RK of $\mathcal{H}_1$

$$\begin{aligned} R_1(x,z) \;=\; & [\log(z)\log(1-x) + \log(x)\log(1-z)]I(x\wedge z;3,3,0,0) \\ & + \log(1-x)\log(1-z)I(x\wedge z;2,4,0,0) + \log(x)\log(z)I(x\wedge z;4,2,0,0) \\ & + I(x\wedge z;2,4,0,2) - (\log(x)+\log(z))I(x\wedge z;3,3,0,1) \\ & - [\log(1-x)+\log(1-z)]I(x\wedge z;2,4,0,1) \\ & + I(x\wedge z;4,2,2,0) - [\log(x)+\log(z)]I(x\wedge z;4,2,1,0) \\ & - [\log(1-x)+\log(1-z)]I(x\wedge z;3,3,1,0) + 2I(x\wedge z;3,3,1,1), \end{aligned}$$

where

$$I(y;m_1,m_2,m_3,m_4) = \int_0^y x^{m_1}(1-x)^{m_2}\log(x)^{m_3}\log(1-x)^{m_4}dx.$$

A brief derivation of the RK is given in A.3.

4.4 L-spline for GGIG Family

Shakil, Kibria and Singh [53] proposed the GGIG family of distributions to include some other commonly used distributions such as the inverse Gaussian, generalized inverse Gaussian (GIG), Rayleigh and half-normal distributions which are not in the Pearson family. A GGIG density function $f(x)$ is the solution to the following differential equation

$$\frac{1}{f(x)}\frac{df(x)}{dx} = \frac{a_0 + a_p x^p + a_{2p} x^{2p}}{x^{p+1}}, \qquad x > 0. \tag{4.10}$$

The solution to the differential equation (4.10) is $f(x) = Cx^{\tau_1 - 1}\exp(-\tau_2 x^p - \tau_3 x^{-p})$ where $\tau_2 \geq 0$, $\tau_3 \geq 0$, $\tau_1 = a_p + 1$, $\tau_2 = -a_{2p}/p$, $\tau_3 = a_0/p$, and C is the normalizing constant. Then $\eta(x) = (\tau_1 - 1)\log(x) - \tau_2 x^p - \tau_3 x^{-p}$ which satisfies the differential equation $L\eta = 0$ where

$$L = \sum_{k=0}^{p+1} \binom{2p+1}{k} \left(D^k x^{p+1}\right) D^{2p+2-k}. \tag{4.11}$$

The null space $\mathcal{H}_0 = \text{span}\{\log(x), x, \ldots, x^p, x^{-1}, \ldots, x^{-p}\}$.

We now consider the special case with $p = 1$ which includes many commonly used distributions such as the inverse Gaussian (IG) ($\tau_1 = -0.5$), GIG, reciprocal IG ($\tau_1 = 0.5$), hyperbolic ($\tau_1 = 1$), Gamma ($\tau_3 = 0$), inverse Gamma ($\tau_1 = 0$), Erlang ($\tau_1 > 0$ and is an integer, $\tau_3 = 0$), and exponential ($\tau_1 = 1$ and $\tau_3 = 0$). In this case we have $g(x) = (\tau_1 - 1)\log(x) - \tau_2 x - \tau_3 x^{-1}$ and

$$L = D^4 + 6x^{-1}D^3 + 6x^{-2}D^2. \tag{4.12}$$

It is not difficult to show that $\mathcal{H}_0 = \text{span}\{\log(x), x, x^{-1}\}$ and the RK of $\mathcal{H}_1$ is

$$\begin{aligned} R_1(x,z) &= \frac{1}{36xz}(x\wedge z)^9 - \frac{1}{16}(x\wedge z)^8\log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) \\ &+\frac{1}{16}(x\wedge z)^8\left(\frac{1}{8x}+\frac{1}{8z}+\frac{1}{z}\log(x)+\frac{1}{x}\log(z)\right) \\ &-\frac{1}{7}(x\wedge z)^7\log(x\wedge z)\left(\frac{2}{7}+\log(x)+\log(z)\right)+\frac{1}{7}(x\wedge z)^7\log(x\wedge z)^2 \\ &+\frac{1}{7}(x\wedge z)^7\left(\frac{2}{49}-\frac{z}{4x}-\frac{x}{4z}+\frac{1}{7}\log(x)+\frac{1}{7}\log(z)+\log(x)\log(z)\right) \\ &+\frac{1}{12}(x+z)(x\wedge z)^6\log(x\wedge z)-\frac{1}{12}(\frac{x}{6}+x\log(z)+\frac{z}{6}+z\log(x))(x\wedge z)^6+\frac{1}{20}(x\wedge z)^5xz. \end{aligned}$$

A brief derivation of the RK is given in A.4.

4.5 L-spline for Inverse Gamma Distribution

The inverse gamma (IGM(α,β)) distribution's probability density function is defined over the support $x > 0$

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right), \quad \alpha>0, \beta>0. \tag{4.13}$$

The logit transformation of the density is $\eta(x) = (-\alpha-1)\log(x) - \beta/x$. It's not hard to see the inverse Gamma distribution is neither in the Pearson family, nor in the GGIG family.

Now consider the L-spline with model space $\eta(x) \in W^3_{20}[a,b]$ and differential operator

$$L = x^2D^3 + 4xD^2 + 2D. \tag{4.14}$$

It can be shown that the corresponding null space is $\mathcal{H}_0 = \text{span}\{\log(x), \frac{1}{x}\}$. And the RK

of $\mathcal{H}_1$ is

$$\begin{aligned}
R_1(x,z) = &\int_0^T G(x,s)G(z,s)ds \\
=&\frac{1}{7xz}(x\wedge z)^7 - \frac{1}{6}(x\wedge z)^6\log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) - \\
&\frac{1}{6}(x\wedge z)^6\left(\frac{5}{6x}+\frac{5}{6z}-\frac{1}{z}\log(x)-\frac{1}{x}\log(z)\right) + \\
&\frac{1}{5}(x\wedge z)^5\log(x\wedge z)\left(\frac{8}{5}-\log(x)-\log(z)\right)+\frac{1}{5}(x\wedge z)^5\log(x\wedge z)^2 + \\
&\frac{1}{5}(x\wedge z)^5\left(\frac{17}{25}-\frac{4}{5}\log(x)-\frac{4}{5}\log(z)+\log(x)\log(z)\right).
\end{aligned}$$

A brief derivation of the RK is given in Appendix A.5.

4.6 Inference of Density Using L-splines

Effective assessment of goodness-of-fit (GOF) and formal inference for a density function is critical in applications (Romantsova [54], Del Castillo & Puig [55], Lehmann & Ramano [56]). In this section we consider the problem of deciding whether the density belongs to a parametric family of distributions. Let $X_1,\ldots,X_n$ be iid samples with a density $f(x)$ on an interval $[a,b]$. We consider the null hypothesis $H_0: f\in\mathfrak{F}_0$ versus the alternative hypothesis $H_1: f\notin\mathfrak{F}_0$ where $\mathfrak{F}_0$ is a specific family of distributions. We assume that there exists a differential operator L as in (4.1) such that $L\eta=0$ for all $f\in\mathfrak{F}_0$ where η is the logistic transformation f. Note that the null hypothesis H_0 is equivalent to $\eta\in\mathcal{H}_0$.

4.6.1 Modified Anderson-Darling, Cramer-von Mises and Kolmogorov-Smirnov tests

A quadratic norm statistic based on the empirical distribution (EDF) is defined as

$$Q = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \, w(x) \, dF_0(x),$$

where F_n is the EDF, F_0 is an estimate of the cumulative density function (CDF) under the null hypothesis, and $w(x)$ is a weight function. Two well-known special cases are the Anderson-Darling (AD) and Cramer-von Mises (CVM) statistics with $w(x) = [F_0(x)\ (1 - F_0(x))]^{-1}$ and $w(x) = 1$ respectively (Stephen [51]).

Denote the CDF associated with the L-spline estimate of the density function as $F_s(x)$. Since L-splines with penalties constructed from specific families of distributions may provide better estimates of density functions (see Section 4.7), a natural extension of the AD and CVM statistics is to replace the EDF F_n in the quadratic norm statistic and weight function by F_s. The resulting modified testing methods are refereed to as AD-L and CVM-L.

Kolmogorov-Smirnov test statistic is defined as

$$\text{KS} = \sup_x |F_n(x) - F_0|. \tag{4.15}$$

Again, we can construct a new test statistic by replacing F_n with $F_s(x)$. The resulting modified testing method is referred to as KS-L.

4.6.2 Likelihood ratio and Kullback-Leibler tests

The likelihood ratio (LR) statistic is

$$\text{LRT} = 2(l_s - l_0) \tag{4.16}$$

where l_s is the log-likelihood with the L-spline density estimate, and l_0 is the log-likelihood with MLE estimates of the parameters under the null hypothesis.

The KL distance between two density functions f_1 and f_2 is defined as

$$\text{KL}(f_1, f_2) = \int_a^b f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \tag{4.17}$$

Let f_0 be the estimated density under the null hypothesis, and f_s be the L-spline estimate of the density function. We will then use the KL distance between f_0 and f_s, $\text{KL}(f_0, f_s)$, as the KL test statistic.

4.7 Simulations

In this section, we conduct simulations to evaluate the proposed estimation and inference methods and compare them with existing methods. The function *ssden* in the R package *gss* is used to compute smoothing spline estimates of density functions (Gu [57]).

We will compare the estimation performance between the L-spline and cubic spline models. Denote f as the true density and $\hat{f}$ as an estimate. We will use the KL distance $\text{KL}(f, \hat{f})$ to assess the performance of estimation. We will use the generalized

decomposition

$$\mathrm{E}(\mathrm{KL}(f,\hat{f})) = \mathrm{KL}(f,\bar{f}) + \mathrm{E}(\mathrm{KL}(\bar{f},\hat{f})) = \text{bias} \ + \ \text{variance} \tag{4.18}$$

proposed by Heskes [58] to evaluate the bias-variance trade-off where $\bar{f} = \exp[\mathrm{E}(\log \hat{f})]/Z$ and Z is a normalization constant.

For density inference we will consider eight methods: Anderson-Darling (AD), Cramer-von Mises (CVM), Kolmogorov-Smirnov (KS), modified AD (AD-L), modified CVM (CVM-L), modified KS (KS-L), likelihood ratio (LR) and Kullback-Leibler (KL) tests. We will use the bootstrap method to approximate null distributions for all tests where the number of bootstrap samples is set to be 1000.

We will present results for two distributions, Gamma and inverse Gaussian, as the favored parametric models. We will consider three sample sizes, $n = 100$, $n = 200$ and $n = 300$. In addition, we will also present the density estimation results for two more distributions, Beta and inverse Gamma, as the favored parametric models with sample size 100. Results for other distributions and sample sizes are similar. We generate 100 data replicates for each simulation setting.

4.7.1 Gamma distribution as the favored parametric model

The generalized Gamma family has the density function

$$f(x;\alpha,\beta,\delta) = \frac{\delta\beta^{\alpha}}{\Gamma(\alpha/\delta)} x^{\alpha-1} e^{-(\beta x)^{\delta}}, \qquad \alpha > 0,\ \beta > 0,\ \delta > 0, \quad x > 0. \tag{4.19}$$

The Gamma distribution Gamma(α,β) is a special case with $\delta = 1$. We set $\alpha = 2$ and $\beta = 1$ in our simulations, and consider three choices of δ: $\delta = 1$, $\delta = 2$, and $\delta = 3$ which reflect different degree of closeness to the Gamma distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (4.7) and the cubic spline estimate of the density. Table 4.1 lists biases, variances, and KL distances for the L-spline and cubic spline estimates under different simulation settings. The L-spline with model-based penalty has smaller biases, variances, and KL distances than the cubic spline when $\delta = 1, \delta = 2$ and $\delta = 3$ when sample size is big. As expected, the improvement is larger when the true distribution is closer to the Gamma distribution.

δ	Model	n=100			n=200			n=300		
		Bias	Var	KL	Bias	Var	KL	Bias	Var	KL
1	Cubic	15.84	19.64	35.48	10.29	13.39	23.68	7.68	10.48	18.16
	L-spline	0.94	13.31	14.25	0.53	6.20	6.73	0.13	4.61	4.74
2	Cubic	7.61	15.07	22.68	5.98	9.30	15.28	4.14	7.07	11.21
	L-spline	1.78	16.05	17.83	0.92	9.76	10.67	0.89	6.22	7.11
3	Cubic	3.18	17.40	20.58	3.05	8.13	11.18	1.89	6.78	8.67
	L-spline	3.17	17.62	20.79	2.37	9.52	11.89	1.41	6.36	7.77

Table 4.1: Biases, variances, and KL distances in 10^{-3} with the generalized Gamma distribution.

For density inference we consider the null hypothesis that the distribution is Gamma. Table 4.2 lists powers of eights test methods with significance level set at 5%. The powers are the probability of type I error when $\delta = 1$. It is clear that all methods have probability of type I error smaller or close to 5%. With the EDF being replaced by the L-spline estimate, the modified AD, CVM and KS tests in general have larger powers than those from the original tests.

Table 4.3 lists more simulation results for testing the null hypothesis of a Gamma distribution against one of the distributions listed below:

1. The inverse Gaussian distributions defined in (4.21). We set $\kappa = 1$ and denote the density as $IG(\mu)$.

δ	Sample Size	AD	AD-L	CVM	CVM-L	KS	KS-L	LR	KL
	100	0.19	0.10	0.18	0.04	0.14	0.16	0.15	**0.27**
0.6	200	0.37	0.29	0.34	0.19	0.31	0.36	0.32	**0.46**
	300	0.58	0.37	0.58	0.47	0.38	0.55	0.54	**0.64**
	100	0.06	0.06	0.05	0.05	0.05	0.04	0.06	0.05
1	200	0.04	0.05	0.06	0.04	0.05	0.04	0.05	0.05
	300	0.04	0.02	0.04	0.01	0.04	0.01	0.01	0.01
	100	0.13	**0.17**	0.15	0.16	**0.17**	0.14	**0.17**	0.14
2	200	0.38	**0.48**	0.30	0.47	0.24	0.46	0.45	0.42
	300	0.47	**0.56**	0.45	**0.56**	0.33	**0.56**	0.53	0.53
	100	0.25	**0.32**	0.22	0.31	0.17	**0.32**	0.31	0.31
3	200	0.48	**0.67**	0.39	0.64	0.32	0.64	0.66	0.66
	300	0.66	**0.77**	0.61	0.76	0.53	0.75	0.75	0.76

Table 4.2: Powers of eight test methods for the Gamma distribution.

2. The lognormal distribution with density

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}. \tag{4.20}$$

   We set $\mu = 0$, and denote the density as $LN(\sigma)$.

3. The Gompertz distribution with density

$$f(x;\eta,b) = b\eta e^{bx} e^{\eta} \exp\left(-\eta e^{bx}\right) b\eta e^{bx} e^{\eta} \exp\left(-\eta e^{bx}\right).$$

   We set $b = 1$, and denote the density as $GO(1/\eta)$.

4. The linear failure rate distribution with density

$$f(x;\theta) = (1+\theta x)\exp\left(-x - \frac{\theta x^2}{2}\right).$$

   We denote it as $LF(\theta)$.

We also calculate the skewness of each distribution. When the distributions under

Distribution	Size	AD	AD-L	CVM	CVM-L	KS	KS-L	LR	KL
IG(1)	30	0.36	0.25	0.32	0.37	0.29	0.38	0.32	**0.38**
	50	0.51	0.42	0.41	0.53	0.39	0.54	0.53	**0.58**
	100	0.83	0.81	0.81	0.84	0.66	0.85	0.87	**0.89**
IG(1.5)	30	0.20	0.23	0.19	0.24	0.19	0.23	0.18	**0.26**
	50	0.30	0.33	0.27	0.37	0.21	0.38	0.28	**0.40**
	100	0.68	0.69	0.65	0.72	0.48	0.73	0.74	**0.77**
LN(0.8)	30	0.21	0.21	0.17	**0.31**	0.10	0.31	0.21	0.30
	50	0.30	0.33	0.30	0.39	0.25	0.41	0.31	**0.42**
	100	0.60	0.60	0.57	0.60	0.47	0.64	0.61	**0.68**
GO(2)	30	0.32	**0.54**	0.31	0.43	0.28	0.42	0.30	0.30
	50	0.59	**0.79**	0.57	0.75	0.48	0.78	0.64	0.69
	100	0.91	**0.99**	0.88	**0.99**	0.70	**0.99**	0.95	0.98
GO(4)	30	0.49	**0.66**	0.45	0.53	0.39	0.52	0.47	0.41
	50	0.70	**0.85**	0.68	0.81	0.49	0.80	0.75	0.74
	100	0.96	**1.00**	0.96	**1.00**	0.91	**1.00**	0.98	0.99
LF(2)	30	0.15	**0.24**	0.15	0.20	0.15	0.17	0.21	0.14
	50	0.24	**0.43**	0.20	0.29	0.18	0.32	0.26	0.23
	100	0.42	**0.60**	0.40	0.58	0.39	0.58	0.50	0.50
LF(4)	30	0.19	**0.32**	0.18	0.16	0.16	0.23	0.18	0.13
	50	0.16	**0.35**	0.16	0.24	0.11	0.26	0.16	0.14
	100	0.58	**0.81**	0.51	0.74	0.41	0.80	0.63	0.65

Table 4.3: Powers of eight test methods for the Gamma distribution under different alternatives.

the alternative are GG(0.6,2), IG(1), IG(1.5) and LN(0.8) with which the skewness is greater than the Gamma distribution under the null (GG(1,2)), the KL statistic is more powerful. When the distributions under the alternative are GG(2,2), GG(3,2), GO(2), GO(4), LF(2) and LF(4) whose skewness is less than the Gamma distribution (GG(1,2)), the AD-L statistic is more powerful.

p	Model	n=100			n=200			n=300		
		Bias	Var	KL	Bias	Var	KL	Bias	Var	KL
1	Cubic	40.51	2.32	42.84	50.69	1.58	52.27	43.86	1.11	44.97
	L-spline	7.51	1.67	9.18	4.72	0.88	5.60	1.85	0.54	2.39
2	Cubic	44.39	1.72	46.11	44.55	1.01	45.55	45.91	0.82	46.72
	L-spline	13.59	1.49	15.07	9.07	0.89	9.96	4.07	0.82	4.89
3	Cubic	46.11	1.44	47.55	51.23	0.85	52.08	54.30	0.70	55.00
	L-spline	65.28	1.86	67.15	29.34	1.79	31.13	35.74	0.45	36.19

Table 4.4: Biases, variances, and KL distances in 10^{-2} with the GIGG family.

4.7.2 Inverse Gaussian distribution as the favored parametric model

The inverse Gaussian (IG) has density function

$$f(x;\mu,\kappa) = \Big(\frac{\kappa}{2\pi x^3}\Big)^{1/2} \exp\left\{\frac{-\kappa(x-\mu)^2}{2\mu^2 x}\right\}, \quad x>0, \tag{4.21}$$

where $\mu > 0$ is the mean and $\kappa > 0$ is the shape parameter. It belongs to the GGIG family with $p = 1$, $\tau_1 = -0.5$, $\tau_2 = 0.5\kappa/\mu^2$, and $\tau_3 = \kappa/2$. We set $\tau_2 = \tau_3 = 2$ in our simulations, and consider three choices of p: $p = 1$, $p = 2$ and $p = 3$ in (4.10), which reflect different degrees of closeness to the inverse Gaussian distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (4.12) and the cubic spline estimate of the density. Table 4.4 lists biases, variances, and KL distances for the L-spline and cubic spline estimates under different simulation settings. The L-spline with model-based penalty has smaller biases, variances, and KL distances than the cubic spline for all settings except when $p = 3$ and $n = 100$.

For density inference we consider the null hypothesis that the distribution is IG. We generate iid samples from the generalized inverse Gaussian (GIG) density

$$f(x) = \frac{(\alpha_0/\alpha_1)^{\zeta/2}}{2K_\zeta(\sqrt{\alpha_0\alpha_1})} x^{(\zeta-1)} e^{-\frac{(\alpha_0 x+\alpha_1/x)}{2}}, \quad \alpha_0>0,\ \alpha_1>0, x>0, \tag{4.22}$$

ζ	Sample Size	AD	CVM	CVM-L	KS	KS-L	LR	KL
	300	0.05	0.03	0.06	0.06	0.06	0.03	0.08
-0.5	200	0.06	0.05	0.01	0.05	0.02	0.04	0.02
	100	0.04	0.04	0.02	0.03	0.03	0.03	0.05
	300	0.54	0.48	**0.67**	0.36	**0.67**	0.54	0.54
3	200	0.37	0.32	**0.44**	0.29	**0.42**	0.35	0.35
	100	0.19	0.2	**0.24**	0.1	0.22	0.15	0.16
	300	0.34	0.32	**0.37**	0.22	0.36	0.26	0.28
2	200	0.34	0.29	**0.38**	0.25	**0.38**	0.25	0.29
	100	0.16	0.15	0.18	0.11	**0.19**	0.12	0.08
	300	0.19	0.21	**0.34**	0.18	**0.34**	0.21	0.21
-3	200	0.16	0.17	**0.23**	0.15	**0.23**	0.17	0.19
	100	0.14	0.12	0.14	0.12	0.14	**0.17**	0.15
	300	0.06	0.06	**0.14**	0.06	**0.14**	0.06	0.06
-2	200	0.13	0.14	**0.15**	0.13	0.14	0.08	0.08
	100	0.12	0.12	0.11	0.08	**0.12**	0.09	0.09

Table 4.5: Powers of seven test methods for the IG distribution.

where K_ζ is a modified Bessel function of the second kind. The IG is a special case of GIG with $\zeta = -0.5$. We set $\alpha_0 = 3$ and $\alpha_1 = 3$ in the simulation, and consider five choices of ζ: $\zeta = -3, -2, -0.5, 2$ and 3 which reflect different degrees of departure from the IG distribution. Table 4.5 lists powers of seven test methods with significance level set at 5%. The AD-L statistic cannot be calculated since the estimate of $F_0(x)\,(1 - F_0(x))$ is close to zero. The powers are the probability of type I error when $\zeta = -0.5$. It is shown that all methods have type I error smaller or close to 5%. Again, the modified CVM and KS tests have larger powers than those from the original tests.

θ	L-spline			Cubic		
	KL	Var	Bias	KL	Var	Bias
0.25	17.38	11.66	5.72	65.62	30.91	34.71
0.5	12.82	12.30	0.52	22.04	12.06	9.98
0.75	12.63	12.12	0.52	27.17	15.43	11.74
1	11.43	11.03	0.40	25.80	15.94	9.86
2	13.13	12.58	0.54	28.41	16.62	11.79
3	14.50	13.80	0.70	32.45	19.00	13.45
5	11.74	11.24	0.50	29.51	16.81	12.69

Table 4.6: Biases, variances, and KL distances in 10^{-3} with the Generalised Beta

4.7.3 Beta distribution as the favored parametric model

The generalized Beta distribution used in the simulations has probability density function

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\theta(x/s)^{ac}[1-(x/s)^{\theta}]^{\beta-1}}{x}, \quad \alpha > 0, \beta > 0, \theta > 0, s > 0, 0 < x < s. \tag{4.23}$$

The Beta distribution Beta(α, β) is special case with $s = 1$ and $\theta = 1$. We set $\alpha = 3, \beta = 3, s = 1$ in our simulations, and consider seven choices of c: $c = 0.25, 0.5, 0.75, 1, 2, 3, 5$, which reflect different degree of closeness to the Beta distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (4.9) and the cubic spline estimate of the density. Table 4.6 lists biases, variances and KL distances for the L-spline and cubic spline estimates under different simulation settings. The L-spline with model-based penalty has smaller biases, variances, and KL distances than the cubic spline for all settings.

4.7.4 Inverse Gamma distribution as the favored parametric model

The generalized inverse Gamma distribution has density function

$$f(x) = \frac{\gamma}{\beta^{\alpha\gamma+2}\Gamma(\alpha)} x^{-\alpha\gamma-1} \exp\left\{-\left(\frac{\beta}{x}\right)^{\gamma}\right\} \quad \alpha > 0, \beta > 0, \gamma > 0. \tag{4.24}$$

The inverse Gamma distribution IGM(α, β) is a special case with $\gamma = 1$. We set $\alpha = 3, \beta = 1$ in the simulations, and consider the following choices of γ: $\gamma = 0.5, 0.75, 1, 2, 3, 5$ which reflect different degree of closeness to the Gamma distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (4.14) and the cubic spine estimate of the density. Table 4.7 lists biases, variances, and KL distances for L-spline and cubic spline estimates under different simulation settings. The L-spline with model-based penalty has smaller biases, variances and KL distances than the cubic spline when $\gamma = 0.5, 0.75, 1, 2, 3$. As expected, the improvement is larger when the true distribution is closer to the inverse Gamma distribution.

γ	L-spline			Cubic		
	KL	Var	Bias	KL	Var	Bias
0.5	27.59	18.04	9.55	140.11	37.23	102.88
0.75	18.43	16.52	1.91	173.59	33.72	139.87
1	15.17	14.41	0.75	118.26	41.92	76.34
2	18.73	15.30	3.43	47.86	26.93	20.93
3	20.37	15.54	4.83	35.35	20.46	14.89
5	106.68	83.96	22.72	29.17	20.35	8.82

Table 4.7: Biases, variances and KL distances in 10^{-3} with the generalised inverse Gamma distribution

4.8 Conclusion

In this chapter, we proposed model-based penalties for smoothing spline density estimation and inference. The model-based penalties successfully incorporate indefinite prior information about the density in density estimation and inference process. Two examples, respectively from the Pearson and GGIG family, are used to show the derivation of the penalties. The simulation results show the substantial reduction of the KL divergence, including both bias and variance, of density estimation with the new model-based penalties, and power gain using the L-spline based Anderson-Darling (AD), Cramer-von Mises (CVM), Kolmogorov-Smirnov (KS), LRT and KL tests.

Chapter 5

Semiparametric Density Estimation with Smoothing Spline

5.1 Introduction

Let X_i, $i = 1, ..., n$, be independent and identically distributed (iid) random samples from a probability density $f(x)$ on a general domain $\mathcal{X}$. We are interested in the estimation and inference of $f(x)$ from the observations. When some parametric form of $f(x)$ is assumed, say $f \in P_{\boldsymbol{\theta}} = \{f(x; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where $f(x, \boldsymbol{\theta})$ is known up to a finite-dimensional parameter $\boldsymbol{\theta}$, density estimation reduces to parameter estimation, for which the maximum likelihood method is the standard technique possessing many favorable properties. When a parametric form is not available, nonparametric methods such as kernel density estimation (Silverman [5]) and smoothing spline density estimation have been developed (Gu [11]).

Often in practice it is desirable to model some components of the density function parametrically while leaving other components unspecified. Many semiparametric density

models have been proposed for different purposes, some of which have been reviewed in Chapter 1. The objective of this chapter is to study a general semiparametric density model, develop estimation methods and computational procedures, and study theoretical properties.

5.2 Semiparametric Density Models

We consider the following general semiparametric density model

$$f(x) = \frac{\exp\{\eta(x;\boldsymbol{\theta},h)\}}{\int \exp\{\eta(x;\boldsymbol{\theta},h)\}\,dx}, \tag{5.1}$$

where η is a known function of x with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and $h \in \mathcal{S}$, and $\mathcal{S}$ is an RKHS. The general semiparametric model (5.1) includes several special cases of $\eta(x;\boldsymbol{\theta},h)$ discussed below.

1. η is linear with respect to parametric and nonparametric component,

$$\eta(x;\boldsymbol{\theta},h) = \boldsymbol{\alpha}^T(x)\boldsymbol{\theta} + h(x), \tag{5.2}$$

   where $\boldsymbol{\alpha}(x)$ is a vector of known functions of x. Models in Efron and Tibshiran [15], Lenk [16] and Yang [17] are special cases of (2). These models were proposed for different purposes as discussed in Chapter 1.

2. η is additive and linear with respect to nonparametric component, but nonlinear with respect to parameters,

$$\eta(x;\boldsymbol{\theta},h) = \alpha(x;\boldsymbol{\theta}) + h(x), \tag{5.3}$$

where α is a known function of x with unknown parameters $\boldsymbol{\theta}$. The model proposed by Hjort and Glad [14] is a special case of (5.3).

3. η is a transformation model,

$$\eta(x; \boldsymbol{\theta}, h) = h(\alpha(x; \boldsymbol{\theta})), \tag{5.4}$$

where α is a known differentiable and invertible function with unknown parameters $\boldsymbol{\theta}$. With transformation $Y = \alpha(X; \boldsymbol{\theta})$, the logistic density of Y is $\eta(y) = h(y) - \log\left(|\alpha'(\alpha^{-1}(y; \boldsymbol{\theta}))|\right)$, which becomes an additive model. The location-scale family density estimation belongs to this case with $\alpha(x; \boldsymbol{\theta}) = (x - \mu)/\sigma$, where μ and σ are the location and scale parameters and $\boldsymbol{\theta} = (\mu, \sigma)$.

4. η is a mixture model of two densities,

$$f(x, \pi) = \pi f_1(x, \boldsymbol{\theta}) + (1 - \pi) f_2(x),$$

where $0 \leq \pi \leq 1$. Originally considered by Olkin and Spiegelman [13], this model is often referred to as the two-component mixture model. Different estimation methods and applications can be found in Bordes, Mottelet, Vandekerkhove et al. [59] and Ma, Yao et al. [60].

Some of the existing models discussed above are summarized in Table 5.1.

Publication	Original Model	$\eta(x)$
Olkin and Spiegelman [13]	$g(x,\pi) = \pi f_1(x,\boldsymbol{\theta}) + (1-\pi) f_2(x)$	$\eta(x) = \log\{\pi f_1(x,\boldsymbol{\theta}) + (1-\pi) f_2(x)\}$
Hjort and Glad [14]	$f(x) = f(x,\boldsymbol{\theta}) r(x)$	$\eta(x) = \log\{f(x,\boldsymbol{\theta})\} + \log\{r(x)\}$
Efron and Tibshirani [15]	$g_\theta(y) = g_0(x)\exp(\theta_1 + \boldsymbol{t}^T(x)\boldsymbol{\theta}_2)$	$\eta(x) = \log(g_0(x)) + \theta_1 + \boldsymbol{t}^T(x)\boldsymbol{\theta}_2$
Yang [17]	$f(x\|\boldsymbol{\theta}, h) = \frac{\exp[\boldsymbol{\alpha}^T(x)\boldsymbol{\theta} + h(x)]}{\int_{\mathcal{X}} \exp[\boldsymbol{\alpha}^T(x)\boldsymbol{\theta} + h(x)] dG(x)}$	$\eta(x) = \boldsymbol{\alpha}^T(x)\boldsymbol{\theta} + h(x)$

Table 5.1: Some existing models as special cases of model (5.1).

Often certain conditions are necessary to make model (5.1) identifiable. These identifiability conditions depend on specific models. We assume that model (5.1) is identifiable, and discuss identifiability conditions for specific models in the following sections. We will develop different estimation procedures in Section 5.3 and asymptotic properties for model (5.2) in Section 5.4. Simulations are conducted to evaluate the proposed estimation procedures in Section 5.5. Section 5.6 shows applications to several real data sets.

5.3 Estimation

In this section, we first describe an estimation procedure for the linear and additive cases of the semiparametric density model in Section 5.3.1, and then adapt it to the general model in Section 5.3.2. We present estimation procedures for one sample and two sample transformation models in Sections 5.3.3 and 5.3.4 respectively.

5.3.1 Profiled penalized likelihood estimation for the additive model

Since the linear case in (5.2) is a special case of the additive model in (5.3), we will only discuss the estimation procedure for (5.3) here. We estimate $\boldsymbol{\theta}$ and h as minimizers of the following penalized likelihood

$$-\sum_{i=1}^{n}\{\alpha(X_i;\boldsymbol{\theta})+h(X_i)\}+\log\int\exp\{\alpha(x;\boldsymbol{\theta})+h(x)\}dx+\frac{\lambda}{2}J(h), \tag{5.5}$$

where the first and second components in (5.5) correspond to the negative log likelihood, the penalty $J(h)$ is a square (semi) norm, and λ is the smoothing parameter. Consider the space $\mathbb{R}^p\times\mathcal{H}$ for $(\boldsymbol{\theta},h)$, where $\mathcal{H}=\mathcal{H}_0\oplus\mathcal{H}_1$ is an RKHS with R_J as RK of $\mathcal{H}_1$ and $\boldsymbol{\phi}(x)=(\phi_1(x),...,\phi_m(x))^T$ as the vector of basis functions of $\mathcal{H}_0=\{h:J(h)=0\}$.

Fixing $\boldsymbol{\theta}$, as in Gu (2013) we approximate the solution of h to (5.5) by

$$\hat{h}_\theta(x)=\sum_{\nu=1}^{m}d_\nu\phi_\nu(x)+\sum_{j=1}^{q}c_jR_J(Z_j,x)=\boldsymbol{\phi}^T(x)\boldsymbol{d}+\boldsymbol{\xi}^T(x)\boldsymbol{c}, \tag{5.6}$$

where $\{Z_j,j=1,...,q\}$ is a random sample of $\{X_i,i=1,...,n\}$, $\boldsymbol{\xi}(x)=(R_J(Z_1,x),...,R_J(Z_q,x))^T$, $\mathbf{c}=(c_1,...,c_q)^T$ and $\mathbf{d}=(d_1,...,d_m)^T$. The dependence of h on $\boldsymbol{\theta}$ is expressed explicitly. Then the calculation of $\hat{h}_\theta$ reduces to the minimization of

$$\begin{aligned}A(\boldsymbol{c},\boldsymbol{d})=&-\frac{1}{n}\sum_{i=1}^{n}\alpha(X_i,\boldsymbol{\theta})-\frac{1}{n}\mathbf{1}^{\boldsymbol{T}}(S\boldsymbol{d}+R\boldsymbol{c})+\\&\log\int\exp\{\alpha(x,\boldsymbol{\theta})+\boldsymbol{\phi}^T(x)\boldsymbol{d}+\boldsymbol{\xi}^T(x)\boldsymbol{c}\}dx+\frac{\lambda}{2}\boldsymbol{c}^TQ\boldsymbol{c}\end{aligned} \tag{5.7}$$

with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$, where S is an $n\times m$ matrix with the (i,ν)th entry $\phi_\nu(X_i)$, R

is an $n \times q$ matrix with the (i,j)th entry $\xi_j(X_i) = R_J(Z_j, X_i)$, and Q is a $q \times q$ matrix with the (j,k)th entry $R_J(Z_j, Z_k)$. Newton method is applied to obtain $\mathbf{c}$ and $\mathbf{d}$. Define

$$\mu_f(g) = \frac{\int g e^f dx}{\int e^f dx}, \tag{5.8}$$

$$V_f(g,h) = \mu_f(gh) - \mu_f(g)\mu_f(h), \tag{5.9}$$

and denote $V_f(g) = V_f(g,g)$. Let $\tilde{\eta}(x) = \alpha(x;\boldsymbol{\theta}) + \tilde{h}_\theta(x)$, where $\tilde{h}_\theta(x) = \boldsymbol{\phi}^T(x)\tilde{\boldsymbol{d}} + \boldsymbol{\xi}^T(x)\tilde{\boldsymbol{c}}$, $\tilde{\boldsymbol{c}} = (\tilde{c}_1, .., \tilde{c}_q)^T$ and $\tilde{\boldsymbol{d}} = (\tilde{d}_1, ..., \tilde{d}_m)^T$ are $\boldsymbol{c}$ and $\boldsymbol{d}$ calculated at the previous step in the Newton iterative method. Taking derivatives with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$ and evaluated at $\tilde{\eta}_\theta$, we have

$$\begin{aligned}
\frac{\partial A}{\partial \boldsymbol{d}} &= -\frac{1}{n} S^T \mathbf{1} + \mu_{\tilde{\eta}}(\boldsymbol{\phi}), \\
\frac{\partial A}{\partial \boldsymbol{c}} &= -\frac{1}{n} R^T \mathbf{1} + \mu_{\tilde{\eta}}(\boldsymbol{\xi}) + \lambda Q \boldsymbol{c}, \\
\frac{\partial A^2}{\partial \boldsymbol{d} \partial \boldsymbol{d}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T), \\
\frac{\partial A^2}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T) + \lambda Q, \\
\frac{\partial A^2}{\partial \boldsymbol{d} \partial \boldsymbol{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T),
\end{aligned}$$

where $V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$ is an $m \times m$ matrix with the (i,j)th entry $V_{\tilde{\eta}}(\phi_i, \phi_j)$, $V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T)$ is an $m \times q$ matrix with the (i,j)th entry $V_{\tilde{\eta}}(\phi_i, \xi_j)$, $V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\phi}^T)$ is the transpose of $V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T)$, $V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$ is an $q \times q$ matrix with the (i,j)th entry $V_{\tilde{\eta}}(\xi_i, \xi_j)$. Therefore, the Newton updating equation is

$$\begin{pmatrix} V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T) & V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T) \\ V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T) & V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T) + \lambda Q \end{pmatrix} \begin{pmatrix} \boldsymbol{d} - \tilde{\boldsymbol{d}} \\ \boldsymbol{c} - \tilde{\boldsymbol{c}} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{1}/n - \mu_{\tilde{\eta}}(\boldsymbol{\phi}) \\ R^T \mathbf{1}/n - \mu_{\tilde{\eta}}(\boldsymbol{\xi}) - \lambda Q \tilde{\boldsymbol{c}} \end{pmatrix}.$$

Rearranging the terms, we obtain

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n - \mu_{\tilde{\eta}}(\boldsymbol{\phi}) + V_{\phi,\tilde{h}} \\ R^T\mathbf{1}/n - \mu_{\tilde{\eta}}(\boldsymbol{\xi}) + V_{\xi,\tilde{h}} \end{pmatrix}, \tag{5.10}$$

where $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$, $V_{\phi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T)$, $V_{\xi,\phi}$ is the transpose of $V_{\phi,\xi}$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$, $V_{\phi,\tilde{h}} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \tilde{h}) = (V_{\tilde{\eta}}(\phi_1, \tilde{h}), ..., V_{\tilde{\eta}}(\phi_m, \tilde{h}))^T$, and $V_{\xi,\tilde{h}} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \tilde{h}) = (V_{\tilde{\eta}}(\xi_1, \tilde{h}), ..., V_{\tilde{\eta}}(\xi_q, \tilde{h}))^T$. At convergence we denote the estimate of h with fixed $\boldsymbol{\theta}$ as $\hat{h}_\theta$.

We select the smoothing parameter λ by optimizing the KL distance between the density associated with the estimate, $\eta(x; \boldsymbol{\theta}, \hat{h}_\theta)$, and the density associated with the true $\eta(x; \boldsymbol{\theta}, h)$,

$$\begin{aligned} \mathrm{KL}(\eta(x;\boldsymbol{\theta},h), \eta(x;\boldsymbol{\theta},\hat{h}_\theta)) =& \mu_{\eta(x;\boldsymbol{\theta},h)}(\eta(x;\boldsymbol{\theta},\hat{h}_\theta) - \eta(x;\boldsymbol{\theta},h)) - \log \int e^{\eta(x;\boldsymbol{\theta},h)} dx \\ &+ \log \int e^{\eta(x;\boldsymbol{\theta},\hat{h}_\theta)} dx. \end{aligned}$$

Dropping terms not involving $\hat{h}_\theta$, we have the relative Kullback-Leibler distance,

$$\mathrm{RKL}(\eta(x;\boldsymbol{\theta},h), \eta(x;\boldsymbol{\theta},\hat{h}_\theta)) = \mu_{\eta(x;\boldsymbol{\theta},h)}(\eta(x;\boldsymbol{\theta},\hat{h}_\theta)) + \log \int e^{\eta(x;\boldsymbol{\theta},\hat{h}_\theta)} dx. \tag{5.11}$$

The second term is computable, but the first term involves the unknown density. We apply the cross-validation method to estimate $\mu_{\eta(x;\boldsymbol{\theta},h)}(\eta(x;\boldsymbol{\theta},\hat{h}_\theta))$. Let $h_\theta^{[i]}$ denote the minimizer of the delete-one version of (5.5) with the fixed $\boldsymbol{\theta}$

$$-\frac{1}{n-1}\sum_{j\neq i}\{\alpha(X_j;\boldsymbol{\theta}) + h(X_j)\} + \log \int \exp\{\alpha(x;\boldsymbol{\theta}) + h(x)\} dx + \frac{\lambda}{2} J(h). \tag{5.12}$$

For an analytically tractable approximation of $h_\theta^{[i]}$, consider the quadratic approximation of (5.12). For $g_1, g_2 \in \mathcal{H}, \eta_1 = \alpha(x;\boldsymbol{\theta}) + g_1, \eta_2 = \alpha(x;\boldsymbol{\theta}) + g_2$ and a real number r, define

$L_{g_1,g_2}(r) = \log \int \exp(\alpha(x;\boldsymbol{\theta}) + g_1 + r g_2)dx$ as a function of r. It is not hard to show that $\dot{L}_{g_1,g_2}(0) = \mu_{\eta_1}(g_2)$ and $\ddot{L}_{g_1,g_2}(0) = V_{\eta_1}(g_2)$. Setting $g_1 = h^*, g_2 = h - h^*$, and $r = 1$, one has the Taylor expansion

$$\log \int e^{\alpha(x;\boldsymbol{\theta})+h(x)}dx = L_{h^*,h-h^*}(1) \approx L_{h^*,h-\tilde{h}}(0) + \mu_{\eta^*}(h-h^*) + \frac{1}{2}V_{\eta^*}(h-h^*), \quad (5.13)$$

where $\eta^* = \alpha(x;\boldsymbol{\theta}) + h^*$. Substituting the right-hand side of (5.13) for the term $\log \int e^{\alpha(x;\boldsymbol{\theta})+h(x)}dx$ in (5.12) and dropping the term not involving h, we obtain the quadratic approximation at h^*

$$-\frac{1}{n-1}\sum_{j\neq i} h(X_j) + \mu_{\eta^*}(h) - V_{\eta^*}(h^*, h) + \frac{1}{2}V_{\eta^*}(h) + \frac{\lambda}{2}J(h). \quad (5.14)$$

Set $\eta^* = \eta(x;\boldsymbol{\theta}, \hat{h}_\theta)$, and write $\breve{\boldsymbol{\xi}} = (\boldsymbol{\phi}^T, \boldsymbol{\xi}^T)^T$ and $\breve{\boldsymbol{c}} = (\boldsymbol{d}^T, \boldsymbol{c}^T)^T$. Rewrite (5.10) as

$$H\breve{\boldsymbol{c}} = \breve{R}^T\mathbf{1}/n + \boldsymbol{g}, \quad (5.15)$$

where $H = V_{\tilde{\eta}}(\breve{\boldsymbol{\xi}}, \breve{\boldsymbol{\xi}}^T) + \mathrm{diag}(0, \lambda Q)$, $\breve{R}^T = (\breve{\boldsymbol{\xi}}(X_1), \ldots, \breve{\boldsymbol{\xi}}(X_n)) = (S, R)^T$, and $\boldsymbol{g} = V_{\tilde{\eta}}(\breve{\boldsymbol{\xi}}, \tilde{\eta}) - \mu_{\tilde{\eta}}(\breve{\boldsymbol{\xi}})$. The minimizer of (5.14) has the coefficient

$$\breve{\boldsymbol{c}}^{[i]} = H^{-1}\left(\frac{\breve{R}^T\mathbf{1} - \breve{\boldsymbol{\xi}}(X_i)}{n-1} + \boldsymbol{g}\right) = \breve{\boldsymbol{c}} + \frac{H^{-1}\breve{R}^T\mathbf{1}}{n(n-1)} - \frac{H^{-1}\breve{\boldsymbol{\xi}}(X_i)}{n-1}, \quad (5.16)$$

therefore

$$h_\theta^{[i]} = \breve{\boldsymbol{\xi}}^T(X_i)\breve{\boldsymbol{c}}^{[i]} = \breve{\boldsymbol{\xi}}^T(X_i)\breve{\boldsymbol{c}} - \frac{1}{n-1}\breve{\boldsymbol{\xi}}^T(X_i)H^{-1}(\breve{\boldsymbol{\xi}}(X_i) - \breve{R}^T\mathbf{1}/n). \quad (5.17)$$

Notice that $\breve{R}^T\mathbf{1}/n = n^{-1}\sum_{i=1}^n \breve{\boldsymbol{\xi}}(X_i)$. We have the cross-validation estimate of

$\mu_{\eta(x;\boldsymbol{\theta},h)}(\ \eta(x;\boldsymbol{\theta},\hat{h}_\theta))$,

$$\hat{\mu}_{\eta(x;\boldsymbol{\theta},h)}(\eta(x;\boldsymbol{\theta},\hat{h}_\theta)) = \frac{1}{n}\sum_{i=1}^{n}\eta^{[i]}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\eta(X_i;\boldsymbol{\theta},\hat{h}_\theta) - \frac{\mathrm{tr}(P_{\mathbf{1}}^{\perp}\breve{R}H^{-1}\breve{R}^T P_{\mathbf{1}}^{\perp})}{n(n-1)}, \quad (5.18)$$

where $\eta^{[i]}(X_i) = \eta(X_i;\boldsymbol{\theta},h_\theta^{[i]})$, $P_{\mathbf{1}}^{\perp} = I - \mathbf{1}\mathbf{1}^T/n$, and the corresponding estimate of the relative Kullback-Leibler distance,

$$V(\lambda,\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\eta(X_i;\boldsymbol{\theta},\hat{h}_\theta) + \log\int e^{\eta(x;\boldsymbol{\theta},\hat{h}_\theta)}dx + \alpha\frac{\mathrm{tr}(P_{\mathbf{1}}^{\perp}\breve{R}H^{-1}\breve{R}^T P_{\mathbf{1}}^{\perp})}{n(n-1)}, \quad (5.19)$$

where $\alpha = 1$. We may set a larger α (e.g. $\alpha = 1.4$ as in Gu [11]) to prevent under-smoothing.

With fixed $\boldsymbol{\theta}$, minimizing (5) is equivalent to minimizing the following penalized weighted likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{h(X_i) - \log\int w(x)e^{h(x)}dx\right\} + \frac{\lambda}{2}J(h), \quad (5.20)$$

where $w(x) = \exp(\alpha(x,\boldsymbol{\theta}))$. Therefore, the estimate of h can be calculated by calling the *ssden* function in R package *gss* with cross-validation estimate of λ and option *bias* to specify the weights.

Plugging $\hat{h}_\theta$ into (5.5), we have the profiled penalized likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\{\alpha(X_i;\boldsymbol{\theta}) + \hat{h}_\theta(X_i)\} + \log\int\exp\{\alpha(x;\boldsymbol{\theta}) + \hat{h}_\theta(x)\}dx + \frac{\lambda}{2}J(\hat{h}_\theta). \quad (5.21)$$

Then the estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is the minimizer of (5.21). The minimization is achieved by the line search algorithm in Nelder and Mead [61]. The final estimate of h is $\hat{h}_{\hat{\theta}}$.

5.3.2 Extended Gauss-Newton procedure for the general semiparametric density model

We will extend the Gauss-Newton procedure for the estimation of $\boldsymbol{\theta}$ and h for the general semiparametric density model (5.1). Let $\boldsymbol{\theta}_-$ and h_- be the current estimates of $\boldsymbol{\theta}$ and h. We assume that the Fréchet derivative of η with respect to h at $\boldsymbol{\theta}_-$ and h_-, $D\eta(x;\boldsymbol{\theta},h)|_{h=h_-,\boldsymbol{\theta}=\boldsymbol{\theta}_-} \equiv \mathcal{L}_x$, exists and is bounded, and the partial derivative of η with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_-$ and h_-, $\frac{\partial\eta(x;\boldsymbol{\theta},h)}{\partial\boldsymbol{\theta}}|_{h=h_-,\boldsymbol{\theta}=\boldsymbol{\theta}_-} \equiv \boldsymbol{\alpha}_-(x)$, exists. Approximating η by its first order Taylor expansion at $\boldsymbol{\theta}_-$ and h_-, we have

$$\begin{aligned}\eta(x;\boldsymbol{\theta},h) &\approx \eta(x;\boldsymbol{\theta}_-,h_-) + \mathcal{L}_x(h-h_-) + \boldsymbol{\alpha}_-^T(x)(\boldsymbol{\theta}-\boldsymbol{\theta}_-)\\ &\equiv \boldsymbol{\alpha}_-^T(x)\boldsymbol{\theta} + \mathcal{L}_x h + r(x;\boldsymbol{\theta}_-,h_-),\end{aligned} \tag{5.22}$$

where $r(x;\boldsymbol{\theta}_-,h_-) = \eta(x;\boldsymbol{\theta}_-,h_-) - \boldsymbol{\alpha}_-^T(x)\boldsymbol{\theta}_- - \mathcal{L}_x h_-$. Thus, we approximate the η function by $\boldsymbol{\alpha}_-^T(x)\boldsymbol{\theta}_- + \mathcal{L}_x h + r(x;\boldsymbol{\theta}_-,h_-)$ which is linear in $\boldsymbol{\theta}$ and h as in model (5.2) with fixed $r(x;\boldsymbol{\theta}_-,h_-)$. We can set $w(x) = \exp(r(x;\boldsymbol{\theta}_-,h_-))$ as a weight function to update the estimates of $\boldsymbol{\theta}$ and h by minimizing the following penalized weighted likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\alpha}_-^T(X_i)\boldsymbol{\theta} + \mathcal{L}_{X_i}h - \log\int_{\mathcal{X}} w(x)e^{\boldsymbol{\alpha}_-^T(x)\boldsymbol{\theta}+\mathcal{L}_x h}dx\right\} + \frac{\lambda}{2}J(h). \tag{5.23}$$

The estimation procedure in Section 5.3.1 can be extended to the case that involves bounded linear functionals (Wang [28]). The following algorithm summarizes the whole procedure for the general model (1).

Algorithm 1

1. Provide initial values $\boldsymbol{\theta}_0$ and h_0 for the parameter $\boldsymbol{\theta}$ and function h.

2. At iteration k, based on current estimates $\boldsymbol{\theta}_k$ and h_k, derive $\boldsymbol{\alpha}_k(x)$, $\mathcal{L}_x$ and $r(x;\boldsymbol{\theta}_k, h_k)$. Then update $\boldsymbol{\theta}$ and h by applying the estimation procedure for linear case to solve (5.23) with the weight function $w(x) = \exp(r(x;\boldsymbol{\theta}_k, h_k))$.

3. Repeat step 2 until convergence.

5.3.3 Backfitting procedure for the transformation density model

Assume that $\eta(x;\boldsymbol{\theta}, h) = h(t(x;\boldsymbol{\theta}))$, where $t(x;\boldsymbol{\theta})$ is a known invertible and differentiable function. We propose the following backfitting procedure for the transformation model.

Algorithm 2

1. Provide initial value $\boldsymbol{\theta}_0$ for the parameter $\boldsymbol{\theta}$.

2. (a) At iteration k, based on current estimates $\boldsymbol{\theta}_k$, transform the data using $Y = t(X;\boldsymbol{\theta}_k)$ and we have $\eta_Y(y) = h(y) - \log(|t'(t^{-1}(y;\boldsymbol{\theta}_k))|)$. With the transformed data, update h by minimizing the penalized likelihood

$$-\sum_{i=1}^{n} h(y_i) + \log \int \exp\{h(y)\}\, w(y)dy + \frac{\lambda}{2} J(h) \tag{5.24}$$

with the weight function $w(y) = 1/(|t'(t^{-1}(y;\boldsymbol{\theta}_k))|)$, where λ is selected by the cross-validation method (Gu [11]). The minimization problem (5.24) is solved using the *ssden* function with *bias* option to specify the weight function. Denote the updated estimate as h_{k+1}.

(b) Update $\boldsymbol{\theta}$ as the MLE based on the likelihood of $X_1, ..., X_n$ with $h = h_{k+1}$.

3. Repeat step 2 until convergence.

5.3.4 Estimation for two-sample density models

The two sample location-scale family distributions were studied in Potgieter and Lombard [62]. They considered the nonparametric estimation of the parameters based on asymptotic likelihood, and showed that the estimators are often near optimal when compared to fully parametric methods. We consider the following two-sample problem. Suppose that $X_1, ..., X_{n_1} \overset{iid}{\sim} f_1$ and $Y_1, ..., Y_{n_2} \overset{iid}{\sim} f_2$, and X and Y have the same density after certain transformation. Specifically, $Y_i^* = t(Y_i; \boldsymbol{\theta}) \overset{iid}{\sim} f_1$, where t is a differentiable and invertible transformation of Y with unknown parameters $\boldsymbol{\theta}$. Then we have a semiparametric model with parameters $\boldsymbol{\theta}$ and nonparametric function h, where h is the logistic transformation of f_1. Fixing $\boldsymbol{\theta}$, transform the second sample with $Y_i^* = t(Y_i; \boldsymbol{\theta})$. Then the loglikelihood for the concatenated sample $\boldsymbol{Z} = (X_1, ...X_{n_1}, Y_1^*, .., Y_{n_2}^*)$ does not depend on $\boldsymbol{\theta}$. Estimate the density for $\boldsymbol{Z}$ using smoothing spline to obtain $\hat{h}(\cdot)$, where the smoothing parameter λ is selected through minimizing the cross-validated relative KL. Plugging $\hat{h}$ back, we obtain the penalized profiled likelihood function,

$$\begin{aligned} pl(\boldsymbol{\theta}) &= -\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{h}(X_i) + \log\int_{\mathcal{X}} \exp\left\{\hat{h}(x)\right\} dx \\ &\quad -\frac{1}{n_2}\sum_{j=1}^{n_2}\left\{\hat{h}\left(t(Y_i;\boldsymbol{\theta})\right) + \log(|t'(Y_i;\boldsymbol{\theta})|)\right\} \\ &\quad + \log\int_{\mathcal{Y}} \exp\left\{\hat{h}\left(t(y;\boldsymbol{\theta})\right) + \log(|t'(y;\boldsymbol{\theta})|)\right\} dy + \frac{\lambda}{2}J(\hat{h}). \end{aligned} \tag{5.25}$$

Minimize the penalized profiled likelihood with respect to $\boldsymbol{\theta}$ to obtain $\hat{\boldsymbol{\theta}}$, and the final estimate of h is $\hat{h}_{\hat{\theta}}$.

5.4 Joint Consistency and Asymptotic Normality

In this section, we develop some theoretical properties of the semiparametric estimation for the linear case in (5.2).

5.4.1 Notations and penalized likelihood estimation

Let the parameter space for $\eta \equiv (\boldsymbol{\theta}, h)$ be $\mathcal{Q} \equiv \mathbb{R}^p \times \mathcal{S}$, where $\mathcal{S}$ is a subspace of $W_{20}^m(\mathbb{I}) = W_2^m(\mathbb{I}) \ominus \{1\}$ and assumed to be an RKHS, and $W_2^m(\mathbb{I})$ is the mth order Sobolev space on $\mathbb{I} = [0, 1]$ defined as

$$W_2^m(\mathbb{I}) \equiv \{h : \mathbb{I} \mapsto \mathbb{R} | h^{(j)} \text{ is absolutely continuous for } j = 0, 1, ..., m-1 \text{ and } h^{(m)} \in L_2(\mathbb{I})\}. \tag{5.26}$$

With some abuse of notation, we use η to denote both the unknown parameters, $\eta = (\boldsymbol{\theta}, h)$, and the logistic density function $\eta(x; \boldsymbol{\theta}, h) \in \mathcal{F} \equiv \{\eta(x) = \boldsymbol{\alpha}(x)^T\boldsymbol{\theta} + h(x) : (\boldsymbol{\theta}, h) \in \mathcal{Q}, x \in \mathbb{I}\}$, where $\boldsymbol{\alpha}(x) = (\alpha_1(x), ..., \alpha_p(x))^T$ is a vector of bounded functions of $x \in \mathbb{I}$. We remove constant functions from $W_2^m(\mathbb{I})$ to make the logistic transformation one-to-one (Gu, 2013). For the same reason, none of the element of $\boldsymbol{\alpha}(x)$ is a constant. The choice of space $\mathcal{S}$ depends on $\boldsymbol{\alpha}$. For example, if $p = 1$ and $\alpha(x) = x$, we need to remove linear functions from $W_{20}^m(\mathbb{I})$ when $m > 1$. If $p = 2$ and $\boldsymbol{\alpha}(x) = (x, x^2)^T$, then we need to remove linear function from $W_{20}^m(\mathbb{I})$ when $m = 2$. When $m > 2$ we need to remove both linear and quadratic functions from $W_{20}^m(\mathbb{I})$. In general, any components of $\boldsymbol{\alpha}$ that belong to the null space of $W_{20}^m(\mathbb{I})$ are removed for identifiability (Theorem 2.9 of Gu 2013).
We will consider $l(a; x) = \log f(x; a)$ as a general function. Let

$$l_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} \eta(X_i; \boldsymbol{\theta}, h(X_i)) - \log \int_{\chi} e^{\eta(x; \boldsymbol{\theta}, h(x))} dx \tag{5.27}$$

be the loglikelihood function. The penalized semiparametric estimator is the maximizer of the penalized likelihood

$$l_{n,\lambda}(\eta) = l_n(\eta) - (\lambda/2)J(h,h), \tag{5.28}$$

where the penalty $J(h,h)$ is a squared (semi) norm, and λ is the smoothing parameter. Write $\hat{\eta}_{n,\lambda} = (\hat{\boldsymbol{\theta}}_{n,\lambda}, \hat{h}_{n,\lambda})$ as the minimizer of (5.28).

Assume the maximizer of the loglikelihood function exists in $\mathcal{S}_0 \equiv \{h : J(h) = 0\}$. To guarantee the existence and uniqueness of $\boldsymbol{\theta}$ and h, we need the loglikelihood function is concave with respect to $\boldsymbol{\theta}$ and h (Theorem 2.9 of Gu [11]). This is true when $\eta(x;\boldsymbol{\theta},h)$ is linear in $\boldsymbol{\theta}$ and h. When $\eta(x;\boldsymbol{\theta},h)$ is a nonlinear function of $\boldsymbol{\theta}$ and h, the following lemma establishes the existence and uniqueness of the maximizer to (5.28) under some conditions.

Lemma 1 *The loglikelihood l_n is concave for $\boldsymbol{\theta}$ and h when $l_n(\eta)$ is decreasing with η and η is convex with respect to $\boldsymbol{\theta}$ and h,or $l_n(\eta)$ is increasing with η and η is concave with respect to $\boldsymbol{\theta}$ and h.*

Proof: Here we only prove the first case when $l_n(\eta)$ is increasing with η and η is convex with respect to $\boldsymbol{\theta}$ and h. The proof for the second case is similar.

Since η is convex with respect to $\boldsymbol{\theta}$ and h, for any $p, q > 0$, $p + q = 1$ and $h_1, h_2 \in \mathcal{S}$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$, we have

$$\eta(x; p(\boldsymbol{\theta}_1, h_1) + q(\boldsymbol{\theta}_2, h_2)) \leq p\eta(x;\boldsymbol{\theta}_1, h_1) + q\eta(x;\boldsymbol{\theta}_2, h_2).$$

Then as l_n is decreasing with η, we have

$$l_n(\eta(x; p(\boldsymbol{\theta}_1, h_1) + q(\boldsymbol{\theta}_2, h_2)) \geq l_n(p\eta(x;\boldsymbol{\theta}_1, h_1) + q\eta(x;\boldsymbol{\theta}_2, h_2)).$$

By Holder's inequality, for $\eta_1, \eta_2 \in \mathcal{Q}$,

$$\log \int_0^1 e^{p\eta_1 + q\eta_2} \leq p \log \int_0^1 e^{\eta_1} dx + q \log \int_0^1 e^{\eta_2} dx.$$

Therefore,

$$l_n(p\eta(\boldsymbol{\theta}_1, h_1; x) + q\eta(\boldsymbol{\theta}_2, h_2; x)) \geq p l_n(\eta(\boldsymbol{\theta}_1, h_1; x)) + q l_n(\eta(\boldsymbol{\theta}_2, h_2; x)).$$

Thus, $l_n(\eta(x; \boldsymbol{\theta}, h))$ is jointly concave with respect to $\boldsymbol{\theta}$ and h. ■

5.4.2 Construction of inner product and representers

In this section, we adapt the RKHS framework to our semiparametric setup. Let $\eta_0 = (\boldsymbol{\theta}_0, h_0)$ be the true set of parameters. For simplicity, we denote $V_{\eta_0}(\cdot, \cdot)$ and $\mu_{\eta_0}(\cdot)$ as $V(\cdot, \cdot)$ and $\mu(\cdot)$ respectively. Note that V is a quadratic functional that defines an interpretable metric so that a small $V(\hat{\eta}, \eta)$ indicates that $\hat{\eta}$ is a good estimate of η. Denote $V(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ as a $p \times p$ matrix with the (i, j)th entry $V(\alpha_i, \alpha_j)$, and $V(\boldsymbol{\alpha}, h)$ as a p-vector with the ith entry $V(\alpha_i, h)$. For any $(\boldsymbol{\theta}, h), (\tilde{\boldsymbol{\theta}}, \tilde{h}) \in \mathcal{Q}$, we define the inner product on $\mathcal{Q}$ as

$$\langle \eta, \tilde{\eta} \rangle \quad = \quad \langle (\boldsymbol{\theta}, h), (\tilde{\boldsymbol{\theta}}, \tilde{h}) \rangle = V(\eta, \tilde{\eta}) + \lambda J(h, \tilde{h}), \tag{5.29}$$

whose validity is trivial, and the norm as

$$\|(\boldsymbol{\theta}, h)\|^2 = \langle (\boldsymbol{\theta}, h), (\boldsymbol{\theta}, h) \rangle. \tag{5.30}$$

Under this norm and for any $x \in [0,1]$, we will find expressions of a vector $R_x \in \mathcal{Q}$ and a linear operator $P_\lambda : \mathcal{Q} \mapsto \mathcal{Q}$ such that

$$\langle R_x, \eta \rangle = \boldsymbol{\alpha}(x)^T \boldsymbol{\theta} + h(x) \quad \text{for any } \eta \in \mathcal{Q} \tag{5.31}$$

and

$$\langle P_\lambda \eta, \tilde{\eta} \rangle = \lambda J(h, \tilde{h}) \quad \text{for any } \eta = (\boldsymbol{\theta}, h) \text{ and } \tilde{\eta} = (\tilde{\boldsymbol{\theta}}, \tilde{h}) \in \mathcal{Q}. \tag{5.32}$$

Let K_x be the RK of $\mathcal{S}$ endowed with the inner product $\langle h, \tilde{h} \rangle_1 = V(h, \tilde{h}) + \langle W_\lambda h, \tilde{h} \rangle_1$. We have $\langle K_x, h \rangle_1 = h(x)$. Let $W_\lambda : \mathcal{S} \mapsto \mathcal{S}$ be a nonnegative definite self-adjoint operator satisfying $\langle W_\lambda h, \tilde{h} \rangle_1 = \lambda J(h, \tilde{h})$ for any $h, \tilde{h} \in \mathcal{S}$, where the existence of W_λ can be shown similarly as Lemma S.2 in the supplement material of Cheng and Shang [63]. We have $\|h\|_1^2 = \langle h, h \rangle_1$.

For $a_\nu > 0$ and $b_\nu > 0$, we denote $a_\nu \asymp b_\nu$ if $a_\nu / b_\nu \to c$ as $\nu \to \infty$, where $c > 0$. Denote the sup-norm of $h \in \mathcal{S}$ as $\|h\|_{\sup} = \sup_{x \in \mathbb{I}} |h(x)|$. We denote $\mathbb{N}$ as the set of natural numbers.

The following assumption about the eigensystem in the space $\mathcal{S}$ is standard in the smoothing spline literature.

Assumption 1 *There exists a nondecreasing real sequence $\gamma_\nu \asymp \nu^{2m}$, and a sequence of real-valued functions $b_\nu \in \mathcal{S}$, $\nu \in \mathbb{N}$ satisfying $\sup_{\nu \in \mathbb{N}} \|b_\nu\|_{\sup} < \infty$ such that $V(b_\mu, b_\nu) = \delta_{\mu\nu}$ and $J(b_\mu, b_\nu) = \gamma_\mu \delta_{\mu\nu}$ for any $\mu, \nu \in \mathbb{N}$, where $\delta_{\mu\nu}$ is the Kronecker's delta. Furthermore, any $h \in \mathcal{S}$ can be expressed as a Fourier expansion $h = \sum_\nu V(h, b_\nu) b_\nu$ under the $\|\cdot\|$-norm defined in* (5.30).

Proposition 1 *Under Assumption 1, $\|h\|_1, W_\lambda b_\nu(\cdot)$ and $K_x(\cdot)$ have the following explicit*

expressions

$$\|h\|_1^2 = \sum_\nu |V(h,b_\nu)|^2(1+\lambda\gamma_\nu), \tag{5.33}$$

$$W_\lambda b_\nu(\cdot) = \frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu} b_\nu(\cdot), \tag{5.34}$$

$$K_x(\cdot) = \sum_\nu \frac{b_\nu(x)}{1+\lambda\gamma_\nu} b_\nu(\cdot). \tag{5.35}$$

Proof: Write $h = \sum_\nu V(h,b_\nu)b_\nu$. We have

$$\begin{aligned}
\|h\|_1^2 &= V(h,h) + \lambda J(h,h) \\
&= V(\sum_\nu V(h,b_\nu)b_\nu, \sum_\mu V(h,b_\mu)b_\mu) + \lambda J(\sum_\nu V(h,b_\nu)b_\nu, \sum_\mu V(h,b_\mu)b_\mu) \\
&= \sum_\nu\sum_\mu V(h,b_\nu)V(h,b_\mu)V(b_\nu,b_\mu) + \lambda\sum_\nu\sum_\mu V(h,b_\nu)V(h,b_\mu)J(b_\nu,b_\mu) \\
&= \sum_\nu V(h,b_\nu)V(h,b_\nu) + \lambda\sum_\nu V(h,b_\nu)V(h,b_\nu)\gamma_\nu \\
&= \sum_\nu |V(h,b_\nu)|^2(1+\lambda\gamma_\nu).
\end{aligned}$$

Proofs for (5.34) and (5.35) are similar, so we show the proof for (5.34) only. For any function $h \in \mathcal{S}$, the coefficient of basis function b_ν is $V(h,b_\nu)$. Therefore, the coefficient of b_ν for function $W_\lambda b_\nu$ is

$$\begin{aligned}
V(W_\lambda b_\nu, b_\nu) &= \langle W_\lambda b_\nu, b_\nu\rangle_1 - \lambda J(W_\lambda b_\nu, b_\nu) \\
&= \lambda J(b_\nu,b_\nu) - \lambda J(\sum_\mu V(W_\lambda b_\nu, b_\mu)b_\mu, b_\nu) \\
&= \lambda\gamma_\nu - \lambda\sum_\mu V(W_\lambda b_\nu, b_\mu)J(b_\mu,b_\nu) \\
&= \lambda\gamma_\nu - \lambda\gamma_\nu V(W_\lambda b_\nu, b_\nu).
\end{aligned}$$

Thus, $V(W_\lambda b_\nu, b_\nu) = \lambda\gamma_\nu/(1+\lambda\gamma_\nu)$, and $W_\lambda b_\nu(\cdot) = \frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu} b_\nu(\cdot)$. ■

For $k = 1, ..., p$, we denote $V(\alpha_k, h) \equiv \mathcal{A}_k h$. By Assumption 2, α_k has a finite second moment, so we have $|\mathcal{A}_k h| \leq V^{\frac{1}{2}}(\alpha_k, \alpha_k) V^{\frac{1}{2}}(h, h) \leq V^{\frac{1}{2}}(\alpha_k, \alpha_k) \|h\|_1 < \infty$. Therefore, $\mathcal{A}_k$ is linear and bounded. By Rieszs representation theorem, there exists an $A_k \in \mathcal{S}$ such that $\mathcal{A}_k h = \langle A_k, h \rangle_1$ for any $h \in \mathcal{S}$. Then by Assumption 1, following similar arguments as the proof for (5.34), we have

$$A_k(\cdot) = \sum_{\nu} \frac{V(\alpha_k, b_\nu)}{1 + \lambda \gamma_\nu} b_\nu(\cdot), \quad k = 1, ..., p. \tag{5.36}$$

Denote $A = (A_1, ..., A_p)^T$. Note that $V(\boldsymbol{\alpha}, \boldsymbol{\alpha}) - V(A, \boldsymbol{\alpha})$ is a matrix with the (i, j)th entry $V(\alpha_i, \alpha_j) - V(A_i, \alpha_j)$. We have

$$\begin{aligned} V(\boldsymbol{\alpha}, \boldsymbol{\alpha}) - V(A, \boldsymbol{\alpha}) &= V(\boldsymbol{\alpha} - A, \boldsymbol{\alpha} - A) + V(A, \boldsymbol{\alpha}) - V(A, A) \\ &= V(\boldsymbol{\alpha} - A, \boldsymbol{\alpha} - A) + \langle A, A \rangle_1 - V(A, A) \\ &= V(\boldsymbol{\alpha} - A, \boldsymbol{\alpha} - A) + \langle W_\lambda A, A \rangle_1 \\ &\equiv \Omega + \Sigma, \end{aligned} \tag{5.37}$$

where $\Omega = V(\boldsymbol{\alpha} - A, \boldsymbol{\alpha} - A)$ and $\Sigma = \langle W_\lambda A, A \rangle_1$ are $p \times p$ matrices. It is easy to show that Σ is negligible as λ approaches to 0. The following Assumption 2 is a regularity condition similar to Assumption A3 in Cheng and Shang [63].

Assumption 2 *For $k = 1, ..., p$, $\alpha_k(x)$ is a bounded function and $\mu(\alpha_k^2) < \infty$. The matrix $\Omega \equiv V(\boldsymbol{\alpha} - A, \boldsymbol{\alpha} - A)$ is positive definite.*

Under Assumption 2, $(\Omega + \Sigma)^{-1}$ exists for small λ. Denote by $id \in \mathcal{S}$ the identity function, that is $id(x) = x$ for any $x \in \mathbb{I}$. Now we are able to construct R_x and P_λ.

Proposition 2 *Under Assumption 2, R_x can be represented as (T_x, H_x), where*

$$\begin{aligned} T_x &= (\Omega + \Sigma)^{-1}(\boldsymbol{\alpha}(x) - A(x)), \\ H_x &= K_x - A^T (\Omega + \Sigma)^{-1} (\boldsymbol{\alpha}(x) - A(x)). \end{aligned} \tag{5.38}$$

Proof: For any $(\boldsymbol{\theta}, h) \in \mathcal{Q}$, we have

$$\begin{aligned} & \langle (T_x, H_x), \eta \rangle \\ = \; & V\left(\boldsymbol{\alpha}(x)^T T_x + H_x(X), \boldsymbol{\alpha}(x)^T \boldsymbol{\theta} + h(X)\right) + \lambda J(H_x, h) \\ = \; & T_x^T V\left(\boldsymbol{\alpha}(X), \boldsymbol{\alpha}(X)\right) \boldsymbol{\theta} + T_x^T V\left(\boldsymbol{\alpha}(X), h(X)\right) \\ & + V\left(H_x(X), \boldsymbol{\alpha}(X)\right) \boldsymbol{\theta} + V(H_x(X), h(X)) + \lambda J(H_x, h) \\ = \; & T_x^T \left[V\left(\boldsymbol{\alpha}(X), \boldsymbol{\alpha}(X)\right) + V\left(H_x(X), \boldsymbol{\alpha}(X)\right)\right] \boldsymbol{\theta} + T_x^T V\left(\boldsymbol{\alpha}(X), h(X)\right) + \langle H_x, h \rangle_1 \\ = \; & T_x^T \left[V\left(\boldsymbol{\alpha}(X), \boldsymbol{\alpha}(X)\right) + V\left(H_x(X), \boldsymbol{\alpha}(X)\right)\right] \boldsymbol{\theta} + T_x^T \langle A, h \rangle_1 + \langle H_x, h \rangle_1 \\ = \; & \boldsymbol{\alpha}(x)^T \boldsymbol{\theta} + \langle K_x, h \rangle_1, \end{aligned}$$

where the last equality is based on the definition of R_x. Therefore,

$$\begin{cases} T_x^T V\left(\boldsymbol{\alpha}(X), \boldsymbol{\alpha}(X)\right) + V\left(H_x(X), \boldsymbol{\alpha}(X)\right) = \boldsymbol{\alpha}(x), \\ \langle T_x^T A + H_x, h \rangle_1 = \langle K_x, h \rangle_1. \end{cases} \tag{5.39}$$

Plugging $H_x = K_x - T_x^T A$ in the first equation of (5.39), we have

$$\boldsymbol{\alpha}(x) = (\Omega + \Sigma) T_x + A(x). \tag{5.40}$$

Thus we have the results in (5.38). ■

Lemma 2 *There exists a constant $c_m > 0$ such that for any $x \in \mathbb{I}$ and $(\boldsymbol{\theta}, h) \in \mathcal{Q}$,*

$\|R_x\| \leq c_m \lambda^{-\frac{1}{4m}}$, *and* $\|\eta(x)\|_{\sup} \leq c_m \lambda^{-\frac{1}{4m}} \|\eta\|$.

Proof: By definition of R_x

$$\begin{aligned}
\langle R_x, R_x \rangle &= \langle R_x, (T_x, H_x) \rangle \\
&= \boldsymbol{\alpha}(x)^T T_x + H_x(x) \\
&= \boldsymbol{\alpha}(x)^T (\Omega + \Sigma)^{-1} (\boldsymbol{\alpha}(x) - A(x)) + K(x,x) - A(x)^T (\Omega + \Sigma)^{-1} (\boldsymbol{\alpha}(x) - A(x)) \\
&= K(x,x) - (\boldsymbol{\alpha}(x) - A(x))^T (\Omega + \Sigma)^{-1} (\boldsymbol{\alpha}(x) - A(x)) . \qquad (5.41)
\end{aligned}$$

By boundedness of b_μs and the explicit expression of K_x in (5.35), there exists a constant c independent of x such that

$$\begin{aligned}
K(x,x) &= \sum_\nu \frac{\{b_\nu(x)\}^2}{1 + \lambda\gamma_\nu} \\
&\leq C \sum_\nu \frac{1}{1 + \lambda\gamma_\nu} \\
&\asymp C \sum_\nu \frac{1}{1 + \lambda\nu^{2m}} \\
&\leq \lambda^{-1/(2m)} \int_1^\infty \frac{2C}{1 + (\lambda^{1/(2m)}\nu)^{2m}} d(\lambda^{1/(2m)}\nu)^{2m} \\
&\leq c\lambda^{-1/(2m)},
\end{aligned}$$

where the asymptotic equality is based on Assumption 1. On the other hand, for $k = 1, ..., p$,

$$\begin{aligned}
|A_k(x)|^2 &= |V(\alpha_k, K_x)|^2 \\
&= \left| \sum_\nu V(\alpha_k, b_\nu \frac{b_\nu(x)}{1 + \lambda\gamma_\mu}) \right|^2
\end{aligned}$$

$$\begin{aligned} &\leq \sum_{\nu} |V(\alpha_k, b_\nu)|^2 \sum_{\nu} \left| \frac{b_\nu(x)}{1+\lambda\gamma_\mu} \right|^2 \\ &\leq c'\lambda^{-1/(2m)}, \end{aligned}$$

where c' is a constant independent of x. Since $\boldsymbol{\alpha}(x)$ is bounded on $\mathbb{I}$ and $\Omega + \Sigma$ is invertible, there exists a constant $c_m > 0$ independent of x such that

$$\begin{aligned} \|R_x\|^2 &\leq |K(x,x)| + |\boldsymbol{\alpha}^T(x)\,(\Omega+\Sigma)^{-1}\,\boldsymbol{\alpha}(x)| + |A^T(x)\,(\Omega+\Sigma)^{-1}\,A| + \\ &\quad 2|\boldsymbol{\alpha}^T(x)\,(\Omega+\Sigma)^{-1}\,A(x)| \\ &\leq c_m^2\lambda^{-1/(2m)}. \end{aligned}$$

Consequently, $\|R_x\| \leq c_m\lambda^{-1/(4m)}$. ■

Proposition 3 *$P_\lambda\eta = (T_h^*, H_h^*) \in \mathcal{Q}$, where*

$$\begin{aligned} T_h^* &= -(\Omega+\Sigma)^{-1}\,V(\boldsymbol{\alpha}, W_\lambda h), \\ H_h^* &= W_\lambda h + V(\boldsymbol{\alpha}, W_\lambda h)^T\,(\Omega+\Sigma)^{-1}\,A(x). \end{aligned} \tag{5.42}$$

Proof: For any $\tilde{\eta} \in \mathcal{Q}$, from $\langle P_\lambda\eta, \tilde{\eta}\rangle = \lambda J(h, \tilde{h})$,

$$\begin{aligned} \langle P_\lambda\eta, \tilde{\eta}\rangle &= \langle (T_h^*, H_h^*), \tilde{\eta}\rangle \\ &= V\left(\boldsymbol{\alpha}(x)^T T_h^* + H_h^*(X), \boldsymbol{\alpha}(x)^T\tilde{\boldsymbol{\theta}} + \tilde{h}(X)\right) + \lambda J(H_h^*, \tilde{h}) \\ &= (T_h^*)^T V(\boldsymbol{\alpha},\boldsymbol{\alpha})\tilde{\boldsymbol{\theta}} + V(H_h^*, \boldsymbol{\alpha})^T\tilde{\boldsymbol{\theta}} + (T_h^*)^T V(\boldsymbol{\alpha}, h) + V(H_h^*, h) + \lambda J(H_h^*, \tilde{h}) \\ &= \left((T_h^*)^T V(\boldsymbol{\alpha},\boldsymbol{\alpha}) + V(H_h^*, \boldsymbol{\alpha})^T\right)\tilde{\boldsymbol{\theta}} + (T_h^*)^T\,\langle A, h\rangle_1 + \langle H_h^*, h\rangle_1 \\ &= \left((T_h^*)^T V(\boldsymbol{\alpha},\boldsymbol{\alpha}) + V(H_h^*, \boldsymbol{\alpha})^T\right)\tilde{\boldsymbol{\theta}} + \langle (T_h^*)^T A + H_h^*, \tilde{h}\rangle_1. \end{aligned} \tag{5.43}$$

Therefore,

$$\begin{cases} (T_h^*)^T V(\boldsymbol{\alpha}, \boldsymbol{\alpha}) + V(H_h^*, \boldsymbol{\alpha})^T = 0, \\ (T_h^*)^T A + H_h^* = W_\lambda h. \end{cases} \tag{5.44}$$

These lead to (5.42). ■

Note that P_λ is self-adjoint and bounded because of the following inequality:

$$\begin{aligned} \|P_\lambda \eta\| &= \sup_{\|\tilde{\eta}=1\|} |\langle P_\lambda \eta, \tilde{\eta} \rangle| \\ &= \sup_{\|\tilde{\eta}=1\|} |\lambda J(h, \tilde{h})| \leq \sqrt{\lambda J(h,h)} \sup_{\|\tilde{\eta}=1\|} \sqrt{\lambda J(\tilde{h}, \tilde{h})} \leq \|\eta\|. \end{aligned} \tag{5.45}$$

5.4.3 Fréchet derivatives

Next we assume some essential conditions for the likelihood function. Let $\mathcal{I}_0$ be the range for the true function $\eta(x; \boldsymbol{\theta}_0, h_0)$, where $\boldsymbol{\theta}_0$ and h_0 are the true parameters. Assume that $\mathcal{I}_0$ is a compact interval. Denote the first-, second- and third-derivative of $l(a; x)$ with respect to a as l'_a, l''_a and l'''_a.

Assumption 3 *(a) The loglikelihood function $l(a; x)$ is three times continuously differentiable and concave with respect to a. There exists a bounded open interval $\mathcal{I} \supset \mathcal{I}_0$ and positive constants C_0 and C_1 s.t.*

$$E\left\{\exp\left(\sup_{a \in \mathcal{I}} |l''_a(a; x)|/C_0\right)\right\} \leq C_1,$$

and

$$\sup_{a \in \mathcal{I}} |l'''(a; x)| \leq C_1, \qquad a.s.$$

(b) There exists a positive constant C_2 *s.t.* $C_2^{-1} \leq I(a) \equiv E\left\{(l'(\eta_0))^2\right\} = -E\left\{l''_a(\eta_0)\right\} \leq C_2$.

(c) $\epsilon(x) \equiv l'(a;x)$ *satisfies* $E(\epsilon|\eta = \eta_0) = 0$.

Assumption 3 is similar to the Assumption A1 in Cheng and Shang (2016). The Assumption 3(a) implies the slow diverging rate $O_p(\log n)$ of $\max_{1\leq i\leq n} \sup_{a\in\mathcal{I}} |l''_a(X_i;a)|$. Assumption 3(b) imposes boundedness and positive definiteness of the Fisher information.

For later use, we derive the Fréchet derivatives of $l_{n,\lambda}(\eta)$. For $j = 1, 2, 3$, let $\Delta\eta, \Delta\eta_j \in \mathcal{Q}$. The Fréchet derivative of $l_{n,\lambda}(\eta)$ is

$$\begin{aligned} Dl_{n,\lambda}(\eta)\Delta\eta &= \frac{1}{n}\sum_{i=1}^{n} l'_a(\eta;X_i)\langle R_{X_i}, \Delta\eta\rangle - \langle P_\lambda\eta, \Delta\eta\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\langle R_{X_i}, \Delta\eta\rangle - \mu_\eta(\Delta\eta) - \langle P_\lambda\eta, \Delta\eta\rangle \\ &= \langle S_{n,\lambda}(\eta), \Delta\eta\rangle, \end{aligned} \tag{5.46}$$

where

$$S_{n,\lambda}(\eta) = S_n(\eta) - P_\lambda\eta,$$

$$S_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} l'_a(\eta;X_i)R_{X_i}.$$

Since $\hat{\eta}_{n,\lambda}$ is the estimate, $S_{n,\lambda}(\hat{\eta}_{n,\lambda}) = 0$. For the second derivative we have

$$\begin{aligned} D^2 l_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2 &= \frac{1}{n}\sum_{i=1}^{n} l''_a(\eta;X_i)\langle R_{X_i}, \Delta\eta_1\rangle\langle R_{X_i}, \Delta\eta_2\rangle - \langle P_\lambda\Delta\eta_1, \Delta\eta_2\rangle \\ &= -V_\eta(\Delta\eta_1, \Delta\eta_2) - \langle P\Delta\eta_1, \Delta\eta_2\rangle \\ &= \langle DS_{n,\lambda}(\eta)\Delta\eta_1, \Delta\eta_2\rangle, \end{aligned}$$

where

$$DS_{n,\lambda}(\eta) = l_a''(\eta; X)\langle R_X, \eta\rangle R_X.$$

Since $-V_{\eta_0}(\Delta\eta_1, \Delta\eta_2) - \langle P_\lambda \Delta\eta_1, \Delta\eta_2\rangle = -\langle \Delta\eta_1, \Delta\eta_2\rangle$, $DS_{n,\lambda}(\eta_0) = -id$. For the third derivative we have

$$\begin{aligned} D^3 l_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2\Delta\eta_3 &= \frac{1}{n}\sum_{i=1}^{n} l_a'''(\eta; X_i)\langle R_{X_i}, \Delta\eta_1\rangle\langle R_{X_i}, \Delta\eta_2\rangle\langle R_{X_i}, \Delta\eta_3\rangle \\ &= -\{\mu(\Delta\eta_1\Delta\eta_2\Delta\eta_3) - \mu(\Delta\eta_1\Delta\eta_2)\mu(\Delta\eta_3) - \mu(\Delta\eta_1\Delta\eta_3)\mu(\Delta\eta_2) \\ &\quad -\mu(\Delta\eta_2\Delta\eta_3)\mu(\Delta\eta_1) + 2\mu(\Delta\eta_1)\mu(\Delta\eta_2)\mu(\Delta\eta_3)\} \\ &= \langle D^2 S_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2, \Delta\eta_3\rangle, \end{aligned}$$

where

$$D^2 S_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2 = l_a'''(\eta, X)\langle R_X, \Delta\eta_1\rangle\langle R_X, \Delta\eta_2\rangle R_X.$$

The Fréchet derivatives of $S_{n,\lambda}$ and $DS_{n,\lambda}$ are respectively denoted as $DS_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2$ and $D^2 S_{n,\lambda}(\eta)\Delta\eta_1\Delta\eta_2\Delta\eta_3$. Define $S(\eta) = E\{S_n(\eta)\}, S_\lambda(\eta) = E(S_{n,\lambda}(\eta)) = S(\eta) - P_\lambda(\eta)$ and $DS_\lambda(\eta) = DS(\eta) - P_\lambda$, where $DS(\eta)\Delta\eta_1\Delta\eta_2 = -V_\eta(\Delta\eta_1, \Delta\eta_2)$.

5.4.4 Convergence Rate

Theorem 1 *Suppose Assumptions 1– 3 hold, and $\lambda = o(1)$ as $n \to \infty$. Then $\|\hat{\eta}_{n,\lambda} - \eta_0\| = O_p(n^{-1/2}\lambda^{-1/(4m)} + \lambda^{1/2})$.*

Proof: Recall that $\hat{\eta}_{n,\lambda}$ is the semiparametric estimates and η_0 is the true set of parameters. In this section, we show that there exists a unique element $\hat{\eta}_{n,\lambda} \in \mathcal{Q}$ satisfying $S_{n,\lambda}(\hat{\eta}_{n,\lambda}) = 0$ and $\|\hat{\eta}_{n,\lambda} - \eta_0\| = O_p(r)$ where $r = \lambda^{\frac{1}{2}} + n^{-1/2}\lambda^{-\frac{1}{4m}}$.

Define an operator $T(\eta) = \eta + S_{n,\lambda}(\eta_0 + \eta)$. The proof involves two steps. The first step

shows the operator $T(\mathbb{B}(r)) \subset \mathbb{B}(r)$, where $\mathbb{B}(r) = \{\eta \in \mathcal{Q} : \|\eta\| \leq r\}$ be a small ball of radius r. And the second step shows T is a contraction mapping on $\mathbb{B}(r)$ so that there exists a unique $\eta_\lambda^* \in \mathbb{B}(r)$ s.t. $T(\eta_\lambda^*) = \eta_\lambda^*$.

Step I. Notice that

$$\begin{aligned}
\|T(\eta)\| &= \|\eta + S_{n,\lambda}(\eta_0 + \eta)\| \\
&= \|\eta + S_{n,\lambda}(\eta_0 + \eta) - S_{n,\lambda}(\eta_0) + S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0) + S_\lambda(\eta_0)\| \\
&\leq \|\eta + S_{n,\lambda}(\eta_0 + \eta) - S_{n,\lambda}(\eta_0)\| + \|S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0)\| + \|S_\lambda(\eta_0)\| \\
&\equiv \|I_1\| + \|I_2\| + \|I_3\|. \qquad (5.47)
\end{aligned}$$

Now we deal with I_1, I_2 and I_3 one by one.

For I_3, since $S(\eta_0) = E(S_n(\eta_0)) = 0$, we have $S_\lambda(\eta_0) = -P_\lambda \eta_0$, and consequently

$$\|I_3\| = \|P_\lambda \eta_0\| \leq \sqrt{\lambda J(\eta_0, \eta_0)}. \qquad (5.48)$$

Therefore, $\|I_3\| = O(\lambda^{1/2})$.

For I_2, we have

$$\begin{aligned}
& E\left\{\|S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0)\|^2\right\} \\
= \; & E\left\{\|\frac{1}{n}\sum_{i=1}^n l'_\eta(\langle R_{X_i}, \eta_0\rangle; X_i) R_{X_i} - E(l'_\eta(\langle R_{X_i}, \eta_0\rangle; X_i) R_{X_i})\|^2\right\} \\
= \; & \frac{1}{n^2}\sum_{i=1}^n E\left\{\|l'_\eta(\langle R_{X_i}, \eta_0\rangle; X_i) R_{X_i} - E\left\{l'_\eta(\langle R_{X_i}, \eta_0\rangle; X_i) R_{X_i}\right\}\|^2\right\} \\
\leq \; & \frac{1}{n} E\left\{|l'_\eta(\langle R_{X_1}, \eta_0\rangle; X_1)|^2 \|R_{X_1}\|^2\right\}
\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{n} c_m^2 \lambda^{-\frac{1}{2m}} E\left\{|l'_\eta(\langle R_{X_1}, \eta_0\rangle; X_1)|^2\right\} \\ &\leq \frac{1}{n} c_m^2 C_2 \lambda^{-\frac{1}{2m}}, \end{aligned}$$

where the second equality results from the fact that X_i's are iid, and the last inequality is by Assumption 3(b). Therefore,

$$E\left\{\|S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0)\|^2\right\} = O(n^{-1}\lambda^{-\frac{1}{2m}}). \tag{5.49}$$

By Markov inequality, $\|I_2\| = \|S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0)\| = O_P\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{4m}}\right)$.

For I_1, by Assumption 3(a), we have

$$\begin{aligned} \|I_1\| &= \|\eta + S_{n,\lambda}(\eta_0 + \eta) - S_{n,\lambda}(\eta_0)\| \\ &= \|\eta + DS_{n,\lambda}(\eta_0)\eta + \int_0^1\int_0^1 sD^2S_{n,\lambda}(\eta_0 + ss'\eta)\eta\eta ds ds'\| \\ &= \|\int_0^1\int_0^1 sD^2S_{n,\lambda}(\eta_0 + ss'\eta)\eta\eta ds ds'\| \\ &\leq \int_0^1\int_0^1 s\|D^2S_{n,\lambda}(\eta_0 + ss'\eta)\eta\eta\| ds ds' \\ &= \int_0^1\int_0^1 s \sup_{\|\tilde\eta\|=1} |\left\langle D^2S_{n,\lambda}(\eta_0 + ss'\eta)\eta\eta, \tilde\eta\right\rangle| ds ds' \\ &= \int_0^1\int_0^1 s \sup_{\|\tilde\eta\|=1} |l'''_\eta(\eta_0 + ss'\eta)\langle R_x, \eta\rangle\langle R_x, \eta\rangle\langle R_x, \tilde\eta\rangle| ds ds' \end{aligned}$$

$$\begin{aligned} &\leq \int_0^1\int_0^1 s \sup_{\|\tilde\eta\|=1} |l'''_\eta(\eta_0 + ss'\eta)|\|R_x\|^3\|\eta\|^2\|\tilde\eta\| ds ds' \\ &\leq C_1(c_m\lambda^{-1/(4m)})^3\|\eta\|^2 \\ &= C_3\lambda^{-3/(4m)}\|\eta\|^2, \end{aligned}$$

where $C_3 = c_m C_1$, and the third equality is because $DS_{n,\lambda}(\eta_0) = -id$.

Now let $r = 2(\|I_2\| + \|I_3\|) = O_p(n^{-\frac{1}{2}}\lambda^{-\frac{1}{4m}} + \lambda^{\frac{1}{2}})$. Then for any $\eta \in \mathbb{B}(r)$, there exists a positive constant c' so that $\|\eta\| \leq C_4(n^{-\frac{1}{2}}\lambda^{-\frac{1}{4m}} + \lambda^{\frac{1}{2}})$. Therefore, as probability approaches to 1,

$$\begin{aligned}
\|T(\eta)\| &\leq \|I_1\| + \|I_2\| + \|I_3\| \\
&\leq C_3\lambda^{-3/(4m)}\|\eta\|^2 + \frac{r}{2} \\
&\leq C_3c'\lambda^{-3/(4m)}(n^{-\frac{1}{2}}\lambda^{-\frac{1}{4m}} + \lambda^{\frac{1}{2}})\|\eta\| + \frac{r}{2} \\
&= c^*(n^{-1/2}\lambda^{-1/m} + \lambda^{1/2-3/(4m)})\|\eta\| + \frac{r}{2} \\
&\leq r,
\end{aligned}$$

where $c^* = C_3c'$, and the last inequality is true when $m > 3/2$ and $n^{-1/2}\lambda^{-1/m} = o(1)$, as $n \to \infty$. Thus, $T(\mathbb{B}(r)) \subset \mathbb{B}(r)$.

Step II. We show that T is a contraction mapping. For any $\eta_1, \eta_2 \in \mathbb{B}(r)$, applying Taylor's expansion, we have

$$\begin{aligned}
&\|T(\eta_1) - T(\eta_2)\| \\
= \;& \|\eta_1 - \eta_2 + S_{n,\lambda}(\eta_0 + \eta_1) - S_{n,\lambda}(\eta_0 + \eta_2)\| \\
= \;& \|\int_0^1 \{DS_{n,\lambda}(\eta_0 + \eta_2 + s(\eta_1 - \eta_2)) - DS_{n,\lambda}(\eta_0)\}(\eta_1 - \eta_2)ds\| \\
= \;& \|\int_0^1\int_0^1 sD^2S_{n,\lambda}(\eta_0 + s'(\eta_2 + s(\eta_0 - \eta_2)))(\eta_1 - \eta_2)(\eta_2 + s(\eta_1 - \eta_2))dsds'\| \\
\leq \;& \int_0^1\int_0^1 s\|D^2S_{n,\lambda}(\eta_0 + s'(\eta_2 + s(\eta_0 - \eta_2)))(\eta_1 - \eta_2)(\eta_2 + s(\eta_1 - \eta_2))\|dsds' \\
= \;& \int_0^1\int_0^1 s \sup_{\|\tilde{\eta}\|=1} |\left\langle D^2S_{n,\lambda}(\eta_0 + s'(\eta_2 + s(\eta_0 - \eta_2)))(\eta_1 - \eta_2)(\eta_2 + s(\eta_1 - \eta_2)), \tilde{\eta}\right\rangle| dsds' \\
= \;& \int_0^1\int_0^1 s \sup_{\|\tilde{\eta}\|=1} |l'''_\eta(\eta_0 + s'(\eta_2 + s(\eta_0 - \eta_2)))\langle(\eta_1 - \eta_2), R_x\rangle\langle(\eta_2 + s(\eta_1 - \eta_2)), R_x\rangle
\end{aligned}$$

$$
\begin{aligned}
& \langle R_x, \tilde{\eta} \rangle \, |dsds' \\
\leq \; & C\|\eta_1 - \eta_2\|(\|\eta_2\| + \|\eta_1\| + \|\eta_2\|)\|R_x\|^3 \\
\leq \; & 3rCc_m^3 \lambda^{-3/(4m)} \|\eta_1 - \eta_2\| \\
< \; & 1/2\|\eta_1 - \eta_2\|,
\end{aligned}
$$

where the last inequality is because $r\lambda^{-3/(4m)} = n^{-1/2}\lambda^{-1/m} + \lambda^{1/2-3/(4m)} = o(1)$ when $m > 3/2$. Therefore, T is a contraction mapping on $\mathbb{B}(r)$. By the contraction mapping theorem (Meir and Keeler [64]), there exists a unique element $\eta'_\lambda \in \mathbb{B}(r)$ such that $T(\eta'_\lambda) = \eta'_\lambda$. Let $\hat{\eta}_{n,\lambda} = \eta_0 + \eta'_\lambda$, then we have $S_{n,\lambda}(\hat{\eta}_{n,\lambda}) = 0$ and $\|\hat{\eta}_{n,\lambda} - \eta_0\| \leq r = O(\lambda^{\frac{1}{2}} + n^{-1/2}\lambda^{-\frac{1}{4m}})$ with probability approaching 1. The proof is completed. ■

5.4.5 Asymptotic Normality

In this section, we establish the asymptotic distribution of the semiparametric estimates under some regularity conditions.

Define $\mathcal{G}_1 = \{g_1(x) = \boldsymbol{\alpha}(x)^T\boldsymbol{\theta} : x \in \mathbb{I}, \|g_1\|_{\sup} \leq 1, \theta \in R^p\}$, and $\mathcal{G}_2 = \{g_2(x) \in \mathcal{S} : \|g_2\|_{\sup} \leq 1, J(g_2, g_2) \leq c_m^{-2}\lambda^{1/(2m)-1}\}$. Let $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 \equiv \{g_1(x) + g_2(x) : g_1 \in \mathcal{G}_1 \text{ and } g_2 \in \mathcal{G}_2\}$. For any $\eta \in \mathcal{G}$, define the empirical processes $Z_n(\eta)$

$$
Z_n(\eta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\psi_n(X_i; \eta)R_{X_i} - E(\psi_n(X; \eta)R_X)] \tag{5.50}
$$

where ψ_n is a real-valued function on $\mathbb{R} \times \mathcal{G}$. It can be shown similarly to Lemma S.3 in Cheng and Shang [63], that if ψ_n satisfies the following Lipschitz continuity:

$$
|\psi_n(X; \eta) - \psi_n(X; \tilde{\eta})| \leq c_m^{-1}\lambda^{1/(4m)}\|\eta - \tilde{\eta}\|_{\sup} \quad \text{for any } \eta, \tilde{\eta} \in \mathcal{G}, \tag{5.51}
$$

then we have

$$\lim_{n\to\infty} P\left(\sum_{\eta\in\mathcal{G}} \frac{\|Z_n(\eta)\|}{\lambda^{-(2m-1)/(8m^2)}+n^{-1/2}} \le (5\log\log n)^{1/2}\right) = 1. \tag{5.52}$$

Lemma 3 *Under Assumptions 2, there exists a positive constant C_v such that for any $\eta \in \mathcal{Q}$, $\|\boldsymbol{\theta}\|_{l_2} \le C_v\|\eta\|$ where $\|\boldsymbol{\theta}\|_{l_2} \equiv \sqrt{\sum_{i=1}^{p}\theta_i^2}$.*

Proof: For any $\eta \in \mathcal{Q}$,

$$\begin{aligned}
\|\eta\|^2 &= \|(\boldsymbol{\theta}, h)\|^2 \\
&= V(\boldsymbol{\theta}^T\boldsymbol{\alpha}+h, \boldsymbol{\theta}^T\boldsymbol{\alpha}+h) + \lambda J(h,h) \\
&= V(\boldsymbol{\theta}^T(\boldsymbol{\alpha}-A)+\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^T(\boldsymbol{\alpha}-A)+\boldsymbol{\theta}^TA+h) + \lambda J(h,h) \\
&= \boldsymbol{\theta}^T V(\boldsymbol{\alpha}-A, \boldsymbol{\alpha}-A)\boldsymbol{\theta} + V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) \\
&\quad +2\boldsymbol{\theta}^T V(\boldsymbol{\alpha}-A, \boldsymbol{\theta}^TA+h) + \lambda J(h,h),
\end{aligned}$$

where

$$\begin{aligned}
& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T V(\boldsymbol{\alpha}-A, \boldsymbol{\theta}^TA+h) + \lambda J(h,h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T V(\boldsymbol{\alpha}, \boldsymbol{\theta}^TA+h) - 2\boldsymbol{\theta}^T V(A, \boldsymbol{\theta}^TA+h) + \lambda J(h,h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T \langle A, \boldsymbol{\theta}^TA+h\rangle_1 - 2\boldsymbol{\theta}^T V(A, \boldsymbol{\theta}^TA+h) + \lambda J(h,h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T \lambda J(A, \boldsymbol{\theta}^TA+h) + \lambda J(h,h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T \lambda J(A, \boldsymbol{\theta}^TA) + 2\boldsymbol{\theta}^T \lambda J(A,h) + \lambda J(h,h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + 2\boldsymbol{\theta}^T \lambda J(A, \boldsymbol{\theta}^TA) + \boldsymbol{\theta}^T \lambda J(A,h) + \lambda J(\boldsymbol{\theta}^TA+h, h) \\
= \;& V(\boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h) + \boldsymbol{\theta}^T \lambda J(A, \boldsymbol{\theta}^TA) + \lambda J(\boldsymbol{\theta}^TA, \boldsymbol{\theta}^TA+h) + \lambda J(\boldsymbol{\theta}^TA+h, h) \\
= \;& \langle \boldsymbol{\theta}^TA+h, \boldsymbol{\theta}^TA+h\rangle_1 + \lambda J(\boldsymbol{\theta}^TA, \boldsymbol{\theta}^TA).
\end{aligned}$$

Therefore,

$$\begin{aligned}\|\eta\|^2 &= \boldsymbol{\theta}^T\Omega\boldsymbol{\theta} + \langle \boldsymbol{\theta}^T A + h, \boldsymbol{\theta}^T A + h\rangle_1 + \lambda J(\boldsymbol{\theta}^T A, \boldsymbol{\theta}^T A) \\ &\geq \boldsymbol{\theta}^T\Omega\boldsymbol{\theta}.\end{aligned}$$

Since Ω is positive definite, there exists a positive constant C_v and $\|\boldsymbol{\theta}\|_{l_2} \leq C_v\|\eta\|$. ■

Theorem 2 *(Joint Bahadur representation) Suppose that Assumptions 1– 3 are satisfied,* $\lambda = o(1)$ *and* $n\lambda^{1/m} \to \infty$. *Let* $a_n = n^{-1/2}(n^{-1/2}\lambda^{-1/(4m)})\lambda^{-(6m-1)/(8m^2)}(\log\log n)^{1/2} + C_1\lambda^{-1/(4m)}(n^{-1/2}\lambda^{-1/(4m)})/\log n$ *where* C_1 *is defined in Assumption 3. Then we have*

$$\|\hat{\eta}_{n,\lambda} - \eta_0 - S_{n,\lambda}(\eta_0)\| = O_p(a_n \log n).$$

Proof: Denote $\eta = \hat{\eta}_{n,\lambda} - \eta_0 \equiv (\boldsymbol{\theta}, h)$. By Taylor expansion, and the fact that $S_{n,\lambda}(\eta + \eta_0) = 0$ and $D(S_\lambda(\eta_0)) = -id$, we have

$$\begin{aligned}&\|S_n(\eta+\eta_0) - S(\eta+\eta_0) - (S_n(\eta_0) - S(\eta_0))\| \\ =\ &\|S_{n,\lambda}(\eta+\eta_0) - S_\lambda(\eta+\eta_0) - (S_{n,\lambda}(\eta_0) - S_\lambda(\eta_0))\| \\ =\ &\|S_\lambda(\eta_0) - S_\lambda(\eta+\eta_0) - S_{n,\lambda}(\eta_0)\| \\ =\ &\|DS_{n,\lambda}(\eta_0)\eta + \int_0^1\int_0^1 sD^2S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta ds ds' + S_{n,\lambda}(\eta_0)\| \\ =\ &\|-\eta + \int_0^1\int_0^1 sD^2S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta ds ds' + S_{n,\lambda}(\eta_0)\| \\ \geq\ &\|-\eta + S_{n,\lambda}(\eta_0)\| - \|\int_0^1\int_0^1 sD^2S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta ds ds'\|.\end{aligned}$$

Thus,

$$\begin{aligned} & \|\eta - S_{n,\lambda}(\eta_0)\| \\ \leq\ & \|S_n(\eta+\eta_0) - S(\eta+\eta_0) - (S_n(\eta_0) - S(\eta_0))\| + \| \int_0^1 \int_0^1 sD^2 S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta ds ds'\| \\ \equiv\ & \|I_1\| + \|I_2\|. \end{aligned} \tag{5.53}$$

We deal with $\|I_1\|$ and $\|I_2\|$ one by one.

First, since $\|I_2\| \leq \int_0^1 \int_0^1 \|sD^2 S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta\| ds ds'$, we need to find an upper bound for $\|D^2 S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta\|$. Recall that $D^2 S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta = l_a'''(ss'\eta+\eta_0)\eta^2 R_x$. Then by Assumption 3,

$$\begin{aligned} & \|D^2 S_{n,\lambda}(ss'\eta+\eta_0)\eta\eta\| \\ =\ & \|l_a'''(ss'\eta+\eta_0)\eta^2 R_x\| \\ \leq\ & \sup_{a\in\mathcal{I}} |l_a'''(a)| \|\eta\|^2 \|R_x\| \\ \leq\ & C_1 c_m \lambda^{-1/(4m)} \|\eta\|^2, \end{aligned} \tag{5.54}$$

where C_1 can be found in Assumption 3.

Next, we find an upper bound for $\|I_1\|$. By Theorem 1, the event $B_{n1} = \{\|\eta\| \leq r_n \equiv M(n^{-1/2}\lambda^{-1/(4m)} + \lambda^{1/2})\}$ has large probability with some preselected large M.

On the other hand by Assumption 3 and Chebyshev's inequality,

$$\max_{1\leq i\leq n} \sup_{a\in\mathcal{I}} |l_a''(a; X_i)| = O_p(\log n).$$

Then we can find a sufficiently large constant $C > C_0$ such that the event

$$B_{n2} = \{\max_{1\le i\le n} \sup_{a\in\mathcal{I}} |l_a''(a; X_i)| \le C \log n\}$$

has large probability. So $B_n = B_{n1} \cap B_{n2}$ has large probability.

Define $\bar{\eta} \equiv (\bar{\boldsymbol{\theta}}, \bar{h}) = d_n^{-1}\eta/2$, where $\bar{\boldsymbol{\theta}} = d_n^{-1}\boldsymbol{\theta}/2, \bar{h} = d_n^{-1}h/2$, and $d_n = c_m r_n \lambda^{-1/4m}$. Since $\lambda = o(1)$ and $n\lambda^{1/m} \to \infty$, we have $d_n = o(1)$. Then on B_n, by Lemma 2, $\|\bar{\eta}\|_{\sup} \le c_m \lambda^{-1/(4m)} \|\bar{\eta}\| \le 1/2$, which implies for any $x \in \mathbb{I}$, $|\eta(x)| \le 1/2$. And By lemma 3, $\|\boldsymbol{\theta}\|_{l_2} \le C_v r_n$, which implies that $\|\bar{\boldsymbol{\theta}}\|_{l_2} \le C_v' \lambda^{1/(4m)}$. Furthermore since $\boldsymbol{\alpha}(x)$ is bounded, we have $|\bar{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(x)| \le C' \lambda^{1/(4m)}$ for any $x \in \mathbb{I}$. Since $\lambda = o(1)$, we may select a small λ so that $\|\bar{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(x)\|_{\sup} \le 1/2$. Consequently $|\bar{h}(x)| = |\bar{\eta}(x) - \bar{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(x)| \le \|\bar{\boldsymbol{\theta}}^T \boldsymbol{\alpha}(x)\|_{\sup} + \|\bar{\eta}\|_{\sup} \le 1$ for any $x \in \mathbb{I}$. Additionally, observe that

$$\begin{aligned} J(\bar{h}, \bar{h}) &= d_n^{-2}\lambda^{-1}(\lambda J(h,h))/4 \\ &\le d_n^{-2}\lambda^{-1}\|\eta\|^2/4 \\ &\le d_n^{-2}\lambda^{-1} r_n^2/4 \\ &\le c_m^{-2}\lambda^{1/(2m)-1}. \end{aligned}$$

Therefore, when event B_n holds, we have $\bar{\eta} \in \mathcal{G}$. Define

$$\psi(x;\eta) \equiv l_a'(\eta_0 + \eta) - l_a'(\eta_0) = \int_{\eta_0}^{\eta_0+\eta} l_a''(x; a) da.$$

It is not hard to check that

$$I_1 = \frac{1}{n}\sum_{i=1}^{n} \left[\psi(X_i;\eta) R_{X_i} - E_X\left(\psi(X;\eta) R_X\right)\right]. \tag{5.55}$$

Let $\tilde{\psi}_n(x;\bar{\eta}) = 1/2C^{-1}c_m^{-1}(\log n)^{-1}\lambda^{1/(4m)}d_n^{-1}\psi(x;2d_n\bar{\eta})$ and $\psi_n(X_i;\bar{\eta}) = \tilde{\psi}_n(X_i;\bar{\eta})I_{A_i}$, where $A_i = \{\sup_{a\in I}|l_a''(X_i;a)| \le C\log n\}$, for $i = 1, ..., n$, and $B_n \subset \cap_i A_i$.

Next we show ψ_n satisfies the Lipschitz continuity (5.51). For any $\bar{\eta}_1 = (\boldsymbol{\theta}_1, h_1), \bar{\eta}_2 = (\boldsymbol{\theta}_2, h_2) \in \mathcal{G}$ and $x \in \mathbb{I}$, when n is sufficiently large, since $\eta_0(x) \in \mathcal{I}_0$ and $d_n = o(1)$, both $\eta_0(x) + 2d_n\bar{\eta}_1(x)$ and $\eta_0(x) + 2d_n\bar{\eta}_2(x)$ are in $\mathcal{I}$. Therefore,

$$
\begin{aligned}
& |\psi_n(X_i;\bar{\eta}_1) - \psi_n(X_i;\bar{\eta}_2)| \\
= \;& \frac{1}{2}C^{-1}c_m^{-1}(\log n)^{-1}\lambda^{1/(4m)}d_n^{-1}|\psi(X_i;2d_n\bar{\eta}_1) - \psi(X_i;2d_n\bar{\eta}_2)|I_{A_i} \\
= \;& \frac{1}{2}C^{-1}c_m^{-1}(\log n)^{-1}\lambda^{1/(4m)}d_n^{-1}\left|\int_{\eta_0(X_i)}^{\eta_0(X_i)+2d_n(\bar{\eta}_1(X_i))} l_a''(X_i;a)I_{A_i}da\right. \\
& \left. - \int_{\eta_0(X_i)}^{\eta_0(X_i)+2d_n(\bar{\eta}_2(X_i))} l_a''(X_i;a)I_{A_i}da\right| \\
\le \;& C^{-1}c_m^{-1}(\log n)^{-1}\lambda^{1/(4m)}d_n^{-1}d_n\|\bar{\eta}_1 - \bar{\eta}_2\|_{\sup}\sup_{a\in\mathcal{I}}|l_a''(X_i;a)|I_{A_i} \\
\le \;& c_m^{-1}\lambda^{1/(4m)}\|\bar{\eta}_1 - \bar{\eta}_2\|_{\sup}.
\end{aligned}
$$

Then by (5.52), with large probability,

$$
||\sum[\psi_n(X_i;\bar{\eta})R_{X_i} - E\{\psi_n(X;\bar{\eta})R_X\}]|| \le (n^{\frac{1}{2}}h^{-\frac{2m-1}{4m}+1})(5\log\log n)^{\frac{1}{2}}. \tag{5.56}
$$

On the other hand, by Chebyshev's inequity and Assumption 3(a),

$$
P(A_i^c) \le \exp(-\frac{C}{C_0}\log n)E\{\exp(\sup_{a\in I}|l_a''(X_i;a)|/C_0\} \le C_1 n^{-\frac{C}{C_0}}. \tag{5.57}
$$

Since $\lambda = o(1)$ and $n\lambda^{1/m} \to \infty$, we can let C be sufficiently large so that $(\log n)^{-1}n^{-\frac{C}{2C_0}} = o(a_n'\lambda^{1/(4m)d_n^{-1}})$, where $a_n' = n^{-1/2}(n^{-1/2}\lambda^{-1/(4m)} + \lambda^{1/2})\lambda^{-(6m-1)/(8m^2)}(\log\log n)^{1/2}$. Assumption 3 implies $E\{\sup_{a\in\mathcal{I}}|l''(a;x)|\} \le 2C_1C_0^2$, so we have on B_n, $E\{|\psi(X;d_nh)|^2\} \le$

$2C_1C_0^2d_n^2$. Then when n is large, on B_n,

$$\begin{aligned}
& \|E(\psi_n(X;\bar{\eta})R_X) - E(\tilde{\psi}_n(X;\bar{\eta})R_X)\| \\
= \; & \|E(\tilde{\psi}_n(X;\bar{\eta})R_X I_{A_i^c})\| \\
\leq \; & 1/2C^{-1}(\log n)^{-1}d_n^{-1}(E\left\{\psi(X;2d_n\bar{\eta})\right\}^2)^{\frac{1}{2}}P(A_i^c)^{\frac{1}{2}} \\
\leq \; & 2C^{-1}C_0C_1(\log n)^{-1}n^{-\frac{C_1}{2C_0}} \\
= \; & O(a_n'\lambda^{1/(4m)}d_n^{-1}).
\end{aligned}$$

Thus, with $\psi_n(X_i;\eta) = \tilde{\psi}_n(X_i;\eta)$ for $i = 1, ..., n$ on B_n, there exists a large positive constant C' such that

$$\begin{aligned}
\|I_1\| &= \|\frac{1}{n}\sum_{i=1}^{n}[\psi(X_i;\eta)R_{X_i} - E\left(\psi(X;\eta)R_X\right)]\| \\
&= \frac{2Cc_m(\log n)\lambda^{-1/(4m)}d_n}{n}\|\sum_{i=1}^{n}[\tilde{\psi}(X_i;\eta)R_{X_i} - E\left(\tilde{\psi}(X;\eta)R_X\right)]\| \\
&\leq \frac{2Cc_m(\log n)\lambda^{-1/(4m)}d_n}{n}\left\{\|\sum_{i=1}^{n}[\psi(X_i;\eta)R_{X_i} - E\left(\psi(X;\eta)R_X\right)]\|\right. \\
&\quad \left. +n\|E\left(\psi(X;\eta)R_X\right) - E\left(\tilde{\psi}(X;\eta)R_X\right)\|\right\} \\
&\leq \frac{2Cc_m(\log n)\lambda^{-1/(4m)}d_n}{n}[(n^{1/2}\lambda^{-(2m-1)/(8m^2)}+1)(5\log\log n)^{1/2} + nM'a_n'\lambda^{1/(4m)}d_n^{-1}] \\
&= 2Cc_m\log n[d_n(n^{-1/2}\lambda^{-(4m-1)/(8m^2)} + n^{-1}\lambda^{-1/(4m)})(5\log\log n)^{1/2} + M'a_n'] \\
&\leq C'c_ma_n'\log n. \qquad (5.58)
\end{aligned}$$

Therefore, by (5.53), (5.54) and (5.58), with large probability

$$\|\eta - S_{n,\lambda}(\eta_0)\| \leq C'c_ma_n'\log n + C_1c_m\lambda^{-1/(4m)}(n^{-1/2}\lambda^{-1/(4m)} + \lambda^{1/2})^2. \qquad (5.59)$$

This completes the proof. ■

Assumption 4 *For any $x \in \mathbb{I}$, as $n \to \infty$,*

$$\begin{aligned}
&\lambda^{1/(2m)} V(\epsilon(X)K_x, \epsilon(X)K_x) \to \sigma_x^2, \\
&\lambda^{1/(4m)}\, Cov(\epsilon(X)K_X(x), \epsilon(X)(\boldsymbol{\alpha}(X) - A(X))) \to \boldsymbol{\zeta}_x, \\
&(\Omega + \Sigma)^{-1} V\left(\epsilon(X)(\boldsymbol{\alpha}(X) - A(X))\right)(\Omega + \Sigma)^{-1} \to \Omega^*, \\
&\lambda^{1/(4m)} A(x) \to -\boldsymbol{\tau}_x,
\end{aligned} \tag{5.60}$$

where $\boldsymbol{\zeta}_x$ and $\boldsymbol{\tau}_x$ are two p-dimensional functions of x.

Lemma 4 *Let Assumptions 1 – 4 hold. Suppose that as $n \to \infty$, $\lambda = o(1), n\lambda^{1/m} \to \infty$, and $a_n \log n = o(n^{-1/2}\lambda^{1/(4m)})$. Then we have, for any $x \in \mathbb{I}$,*

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0^*) \\ \sqrt{n\lambda^{1/(2m)}}\left(\hat{h}_{n,\lambda}(x) - h_0^*(x)\right) \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Psi^*), \tag{5.61}$$

where

$$\Psi^* = \begin{pmatrix} \Omega^* & \Omega\boldsymbol{\zeta}_x + \Omega^*\boldsymbol{\tau}_x \\ \boldsymbol{\zeta}_x^T\Omega + \boldsymbol{\tau}_x^T\Omega^* & \sigma_x^2 + 2\boldsymbol{\tau}_x^T\Omega\boldsymbol{\zeta}_x + \boldsymbol{\tau}_x^T\Omega^*\boldsymbol{\tau}_x \end{pmatrix}. \tag{5.62}$$

Proof: Define

$$\mathrm{Rem}_n = \hat{\eta}_{n,\lambda} - \eta_0^* - \frac{1}{n}\sum_{i=1}^n \epsilon(X_i)R_{X_i}, \quad \text{and} \quad \mathrm{Rem}_n^\lambda = \hat{\eta}_{n,\lambda}^\lambda - \eta_0^{*\lambda} - \frac{1}{n}\sum_{i=1}^n \epsilon(X_i)R_{X_i}^\lambda,$$

where $R_x = (T_x, H_x)$ is given in (5.38) and

$$\begin{aligned}
\eta_0^* &= \eta_0 - P_\lambda\eta_0, \qquad \eta_0^{*\lambda} = (\boldsymbol{\theta}_0^*, \lambda^{1/(4m)}h_0^*) \\
\hat{\eta}_{n,\lambda}^\lambda &= (\hat{\boldsymbol{\theta}}_{n,\lambda}, \lambda^{1/(4m)}\hat{h}_{n,\lambda}), \qquad R_x^\lambda = (T_x, \lambda^{1/(4m)}H_x).
\end{aligned}$$

By Theorem 2, $\|\text{Rem}_n\| = O_p(a_n \log n)$. Combined with lemma 3, we have

$$\|\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0^* - \frac{1}{n}\sum_{i=1}^{n} \epsilon(X_i) T_{X_i}\|_{l_2} = O_p(a_n \log n).$$

Consequently,

$$\begin{aligned}
\|\text{Rem}_n^\lambda - \lambda^{1/(4m)}\text{Rem}_n\| =& \| \left((1-\lambda^{1/(4m)})(\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0^* - \frac{1}{n}\sum_{i=1}^{n} \epsilon(X_i) T_{X_i}), 0 \right) \| \\
\leq & (1-\lambda^{1/(4m)}) \|\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0^* - \frac{1}{n}\sum_{i=1}^{n} \epsilon(X_i) T_{X_i}\|_{l_2} \\
= & O_p(a_n \log n).
\end{aligned}$$

Since by assumption $a_n \log n = o(n^{-1/2}\lambda^{1/(4m)})$, and the fact that $\lambda^{1/(4m)}\text{Rem}_n = o_p(a_n \log n)$, $\|\text{Rem}_n^\lambda\| = o_p(n^{-1/2}\lambda^{1/(4m)})$.

Next we will use Rem_n^λ to derive the target joint limit distribution of $n^{1/2}\boldsymbol{\alpha}(x)^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*) + n^{1/2}\lambda^{1/(4m)}(\hat{h}_{n,\lambda}(x) - h_0^*(x))$ with the Cramer-Wold device. For any $x \in \mathbb{I}$, it's easy to verify that

$$n^{1/2}\boldsymbol{\alpha}(x)^T(\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0^*) + n^{1/2}\lambda^{1/(4m)}(\hat{h}_{n,\lambda}(x) - h_0^*(x)) = n^{1/2}\langle R_x, \hat{\eta}_{n,\lambda}^\lambda - \eta_0^{*\lambda}\rangle.$$

With the fact that

$$\begin{aligned}
|n^{1/2}\langle R_x, \hat{\eta}_{n,\lambda}^\lambda - \eta_0^{*\lambda}\rangle| \leq & n^{1/2}\|R_x\|\|\text{Rem}_n^\lambda\| \\
= & O_p(n^{1/2}\lambda^{-1/(4m)} a_n \log n) = o_p(1),
\end{aligned}$$

we just need to find the limit distribution of $n^{1/2}\langle R_x, \hat{\eta}^{\lambda}_{n,\lambda} - \eta^{*\lambda}_0\rangle$ which is equivalent to

$$n^{1/2}\sum_{i=1}^{n}\epsilon(X_i)\left(\boldsymbol{\alpha}(x)^T T_{X_i} + \lambda^{1/(4m)} H_{X_i}(x)\right).$$

Next we employ Lindeberg's condition for Central Limit Theorem to find the limit distribution. By Assumption 3 and the representation of T_X and H_X in (5.38),

$$\begin{aligned}&\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x)\\ =&\lambda^{1/(4m)} K_X(x) + \boldsymbol{\alpha}^T(x)\Omega^{-1}\boldsymbol{\alpha}(X).\end{aligned} \tag{5.63}$$

Thus,

$$\begin{aligned}
s_n^2 &\equiv \mathrm{Var}\left(\sum_{i=1}^{n}\epsilon(X_i)(\boldsymbol{\alpha}(x)^T T_{X_i} + \lambda^{1/(4m)} H_{X_i}(x))\right)\\
&= n\mathrm{Var}\left(\epsilon(X)(\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x))\right)\\
&= n\mathrm{Var}\left(\epsilon(X)(\lambda^{1/(4m)} K_X(x) + \boldsymbol{\alpha}^T(x)\Omega^{-1}\boldsymbol{\alpha}(X))\right)\\
&= n\left\{\lambda^{1/(2m)}\mathrm{Var}(\epsilon(X)K_X(x)) + \boldsymbol{\alpha}^T(x)\Omega^{-1}\mathrm{Var}\left(\epsilon(X)\boldsymbol{\alpha}(X)\right)\Omega^{-1}\boldsymbol{\alpha}(x) + \right.\\
&\quad \left. 2\lambda^{1/(4m)}\boldsymbol{\alpha}^T(x)\Omega^{-1}\mathrm{Cov}(\epsilon(X)K_X(x), \epsilon(X)\boldsymbol{\alpha}(X))\right\}.
\end{aligned}$$

By Assumption 4, as $n \to \infty$,

$$\begin{aligned}
s_n^2/n &\to \sigma_x^2 + \boldsymbol{\alpha}^T(x)\Omega^*\boldsymbol{\alpha}(x) + 2\boldsymbol{\alpha}^T(x)\Omega\boldsymbol{\zeta}_x\\
&= (\boldsymbol{\alpha}(x), 1)^T\Psi^*(\boldsymbol{\alpha}(x), 1),
\end{aligned}$$

where Ψ^* is defined in (5.62). Since $\boldsymbol{\alpha}(x)$ is bounded, $s_n^2 \asymp n$.

By the proof of Lemma 2, $|K_X(x)| = O(\lambda^{-1/(2m)})$, and for any $x \in \mathbb{I}$ and $k = 1, ..., p$,

there exists a positive constant c_k' independent of x such that

$$\begin{aligned}|A_k(x)| =&|V(\alpha_k, K_x)|\\ =&\left|\sum_\nu V(\alpha_k, b_\nu \frac{b_\nu(x)}{1+\lambda\gamma_\mu})\right|\\ \le&\left(\sum_\nu |V(\alpha_k, b_\nu)|^2\right)^{1/2}\left(\sum_\nu \left|\frac{b_\nu(x)}{1+\lambda\gamma_\mu}\right|^2\right)^{1/2}\\ \le&c_k'\lambda^{-1/(4m)}.\end{aligned}$$

Thus, we can find a positive constant c' such that

$$|\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x)| \le c'\lambda^{-1/(4m)}, \qquad \text{a.s.}$$

Then for any $\varepsilon > 0$, by $s_n^2 \asymp n$ and the assumption $n\lambda^{1/m} \to \infty$,

$$\begin{aligned}&E\left\{|\epsilon(X)(\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x))|^2 I(|\epsilon(X)(\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x))| \ge \varepsilon s_n)\right\}\\ \le&(c'\lambda^{-1/(4m)})^2 E\{\epsilon(X)^2 I(|\epsilon(X)| \ge \varepsilon s_n \lambda^{1/(4m)}/c')\}\\ \le&(c'\lambda^{-1/(4m)})^2 (E(\epsilon(X)^4))^{1/2}(P(|\epsilon(X)| \ge \varepsilon s_n\lambda^{1/(4m)}/c'))^{1/2}\\ \le&(c'\lambda^{-1/(4m)})^2 (E(\epsilon(X)^4))^{1/2}(\varepsilon^4 s_n^4 \lambda^{1/m})^{-1/2}(E(\epsilon(X)^4))^{1/2}\\ =&\frac{(c')^2 E(\epsilon(X)^4)}{\varepsilon^2 s_n^2 \lambda^{1/m}} \to 0,\end{aligned}$$

where $E(\epsilon(X)^4)$ is assumed to exist. Then as $n \to \infty$,

$$\begin{aligned}&\frac{1}{s_n^2}\sum_{i=1}^n E\left\{|\epsilon(X_i)(\boldsymbol{\alpha}(x)^T T_{X_i} + \lambda^{1/(4m)} H_{X_i}(x))|^2 I|\epsilon(X_i)(\boldsymbol{\alpha}(x)^T T_{X_i} + \lambda^{1/(4m)} H_{X_i}(x))| \ge \varepsilon s_n\right\}\\ =&\frac{n}{s_n^2} E\left\{|\epsilon(X)(\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x))|^2 I|\epsilon(X)(\boldsymbol{\alpha}(x)^T T_X + \lambda^{1/(4m)} H_X(x))| \ge \varepsilon s_n\right\} \to 0.\end{aligned}$$

Thus, Lindeberg's condition is satisfied. Then by central limit theorem, we can have the limit distribution (5.61). This completes the proof. ■

Lemma 5 *Suppose that there exists* $b \in (1/(2m), 1]$ *such that* $\alpha_k(\cdot)$ *satisfies*

$$\sum_{\nu} |V(\alpha_k, b_\nu)|^2 \gamma_\nu^b < \infty, \qquad \textit{for any } k = 1, ..., p. \tag{5.64}$$

Then we have, for any $x \in \mathbb{I}$, $\lambda^{1/(4m)} A(x) = o(1)$ *and* $\lambda^{1/(4m)}(W_\lambda A)(x) = o(1)$. *Moreover, if* $n^{1/2}\lambda^{(1+b)/2} = o(1)$, *then as* $n \to \infty$,

$$\begin{pmatrix} \sqrt{n}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\ \sqrt{n\lambda^{1/(2m)}}\,\{h_0^*(x) - h_0(x) + W_\lambda h_0(x)\} \end{pmatrix} \longrightarrow 0. \tag{5.65}$$

Proof: By the definition of $\boldsymbol{\theta}_0^*, h_0^*$ and the representation of P_λ in (5.42), we have

$$\begin{aligned} &\begin{pmatrix} \sqrt{n}(\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_0) \\ \sqrt{n\lambda^{1/(2m)}}\,\{h_0^*(x) - h_0(x) + W_\lambda h_0(x)\} \end{pmatrix} \\ = &\begin{pmatrix} \sqrt{n}(\Omega + \Sigma_\lambda)^{-1} V(\boldsymbol{\alpha}, W_\lambda h_0) \\ -\sqrt{n\lambda^{1/(2m)}} V(\boldsymbol{\alpha}^T, W_\lambda h_0)(\Omega + \Sigma_\lambda)^{-1} A(x) \end{pmatrix}. \end{aligned} \tag{5.66}$$

Thus, we just need to show (5.66) goes to 0, which proceeds in the following two steps.

(i) Show $\|V(\boldsymbol{\alpha}, W_\lambda h_0)\|_{l_2} = o(n^{-1/2})$. By (5.34),

$$V(\alpha_k, W_\lambda h_0) = \sum_{\mu} V(\alpha_k, b_\mu) V(h_0, b_\mu) \frac{\lambda\gamma_\mu}{1 + \lambda\gamma_\mu},$$

for all $k = 1, ..., p$. Then by Cauchy's inequality, we have

$$\begin{aligned} |V(\alpha_k, W_\lambda h_0)|^2 &\leq \sum_\mu |V(\alpha_k, b_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \sum_\mu |V(h_0, b_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \\ &= J(h_0, h_0 - W_\lambda h_0)\lambda \sum_\mu |V(\alpha_k, b_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \\ &= \text{const} \cdot \lambda \sum_\mu |V(\alpha_k, b_\mu)|^2 \gamma_\mu^b \left(\frac{\lambda\gamma_\mu^{1-b}}{1+\lambda\gamma_\mu} \right) \\ &\leq \text{const} \cdot \lambda^{1+b} \end{aligned}$$

where the last inequality is by $\gamma_\mu \equiv \mu^{2m}$ and $\sum_\nu |V(\alpha_k, b_\nu)|^2 \gamma_\nu^b < \infty$. Therefore, by assumption $n^{1/2}\lambda^{(1+b)/2} = o(1)$, $\|V(\boldsymbol{\alpha}, W_\lambda h_0)\|_{l_2} \equiv \sqrt{\sum_{k=1}^p |V(\alpha_k, W_\lambda h_0)|^2} = o(n^{-1/2})$.

(ii) Show $\lambda^{1/(4m)} A(x) = o(1)$. For any $x \in \mathbb{I}$, with the explicit expression of K_x (5.35), we have

$$A_k(x) = \langle A_k, K_x \rangle_1 = V(\alpha_k, K_x) = \sum_\mu \frac{V(\alpha_k, b_\mu)}{1+\lambda\gamma_\mu} b_\mu(x).$$

By the boundedness of b_μs (Assumption 1) and Cauchy's inequality, for any $x \in \mathbb{I}$,

$$\begin{aligned} |A_k(x)|^2 &\leq \sum_\mu |V(\alpha_k, b_\mu)|^2 (1+\gamma_\mu)^b (b_\mu(x))^2 \sum_\mu \frac{1}{(1+\gamma_\mu)^b (1+\lambda\gamma_\mu)^2} \\ &= O\left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b} \right) = O(1), \end{aligned}$$

where the last equality is because $\sum_\nu |V(\alpha_k, b_\nu|^2 \gamma_\nu^b < \infty$ and $\sum_\mu \frac{1}{(1+\lambda\gamma_\mu)^2}$ converges. Therefore, we have $\|A_k\|_{\sup} = O(1)$, furthermore $\lambda^{1/(4m)} A(x) = o(1)$. On the other

hand,

$$\begin{aligned} & \lambda^{1/(4m)}\mathrm{Cov}(\epsilon(X)K_X(x), \epsilon(X)(\boldsymbol{\alpha}(X) - A(X))) \\ = \; & \lambda^{1/(4m)}V(\epsilon(X)K_X(x), \epsilon(X)\boldsymbol{\alpha}(X)) - \lambda^{1/(4m)}V(\epsilon(X)K_X(x), \epsilon(X)A(X)). \end{aligned}$$

Using similar derivations as above, by (5.35) and (5.36), for any $x \in \mathbb{I}$ and $k = 1, .., p$,

$$\begin{aligned} & |V(\epsilon(X)K_X(x), \epsilon(X)\alpha_k(X))|^2 \\ = \; & |\sum_\mu \frac{b_\mu(x)}{1+\lambda\gamma_\mu} V(\epsilon(X)b_\mu(X), \epsilon(X)\alpha_k(X))|^2 \\ \leq \; & \sum |V(\epsilon(X)b_\mu(X), \epsilon(X)\alpha_k(X))|^2 (1+\gamma_\mu)^b (b_\mu(x))^2 \left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b(1+\lambda\gamma_\mu)^2}\right) \\ \leq \; & O\left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b}\right), \end{aligned}$$

and

$$\begin{aligned} & |V(\epsilon(X)K_X(x), \epsilon(X)A(X))|^2 \\ = \; & |\sum_\mu \frac{b_\mu(x)}{(1+\lambda\gamma_\mu)^2} V(\alpha_k, b_\mu) V(\epsilon b_\nu(X), \epsilon b_\mu(X))|^2 \\ \leq \; & \sum |V(\alpha_k, b_\mu) V(\epsilon b_\nu(X), \epsilon b_\mu(X))|^2 (1+\gamma_\mu)^b (b_\mu(x))^2 \left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b(1+\lambda\gamma_\mu)^2}\right) \\ \leq \; & O\left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b}\right). \end{aligned}$$

Therefore, in Assumption 4 and (5.62), $\boldsymbol{\zeta}_x = o(1)$.

By (i) and (ii), as $n \to \infty$,

$$\begin{aligned} & \begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\theta}}_0^* - \boldsymbol{\theta}_0) \\ \sqrt{n\lambda^{1/(2m)}} \left\{ \hat{h}_0^*(x) - h_0(x) + W_\lambda h_0(x) \right\} \end{pmatrix} \\ = & \begin{pmatrix} \sqrt{n}(\Omega + \Sigma_\lambda)^{-1} V(\boldsymbol{\alpha}, W_\lambda h_0) \\ -\sqrt{n\lambda^{1/(2m)}} V(\boldsymbol{\alpha}^T, W_\lambda h_0)(\Omega + \Sigma_\lambda)^{-1} A(x) \end{pmatrix} \to 0. \end{aligned}$$

■

With Lemma 4 and Lemma 5, we can directly conclude the limit distribution for the semiparametric estimates.

Theorem 3 *(Joint limit distribution). Let Assumptions 1 to 4 hold. Suppose that there exists $b \in (1/(2m), 1]$ such that $\alpha_k(\cdot)$ satisfies*

$$\sum_\nu |V(\alpha_k, b_\nu)|^2 \gamma_\nu^b < \infty, \qquad \text{for any } k = 1, ..., p. \tag{5.67}$$

Moreover, as $n \to \infty$, $\lambda = o(1)$, $n\lambda^{1/m} \to \infty$, $a_n \log n = o(n^{-1/2}\lambda^{1/(4m)})$. Then we have, for any $x \in \mathbb{I}$,

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\theta}}_{n,\lambda} - \boldsymbol{\theta}_0) \\ \sqrt{n\lambda^{1/(2m)}} \left\{ \hat{h}_{n,\lambda}(x) - h_0(x) + W_\lambda h_0(x) \right\} \end{pmatrix} \xrightarrow{d} N(0, \Psi), \tag{5.68}$$

where

$$\Psi = \begin{pmatrix} \Omega^* & 0 \\ 0 & \sigma_x^2 \end{pmatrix}.$$

5.5 Simulations

To evaluate the proposed estimation procedures and algorithms, we conduct several simulations for both additive and nonadditive cases. We use the KL divergence between the estimated density and the true density as introduced in Chapter 4 to evaluate the performance of density estimation, and bias, variance and MSE to evaluate the performance of parameter estimation.

5.5.1 Additive Case

Near Normal distribution

Consider the density function

$$f(x;\mu,\sigma) = \frac{\exp\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} + ax^3\}}{\int_0^1 \exp\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} + ax^3\}dx}$$

where $x \in [0,1]$, $\alpha(x;\mu,\sigma) = -\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}$, $h(x) = ax^3$ and a is a constant. The function $\alpha(x;\mu,\sigma)$ is the logistic transformation of the truncated normal density function with mean μ and standard deviation σ. The constant a controls the departure from the truncated normal distribution. The choice of this density function is motivated by Hjort and Glad [14]. Hjort and Glad [14] considered a modified kernel estimator of f as $f_0(x;\boldsymbol{\theta})r(x)$ where $f_0(x;\boldsymbol{\theta})$ is a parametric density with unknown parameters $\boldsymbol{\theta}$ and $r(x)$ is a nonparametric correction function. The form of modified kernel estimator is a special case of (5.3) with $h(x) = \log r(x)$ and $\alpha(x;\boldsymbol{\theta}) = \log f_0(x;\boldsymbol{\theta})$. Hjort and Glad [14] showed that starting with a parametric density estimate leads to a better estimate of the density function than a direct kernel estimate when the true function is close to the parametric density. Therefore this simulation setting can be regarded as starting with the truncated normal. When a is small, we expect that the semiparametric estimate of the density

performs better than direct nonparametric estimation methods such as kernel and cubic spline. We will consider the additive model (5.3) with $\alpha(x;\mu,\sigma) = -\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}$ and $h \in W_{20}^3(\mathbb{I})$. We use the profile likelihood approach in Section 5.3.1 to compute estimates of $\boldsymbol{\theta} = (\mu, \sigma)$ and h. For comparison, we also implement the approach in Yang [17] which is a special case of model (5.2) with preselected $\boldsymbol{\alpha}(x) = (x, x^2)^T$. Yang's approach is equivalent to the nonparametric smoothing spline estimation with model space $W_{20}^3(\mathbb{I})$ for quintic spline, which is more tailored for this case and thus is expected to behave better than the cubic spline. In the implementation of kernel density estimation and Hjort & Glad's method, we use the Gaussian kernel and compare different methods for bandwidth selection including the approach in Scott [65], the unbiased and biased cross-validation procedures, and the method using pilot estimation of derivatives in Sheather and Jones [66]. We find that the bandwidth obtained by

$$h = 0.9n^{-1/5}\min\{\text{sample standard deviation, sample inter quantile range}/1.34\}, \tag{5.69}$$

provides the best overall performance. We report results with bandwidth (5.69) only. We set $\mu = 0.5$ and $\sigma = 0.2$. We consider 6 choices of a: $a = 0.25, 0.5, 1, 2, 3, 4, 6$, and three sample sizes, $n = 100, 200$ and 300. For each simulation setting, we generate 100 simulation data sets. For each simulated data set, we compute density estimates using the kernel method, Hjort & Glad's (HG), cubic spline, quintic spline (i.e. Yang's method) and our proposed semiparametric (SEMI) method. The simulation results are shown in Tables 5.2, 5.3 and 5.4. As expected, except for $a = 6$ with small sample size, the semiparametric and quintic spline estimates have the smallest KL than nonparametric estimates. Our semiparametric approach performs better than the semiparametric approach in Hjort & Glad (1995). The bias and variance in the estimation of parameters increase when the true density is further away from the truncated normal. Nevertheless, the performance of

the estimation to the density function is not affected by the large biases and variances in the estimation of parameters when a is large. And as expected, as sample size becomes larger, the density and parameters estimations are improved.

a	Method	KL	KLBias	KLvar	|Bias(μ)|	Var(μ)	|Bias(σ)|	Var(σ)
	Kernel	2.56	0.04	2.52				
0.25	HG	2.48	0.47	2.01				
	Cubic	1.53	0.31	1.22				
	Quintic	1.08	0.04	1.04	0.75	2.33	0.05	1.65
	SEMI	1.08	0.04	1.04	0.75	2.34	0.05	1.65
	Kernel	2.67	0.08	2.60				
0.5	HG	2.38	0.12	2.25				
	Cubic	1.53	0.31	1.22				
	Quintic	1.11	0.05	1.06	1.56	2.07	0.43	1.96
	SEMI	1.11	0.05	1.06	1.57	2.07	0.43	1.97
	Kernel	2.81	0.04	2.78				
1	HG	2.30	0.14	2.16				
	Cubic	1.50	0.28	1.23				
	Quintic	0.93	0.03	0.90	4.05	2.37	1.17	1.75
	SEMI	0.93	0.03	0.90	4.05	2.37	1.17	1.75
	Kernel	3.73	0.16	3.57				
2	Cubic	1.75	0.32	1.42				
	HG	2.79	0.56	2.23				
	Quintic	1.19	0.04	1.15	9.47	3.14	3.49	2.78
	SEMI	1.20	0.03	1.16	9.40	3.18	3.47	2.75
	Kernel	7.79	0.83	6.96				
4	HG	4.32	2.19	2.13				
	Cubic	1.54	0.27	1.27				
	Quintic	1.27	0.19	1.08	31.94	9.90	10.01	5.60
	SEMI	1.27	0.19	1.08	32.03	10.02	10.03	5.64
	Kernel	16.06	5.01	11.05				
6	HG	11.73	9.23	2.51				
	Cubic	1.47	0.37	1.10				
	Quintic	1.87	0.68	1.18	82.00	178.91	24.55	15.75
	SEMI	1.84	0.71	1.12	3265.02	19500.91	77.57	266.74

Table 5.2: Performance of different methods (all numbers are in 10^{-2}) with sample size 100.

a	Method	KL	KLBias	KLvar	$\lvert\text{Bias}(\mu)\rvert$	$\text{Var}(\mu)$	$\lvert\text{Bias}(\sigma)\rvert$	$\text{Var}(\sigma)$
	Kernel	1.65	0.03	1.62				
0.25	HG	1.20	0.18	1.03				
	Cubic	0.81	0.21	0.60				
	Quintic	0.42	0.01	0.42	0.95	1.31	0.24	1.24
	SEMI	0.43	0.01	0.42	0.91	1.35	0.23	1.26
	Kernel	1.77	0.04	1.73				
0.5	HG	1.38	0.26	1.12				
	Cubic	0.87	0.17	0.71				
	Quintic	0.54	0.02	0.52	1.60	1.35	0.51	1.48
	SEMI	0.54	0.02	0.52	1.60	1.36	0.51	1.48
	Kernel	1.84	0.07	1.77				
1	HG	1.39	0.43	0.95				
	Cubic	0.81	0.15	0.65				
	Quintic	0.46	0.02	0.44	3.70	1.48	1.24	1.39
	SEMI	0.47	0.02	0.45	3.72	1.52	1.25	1.40
	Kernel	2.56	0.07	2.48				
2	HG	1.37	0.38	0.99				
	Cubic	0.81	0.23	0.58				
	Quintic	0.50	0.04	0.47	9.41	1.80	3.31	1.80
	SEMI	0.50	0.04	0.47	9.42	1.80	3.31	1.80
	Kernel	6.24	0.74	5.50				
4	HG	2.77	1.73	1.04				
	Cubic	0.79	0.18	0.61				
	Quintic	0.63	0.14	0.49	30.84	6.19	10.25	3.88
	SEMI	0.64	0.15	0.49	30.75	6.23	10.23	3.88
	Kernel	12.88	4.31	8.57				
6	HG	8.05	7.04	1.01				
	Cubic	0.94	0.36	0.58				
	Quintic	1.10	0.64	0.46	167.13	453.96	30.29	32.62
	SEMI	1.12	0.66	0.46	155.26	345.87	29.72	29.05

Table 5.3: Performance of different methods (all numbers are in 10^{-2}) with sample size 200.

a	Method	KL	KLBias	KLvar	$\lvert\text{Bias}(\mu)\rvert$	$\text{Var}(\mu)$	$\lvert\text{Bias}(\sigma)\rvert$	$\text{Var}(\sigma)$
	Kernel	0.96	0.03	0.93				
0.25	HG	0.62	0.12	0.50				
	Cubic	0.41	0.11	0.30				
	Quintic	0.20	0.003	0.20	0.89	0.97	0.24	0.78
	SEMI	0.21	0.003	0.20	0.87	1.01	0.23	0.78
	Kernel	1.07	0.02	1.05				
0.5	HG	0.71	0.11	0.60				
	Cubic	0.42	0.09	0.33				
	Quintic	0.24	0.01	0.23	1.75	1.13	0.50	0.80
	SEMI	0.26	0.01	0.25	167.02	1653.18	4.46	39.70
	Kernel	1.25	0.04	1.21				
1	HG	0.67	0.11	0.56				
	Cubic	0.45	0.12	0.33				
	Quintic	0.22	0.01	0.21	4.15	1.11	1.46	0.87
	SEMI	0.24	0.01	0.23	3.87	2.74	1.35	1.32
	Kernel	1.80	0.08	1.72				
2	HG	0.73	0.29	0.45				
	Cubic	0.40	0.13	0.28				
	Quintic	0.22	0.04	0.18	9.19	1.07	3.49	1.17
	SEMI	0.22	0.04	0.18	8.96	1.12	3.41	1.17
	Kernel	4.77	0.67	4.11				
4	HG	1.78	1.29	0.49				
	Cubic	0.34	0.07	0.27				
	Quintic	0.33	0.15	0.18	29.87	3.20	9.75	2.04
	SEMI	0.35	0.16	0.19	29.15	3.23	9.54	2.01
	Kernel	10.46	3.79	6.67				
6	HG	5.95	5.54	0.41				
	Cubic	0.71	0.45	0.26				
	Quintic	0.80	0.64	0.16	97.66	31.31	23.40	7.72
	SEMI	0.82	0.66	0.16	96.30	33.64	23.07	8.20

Table 5.4: Performance of different methods (all numbers are in 10^{-2}) with sample size 500.

Near Gumbel distribution

Consider the density function

$$f(x;\mu,\sigma) = \frac{\exp\{-\frac{x-\mu}{\sigma} - \exp(\frac{x-\mu}{\sigma}) + ax^3\}}{\int_0^1 \exp\{-\frac{x-\mu}{\sigma} - \exp(\frac{x-\mu}{\sigma}) + ax^3\}dx}$$

where $x \in [0,1]$, $\alpha(x;\mu,\sigma) = -\frac{x-\mu}{\sigma} - \exp(\frac{x-\mu}{\sigma})$, $h(x) = ax^3$, and a is a constant. This is equivalent to start with the truncated Gumbel distribution whose logistic transformation is $\alpha(x;\mu,\sigma)$, and the constant a controls the departure from the truncated Gumbel distribution. We will consider the additive model (5.3) with $\alpha(x;\mu,\sigma) = -\frac{x-\mu}{\sigma} - \exp(\frac{x-\mu}{\sigma})$ and $h \in W_{20}^2(\mathbb{I})$. Note that α is non-linear in $\boldsymbol{\theta}$. Therefore, Yang's approach does not apply. We compare our proposed method with the kernel, cubic spline and HG's method, where f_0 in HG's approach is the truncated Gumbel distribution. We use the profile likelihood approach as described in Section 5.3.1 to compute estimates of $\boldsymbol{\theta} = (\mu,\sigma)$ and h.

In the simulations, we set $\mu = 0.5$ and $\sigma = 0.2$. We consider five choices of a: $a = 0.25, 0.5, 1, 2, 4$, and three sample sizes $100, 200, 500$. For each simulation setting, we generate 100 simulated data sets. The simulation results are shown in Tables 5.5, 5.6 and 5.7. As expected, our semiparametric approach performs better than the nonparametric methods when the true density is not far from truncated Gumbel distribution ($a = 0.25, 0.5, 1, 2$). Our method also performs better than the semiparametric method in Hjort and Glad (1995). The bias and variance in the estimation of parameters increase when the true density is further away from the truncated Gumbel. Nevertheless, the performance of the estimation to the density function is not affected by the large biases and variances in the estimation of parameters when a is large. The density and parameters estimations are improved as sample size increases.

a	Method	KL	KLbias	KLvar	$\lvert\text{Bias}(\mu)\rvert$	$\text{Var}(\mu)$	$\lvert\text{Bias}(\sigma)\rvert$	$\text{Var}(\sigma)$
	Kernel	3.50	0.41	3.09				
0.25	HG	3.20	1.12	2.08				
	Cubic	2.12	0.59	1.53				
	SEMI	1.58	0.16	1.42	0.48	4.28	0.27	3.46
	Kernel	3.91	0.36	3.55				
0.5	HG	3.04	1.19	1.85				
	Cubic	2.02	0.86	1.16				
	SEMI	1.52	0.27	1.25	5.26	5.12	2.39	3.87
	Kernel	4.32	0.47	3.85				
1	HG	3.07	1.00	2.07				
	Cubic	2.17	0.69	1.48				
	SEMI	1.28	0.11	1.17	6.52	4.31	2.87	3.52
	Kernel	6.29	0.79	5.50				
2	HG	3.92	1.85	2.07				
	Cubic	2.13	0.65	1.48				
	SEMI	1.70	0.23	1.47	15.30	9.19	7.23	7.12
	Kernel	11.47	1.86	9.61				
4	HG	7.83	5.85	1.98				
	Cubic	2.00	0.65	1.35				
	SEMI	2.33	0.58	1.75	79.50	48.42	35.64	22.92

Table 5.5: Performance of different methods (all numbers are in 10^{-2}) with sample size 100.

a	Method	KL	KLbias	KLvar	$\lvert\text{Bias}(\mu)\rvert$	$\text{Var}(\mu)$	$\lvert\text{Bias}(\sigma)\rvert$	$\text{Var}(\sigma)$
0.25	Kernel	2.60	0.36	2.23				
	HG	1.89	0.68	1.21				
	Cubic	1.32	0.42	0.90				
	SEMI	0.65	0.01	0.63	0.90	3.11	0.41	2.64
0.5	Kernel	2.84	0.33	2.51				
	HG	2.08	0.81	1.27				
	Cubic	1.33	0.37	0.96				
	SEMI	0.98	0.07	0.92	2.05	5.66	0.61	4.17
1	Kernel	3.22	0.49	2.73				
	HG	2.04	1.02	1.02				
	Cubic	1.21	0.52	0.69				
	SEMI	0.64	0.06	0.58	8.32	24.96	4.11	12.56
2	Kernel	4.83	0.63	4.20				
	HG	2.50	1.47	1.03				
	Cubic	1.16	0.39	0.77				
	SEMI	1.19	0.22	0.97	13.13	9.00	5.67	6.68
4	Kernel	9.67	1.88	7.80				
	HG	5.27	4.08	1.19				
	Cubic	1.08	0.37	0.71				
	SEMI	1.31	0.32	0.98	72.11	125.28	30.88	40.44

Table 5.6: Performance of different methods (all numbers are in 10^{-2}) with sample size 200.

a	Method	KL	KLbias	KLvar	$\lvert\text{Bias}(\mu)\rvert$	$\text{Var}(\mu)$	$\lvert\text{Bias}(\sigma)\rvert$	$\text{Var}(\sigma)$
	Kernel	1.72	0.21	1.51				
0.25	HG	1.09	0.49	0.61				
	Cubic	0.63	0.19	0.44				
	SEMI	0.45	0.04	0.41	0.26	4.84	0.69	3.55
	Kernel	1.86	0.27	1.60				
0.5	HG	1.02	0.49	0.53				
	Cubic	0.63	0.19	0.44				
	SEMI	0.40	0.04	0.36	1.06	4.90	0.11	3.49
	Kernel	2.36	0.31	2.04				
1	HG	1.18	0.61	0.58				
	Cubic	0.64	0.21	0.43				
	SEMI	0.44	0.05	0.40	5.33	8.52	1.89	5.28
	Kernel	3.71	0.56	3.15				
2	HG	1.54	1.02	0.52				
	Cubic	0.58	0.20	0.38				
	SEMI	0.48	0.07	0.41	9.34	7.79	3.12	5.26
	Kernel	7.68	1.53	6.15				
4	HG	3.51	3.04	0.46				
	Cubic	0.52	0.19	0.33				
	SEMI	0.65	0.13	0.53	33.09	29.04	12.75	15.80

Table 5.7: Performance of different methods (all numbers are in 10^{-2}) with sample size 500.

5.5.2 Nonadditive Case

Power Transformation

The truncated Weibull distribution has density function

$$f(x;s,k) = \frac{k}{s}\left(\frac{x}{s}\right)^{k-1}\exp\left\{-\left(\frac{x}{s}\right)^{k}\right\}, \tag{5.70}$$

where $x \in [0,1]$, $s > 0$ is the scale parameter and $k > 0$ is the shape parameter. Since $Y = t(X;k) \sim$ truncated $\text{Exp}(s)$ which is independent of k, we use the power transformation

$$t(x;\theta) = x^{\theta}, \quad \theta > 0. \tag{5.71}$$

We consider the transformation model in Section 5.3.3 with $h \in W_{20}^2(\mathbb{I})$ and apply Algorithm 2 to estimate $k > 0$ and h. We also compute cubic spline estimate of the density function for comparison. In the simulations, we set $s = 1$, and consider $k = 1, 2, 3, 5$ where $k = 1$ corresponding to the truncated exponential distribution for which the cubic spline is tailored. We consider three sample sizes 50, 100 and 200. For each simulation setting, we generate 100 data sets. The simulation results are shown in Tables 5.8, 5.9 and 5.10. As expected, cubic spline performs better when $k = 1$. Our semiparametric approach provide can achieve smaller KL than cubic spline when $k = 2, 3, 5$. And as sample size increases, the performance of density and parameter estimation is improved.

k	Method	KL	Bias(KL)	Var(KL)	|Bias(k)|	Var(k)
	Cubic	1.30	0.005	1.29		
1	SEMI	2.15	0.03	2.11	2.73	21.43
	Cubic	3.34	1.03	2.31		
2	SEMI	2.35	0.06	2.30	7.79	45.58
	Cubic	3.31	0.88	2.44		
3	SEMI	2.36	0.04	2.32	8.31	71.95
	Cubic	2.75	0.45	2.30		
5	SEMI	2.25	0.10	2.15	21.90	102.82

Table 5.8: Comparison of the semiparametric method with the cubic spline (in 10^{-2}) when sample size is 50.

k	Method	KL	Bias(KL)	Var(KL)	|Bias(k)|	Var(k)
	Cubic	0.67	0.001	0.67		
1	SEMI	1.08	0.01	1.07	1.51	14.58
	Cubic	2.36	0.75	1.62		
2	SEMI	1.49	0.03	1.46	4.91	36.21
	Cubic	1.88	0.66	1.22		
3	SEMI	1.07	0.05	1.02	8.14	50.25
	Cubic	1.54	0.31	1.23		
5	SEMI	1.20	0.04	1.16	19.21	75.46

Table 5.9: Comparison of the semiparametric method with the cubic spline (in 10^{-2}) when sample size is 100.

k	Method	KL	Bias(KL)	Var(KL)	$\lvert\text{Bias}(k)\rvert$	Var(k)
	Cubic	0.31	0.003	0.31		
1	SEMI	0.49	0.004	0.49	0.71	10.12
	Cubic	1.24	0.47	0.77		
2	SEMI	0.64	0.02	0.62	2.63	24.07
	Cubic	1.00	0.38	0.62		
3	SEMI	0.54	0.008	0.54	2.86	30.00
	Cubic	0.84	0.25	0.59		
5	SEMI	0.51	0.01	0.50	3.33	51.47

Table 5.10: Comparison of the semiparametric method with the cubic spline (in 10^{-2}) when sample size is 200.

Two-Sample density estimation

The Gumbel distribution and logistic distribution are members of location scale family with the location parameter μ and scale parameter σ. The Gumbel distribution is also named as generalized Extreme Value distribution Type-I. It is used to model the distribution of the maximum or the minimum of a number of samples of various distributions. It is also known as the log-Weibull distribution and the double exponential distribution, and is related to the Gompertz distribution. The logistic distribution has the logistic function as the cumulative distribution, which is used in logistic regression and neural networks as the link function.

We generate two independent samples, $X_1, ..., X_{n_1} \overset{iid}{\sim} f(x; 0, 1)$ and $Y_1, ..., Y_{n_2} \overset{iid}{\sim} f(x; \mu, \sigma)$ where f is either the Gumbel distribution or the logistic distribution. We consider two combinations of parameters $(\mu, \sigma) = (2, 1)$ and $(2, 2)$, and four sample sizes $(n_1, n_2) = (100, 100), (100, 200), (200, 100)$ and $(200, 200)$. We fit model (5.4) with $t(x; \boldsymbol{\theta}) = \frac{x-\mu}{\sigma}$ and h belongs to the RKHS for univariate thin-plate spline

$$\mathcal{H} = \left\{ h : \int_{-\infty}^{\infty} (h'')^2 dx < \infty \right\}. \tag{5.72}$$

We estimate the density functions using the procedure described in Section 5.3.4, and report $\mathrm{KL}(f(\cdot;0,1),\hat{f}(\cdot;0,1)) + \mathrm{KL}(f(\cdot;\mu,\sigma),\hat{f}(\cdot;\hat{\mu},\hat{\sigma}))$ for the performance of density estimation. For comparison, we estimate the densities for X_i's and Y_i's using thin-plate spline models separately with logistic of densities belong to $\mathcal{H}$ in (5.72). Denote the separate thin-plate estimates for X and Y samples as $\hat{f}_1$ and $\hat{f}_2$ respectively. We report $\mathrm{KL}(f(\cdot;0,1),\hat{f}_1)+\mathrm{KL}(f(x;\mu,\sigma),\hat{f}_2)$ for the performance of density estimation. We report biases and variances as measures for the estimation performance of parameters. We report $\mathrm{KL}(f(\cdot;0,1),\exp(\hat{h})/\int\exp(\hat{h})dx)$ as a measure of the estimation performance for the function h.

1. Gumbel distribution The Gumbel distribution has the density function

$$f(x) = \frac{1}{\sigma} e^{-(\frac{x-\mu}{\sigma}+e^{-\frac{x-\mu}{\sigma}})}, \quad x \in \mathbb{R}. \tag{5.73}$$

The simulation results are shown in Tables 5.11, 5.12, 5.13 and 5.14. Our estimation procedure provides good estimates of parameters $\boldsymbol{\theta}$ and the function h. The performance improves as sample size increases. The semiparametric density estimates have smaller KLs than those based on separate nonparametric fits to two samples.

n_1	n_2	Performance	$\hat{\mu}$	$\hat{\sigma}$	$\mathrm{KL}_{\hat{h}}$
100	100	|Bias|	2.19	2.84	1.81
		Var	2.79	2.38	2.35
	200	Bias	2.87	0.25	2.11
		Var	2.33	1.46	2.52
200	100	|Bias|	1.01	0.75	1.26
		Var	2.32	1.43	1.60
	200	|Bias|	0.01	0.57	1.18
		Var	1.57	1.12	1.50

Table 5.11: The estimation performance (in 10^{-2}) of parameters and h function when $\mu = 2, \sigma = 1$.

n_1	n_2	Method	KL	Bias(KL)	Var(KL)
100	100	Semiparametric	4.88	1.09	3.79
		Nonparametric	5.82	2.04	3.78
	200	Semiparametric	3.86	1.04	2.82
		Nonparametric	4.75	1.69	3.06
200	100	Semiparametric	3.90	0.69	3.21
		Nonparametric	4.77	1.36	3.40
	200	Semiparametric	2.95	0.63	2.32
		Nonparametric	3.46	1.24	2.22

Table 5.12: The KL divergence (in 10^{-2}) between the true and estimated densities when $\mu = 2, \sigma = 1$.

n_1	n_2	Performance	$\hat{\mu}$	$\hat{\sigma}$	$\mathrm{KL}_{\hat{h}}$
100	100	|Bias|	3.13	2.08	2.29
		Var	16.08	10.40	2.78
	200	|Bias|	2.23	4.54	2.03
		Var	13.24	7.14	2.43
200	100	|Bias|	5.70	0.59	1.37
		Var	11.27	5.74	1.74
	200	|Bias|	2.11	1.99	1.24
		Var	5.08	3.71	1.60

Table 5.13: The estimation performance (in 10^{-2}) of parameters and h function when $\mu = 2, \sigma = 2$.

n_1	n_2	Method	KL	Bias(KL)	Var(KL)
100	100	Semiparametric	5.53	0.99	4.54
		Nonparametric	6.11	1.99	4.12
	200	Semiparametric	4.21	0.78	3.43
		Nonparametric	4.95	1.59	3.36
200	100	Semiparametric	4.06	0.73	3.33
		Nonparametric	4.57	1.63	2.94
	200	Semiparametric	3.01	0.70	2.31
		Nonparametric	3.71	1.26	2.45

Table 5.14: The KL divergence (in 10^{-2}) between the true and estimated densities when $\mu = 2, \sigma = 2$.

2. Logistic distribution The pdf of the logistic distribution is given by

$$f(x;\mu,\sigma) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma\left(1+e^{-\frac{x-\mu}{\sigma}}\right)^2}, \quad x \in \mathbb{R}. \tag{5.74}$$

The simulation results in Tables 5.15, 5.16, 5.17 and 5.18 indicate our semiparametric model provides good estimates of parameter $\boldsymbol{\theta}$ and the function h. It achieves smaller KLs than those based on separate nonparametric fits to two samples. The performance improves as sample size increases.

n_1	n_2	Performance	$\hat{\mu}$	$\hat{\sigma}$	$\mathrm{KL}_{\hat{h}}$
100	100	|Bias|	1.52	0.17	1.43
		Var	5.02	1.81	1.58
	200	|Bias|	0.03	1.07	1.17
		Var	3.73	1.37	1.22
200	100	|Bias|	4.32	1.28	0.89
		Var	5.05	1.91	0.98
	200	|Bias|	1.47	1.27	0.83
		Var	2.71	0.76	0.88

Table 5.15: The estimation performance (in 10^{-2}) of parameters and h function when $\mu = 2, \sigma = 1$.

n_1	n_2	Method	KL	Bias(KL)	Var(KL)
100	100	Semiparametric	3.06	0.25	2.81
		Nonparametric	3.26	0.59	2.67
	200	Semiparametric	2.03	0.11	1.92
		Nonparametric	2.33	0.36	1.97
200	100	Semiparametric	2.64	0.19	2.45
		Nonparametric	2.74	0.47	2.28
	200	Semiparametric	1.77	0.12	1.65
		Nonparametric	1.96	0.27	1.69

Table 5.16: The KL divergence (in 10^{-2}) between the true logistic and estimated densities when $\mu = 2, \sigma = 1$.

n_1	n_2	Performance	$\hat{\mu}$	$\hat{\sigma}$	$\mathrm{KL}_{\hat{h}}$
100	100	|Bias|	6.32	1.01	1.46
		Var	24.73	5.21	1.62
	200	|Bias|	4.63	0.67	1.38
		Var	20.61	5.39	1.52
200	100	|Bias|	7.11	1.37	0.78
		Var	20.87	5.20	0.86
	200	|Bias|	2.30	0.95	0.65
		Var	12.59	3.69	0.74

Table 5.17: The estimation performance (in 10^{-2}) of parameters and h function when $\mu = 2, \sigma = 2$.

n_1	n_2	Method	KL	Bias(KL)	Var(KL)
100	100	Semiparametric	3.06	0.33	2.73
		Nonparametric	3.49	0.67	2.82
	200	Semiparametric	2.39	0.24	2.15
		Nonparametric	2.88	0.50	2.38
200	100	Semiparametric	2.15	0.13	2.03
		Nonparametric	2.47	0.29	2.19
	200	Semiparametric	1.54	0.14	1.40
		Nonparametric	1.84	0.26	1.57

Table 5.18: The KL divergence (in 10^{-2}) between the true logistic density and estimated density when $\mu = 2, \sigma = 2$.

5.6 Examples

In this section, three data sets are used to illustrate the application of our semiparametric method in practice.

5.6.1 Suicide risk data

The data set comprises the lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks reported by Copas and Fryer [67]. Silverman [5] showed that the traditional kernel estimate for this data tends to either have

spurious noise in the tails, or mask the essential detail in the main part if the estimates are smoothed sufficiently to deal with the tail noise. It is also a classical example for which the traditional kernel estimator fails over the boundary region. We fit four models to this data set. As in Yang [17], we consider the semiparametric model (5.2) with $\boldsymbol{\alpha}(x) = (\sin(8\pi x), \sqrt{x})$ and $h \in W_{20}^2(\mathbb{I})$. The estimated parameters $\hat{\boldsymbol{\theta}} = (0.102, 2.731)$. This semiparametric method will be refered to as SEMI-linear. Motivated by Wand, Marron and Ruppert [68], we also consider the power transformation semiparametric estimation described in Section 5.5.2. This semiparametric method will be referred to as SEMI-power. The estimated power parameter is $\hat{\theta} = 0.883$. Two other models are the kernel and cubic spline. The domain for the four models is set as $[0, 750]$. The histogram in Figure 5.1 shows that these data are highly right-skewed and long-tailed. The SEMI-linear model can portray the slight shoulder pattern on the main part and avoid the spurious noise on the tail.

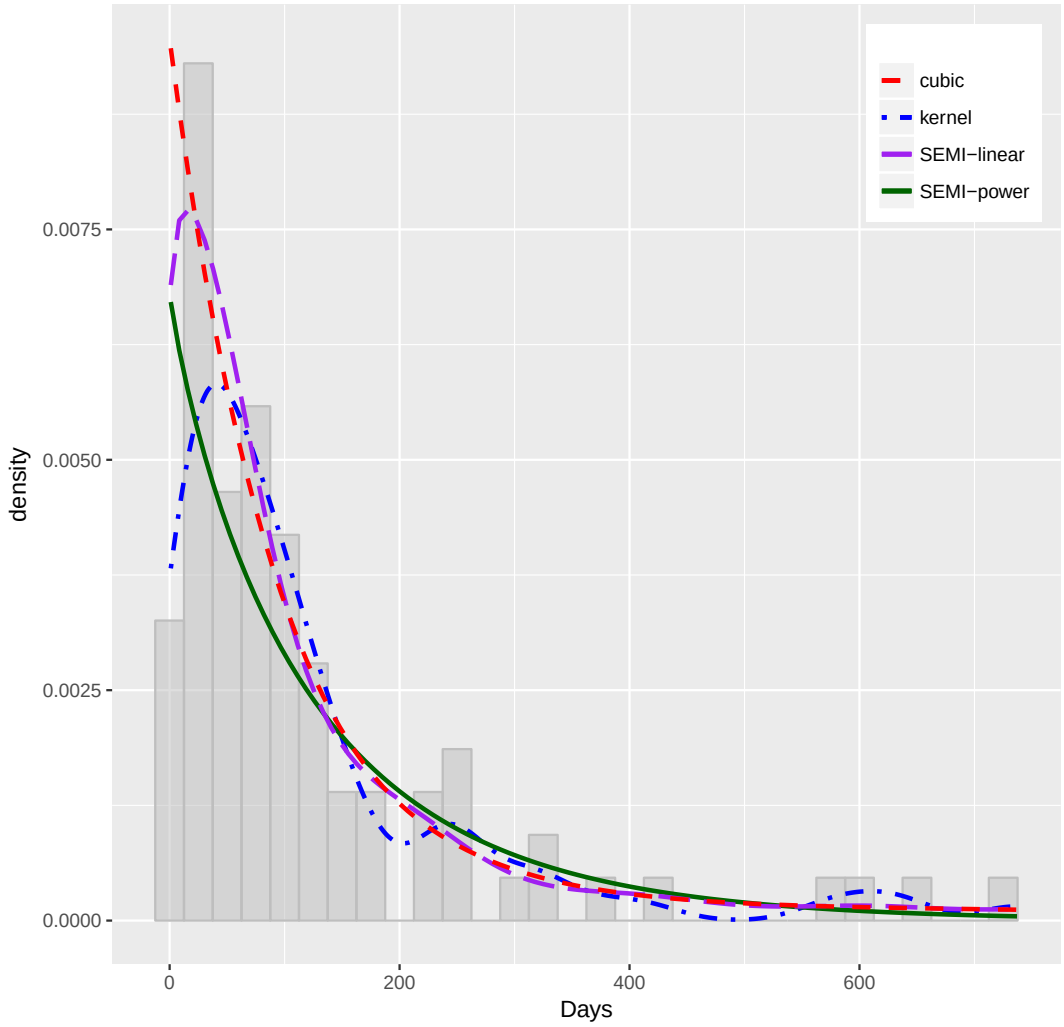

Figure 5.1: Density estimates for the suicide data.

5.6.2 Income data

The income data set has 205 pairs of log.income and age observations on Canadian workers from a 1971 Canadian Census Public Use Tape. It was used in Ruppert, Wand and Carroll [69] to illustrate semiparametric regression. Here we use the log.income only to illustrate the density estimation. The log.income variable is the natural log of the income, and the histogram in Figure 5.2 shows that the distribution is left-skewed with

one mode on the peak and some light bumps on the left tail. For comparison, we fit three models for this data. In addition to the cubic spline and kernel method, we fit the additive semiparametric model (5.3) with $\boldsymbol{\alpha}(x;\boldsymbol{\theta}) = \cos(\theta_1 + 4\pi x) + \theta_2\sqrt{x}$ and $h \in W_{20}^2(\mathbb{I})$. The estimated parameters $\hat{\boldsymbol{\theta}} = (2.67, -2.94)$. Both kernel and semiparametric density estimates pick up the small bump on the left tail. However, the kernel estimate slightly under-estimates the main peak.

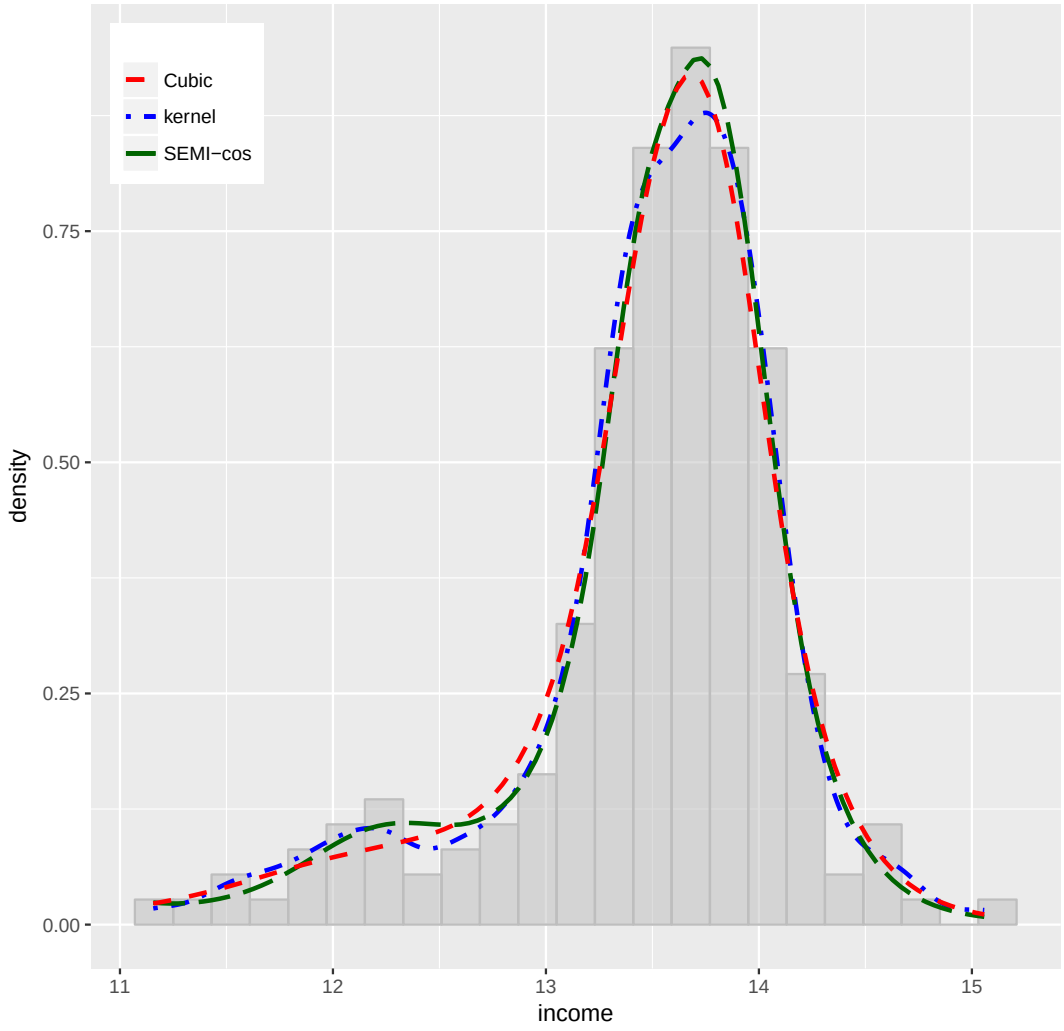

Figure 5.2: Density estimates for the income data.

5.6.3 Old Faithful data

The data set consists of waiting time between eruptions and the duration of the eruption gathered from 272 consecutive eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming, USA (Hardle [70]). Histograms of these two variables suggest that both of the two variables have two modes. Therefore, we consider a semiparametric model with a mixed normal as the parametric component. Consider the logistic density function distribution

$$\eta(x;\mu_1,\sigma_1,\mu_2,\sigma_2,p,h) = \log f_0(x;\mu_1,\sigma_1,\mu_2,\sigma_2,p) + h(x), \tag{5.75}$$

where $h \in W_{20}^2([40,100])$ for the waiting time and $h \in W_{20}^2([1.5,5.5])$ for the duration time, and f_0 is a mixture normal density

$$f_0(x) = p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right). \tag{5.76}$$

Model (5.75) is a special case of the additive model (5.3). Using the profiled likelihood estimation procedure in Section 5.3.1, we obtain the estimated parameters for each variable. We also fit the parametric mixed normal distribution (5.76) for each variable. The estimated parameters are shown in Table 5.19.

Variable	Model	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{p}$
Waiting	Mixed Normal	54.61	5.87	80.09	5.87	0.36
	SEMI	54.54	6.12	80.13	5.85	0.35
Duration	Mixed Normal	2.02	0.24	4.27	0.44	0.35
	SEMI	2.11	0.38	4.14	0.39	0.23

Table 5.19: Estimated parameters using semiparametric density esitmation and mixed normal parametric estimation for Old Faithful Data.

The estimated densities from model (76) are shown in Figures 5.3 and 5.4 for the

two variables respectively. For comparison, we also fit the kernel and cubic spline density estimates. For the variable of waiting time, the estimates from the semiparametric model and mixture normal are quite close suggesting the parametric mixture normal fits data adequately. Both of them capture two modes better than the kernel and cubic spline estimates. For the variable of duration time, the estimates from the semiparametric model and cubic spline are close. And both of them portray the left mode more precise and the right mode more details.

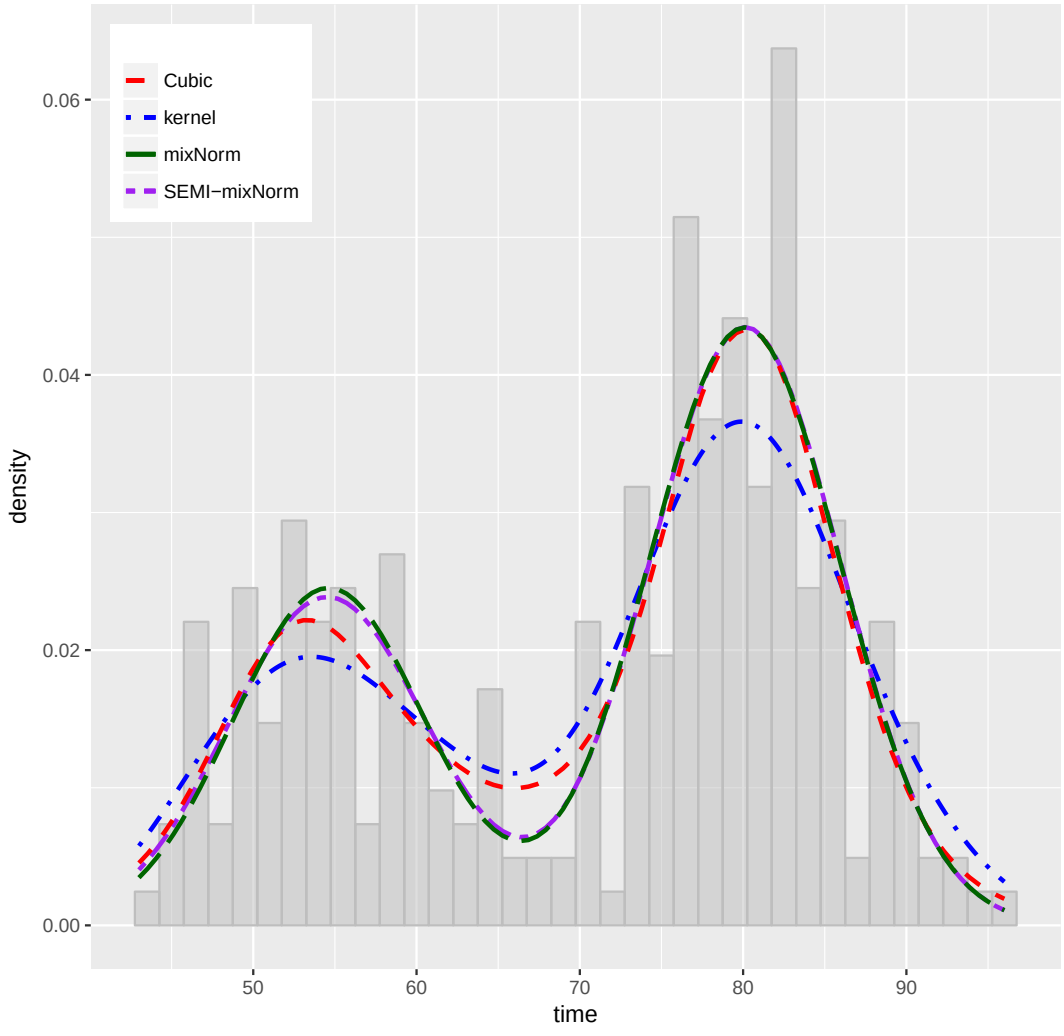

Figure 5.3: Density estimates for the waiting time of faithful data.

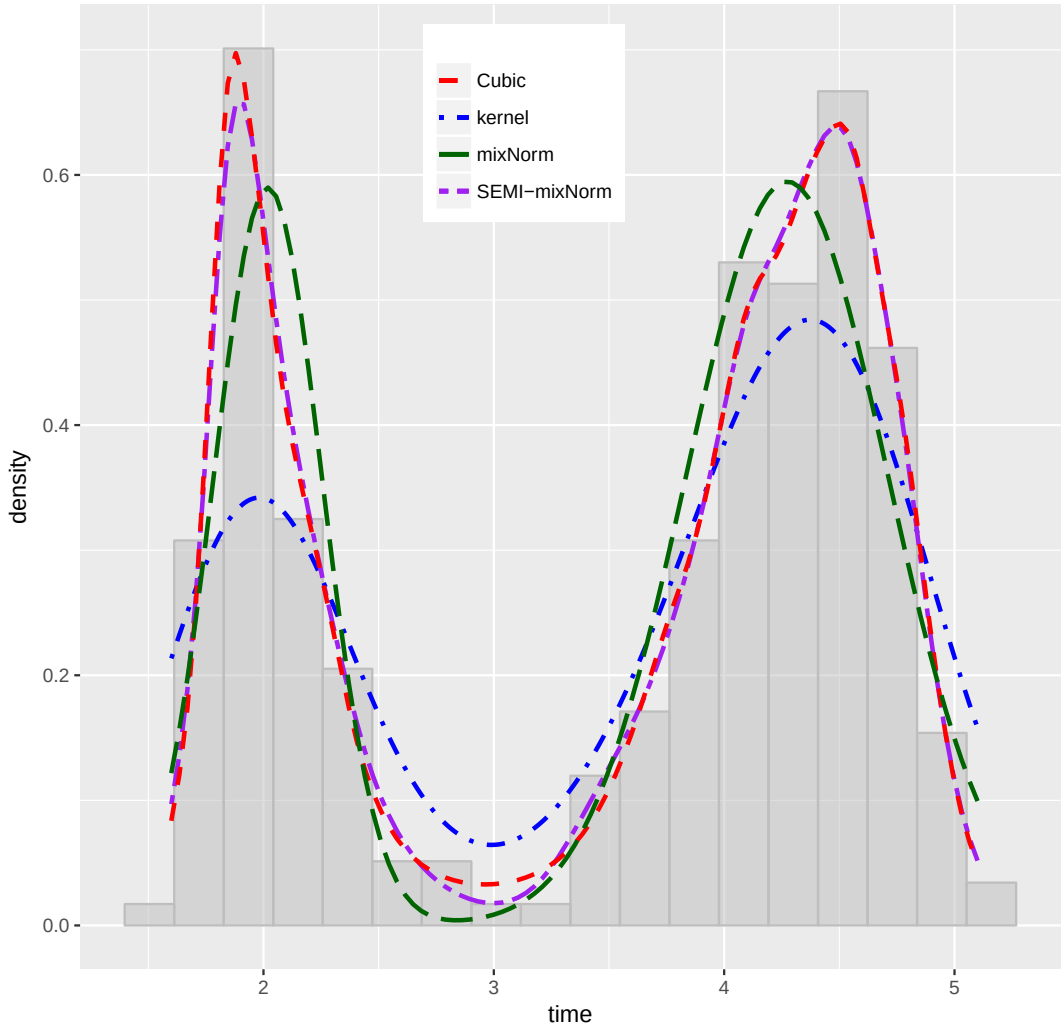

Figure 5.4: Density estimates for the duration time of faithful data.

Appendix A

Reproducing Kernels

A.1 Quintic Spline

A.1.1 RK function for quintic spline

The RK function of $\mathcal{H}_1$ for quintic spline is

$$R(x,y) = k_3(x) \times k_3(y) + k_6(x-y),$$

where

$$k_3(x) = \frac{1}{6}\left((x-.5)^3 - \frac{1}{4}(x-.5)\right)$$
$$k_6(x) = \frac{1}{720}\left((x-.5)^6 - 1.25 \times (x-.5)^4 + \frac{7}{16}(x-.5)^2 - \frac{31}{1344}\right)$$

A.1.2 RKs for tensor product quintic spline

Let

$$k_1(x) = x - .5$$
$$k_2(x) = \frac{1}{2}\left(k_1^2(x) - \frac{1}{12}\right).$$

RK functions for tensor product quintic spline are

$$\begin{aligned}
R_{1,00}(\boldsymbol{x}, \boldsymbol{y}) &= R(x^{(1)}, x^{(2)}), \\
R_{1,01}(\boldsymbol{x}, \boldsymbol{y}) &= R(x^{(1)}, x^{(2)}) \times R_0(y^{(1)}, y^{(2)}), \\
R_{00,1}(\boldsymbol{x}, \boldsymbol{y}) &= R(y^{(1)}, y^{(2)}), \\
R_{01,1}(\boldsymbol{x}, \boldsymbol{y}) &= R_0(x^{(1)}, x^{(2)}) \times R(y^{(1)}, y^{(2)}), \\
R_{1,1}(\boldsymbol{x}, \boldsymbol{y}) &= R(x^{(1)}, x^{(2)}) \times R(y^{(1)}, y^{(2)}),
\end{aligned}$$

where $\boldsymbol{x} = (x^{(1)}, x^{(2)}), \boldsymbol{y} = (y^{(1)}, y^{(2)})$, R(.,.) is the same as in A.1.1, and $R_0(x, y) = k_1(x)k_1(y) + k_2(x)k_2(y)$.

A.2 Derivation of a Reproducing Kernel for the Gamma Distribution

To save space we show a brief derivation of the RK for the Gamma distribution. Since $\mathcal{H}_0 = \{1, x, \log(x)\}$ (the constant function will be removed after this construction), the

Wronskian matrix is

$$W(x) = \begin{bmatrix} 1 & x & \log(x) \\ 0 & 1 & 1/x \\ 0 & 0 & -\frac{1}{x^2} \end{bmatrix}, \tag{A.1}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & -x & -x^2 + x^2\log(x) \\ 0 & 1 & x \\ 0 & 0 & -x^2 \end{bmatrix}. \tag{A.2}$$

The Green function is

$$G(t,s) = -s^2 + s^2\log(s) + ts - s^2\log(t) \quad (s <= t). \tag{A.3}$$

Thus, the RK of $\mathcal{H}_1$ is

$$\begin{aligned} R_1(x,z) = &\int_0^T G(x,s)G(z,s)ds \\ =&(1 + \log(z) + \log(x) + \log(z)\log(x)) \times I_4(x \wedge z) \\ &- (z + x + z\log(x) + x\log(z)) \times I_3(x \wedge z) \\ &+ xzI_2(x \wedge z) \\ &+ I_{4,2}(x \wedge z) \\ &- (2 + \log(z) + \log(x))\, I_{4,1}(x \wedge z) \\ &+ (z + x)I_{3,1}(x \wedge z), \end{aligned} \tag{A.4}$$

where $x \wedge z = \min(x, z)$, and

$$\begin{aligned} I_p(s) &= \int_0^s x^p dx = \frac{1}{p+1}(s)^{p+1} \\ I_{p,k}(s) &= \int_0^s x^p \log(x)^k = \frac{1}{p+1}(s)^{p+1}\log(s)^k - \frac{k}{p+1}I_{p+1,k-1}(s). \end{aligned} \tag{A.5}$$

A.3 Derivation of a Reproducing Kernel for the Beta Distribution

To save space we show a brief derivation of the RK for the Beta distribution. Given the differential operator L in equation (4.9), the Wronskian matrix associated with $\mathcal{H}_0$ is

$$W(x) = \begin{bmatrix} 1 & \log(x) & \log(1-x) \\ 0 & 1/x & 1/(x-1) \\ 0 & -\frac{1}{x^2} & -\frac{1}{(x-1)^2} \end{bmatrix}, \tag{A.6}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & (x-1)^2\log(1-x) - x^2\log(x) & x(x-1)\left[(x-1)\log(1-x) - x\log(x)\right] \\ 0 & x^2 & x^2(x-1) \\ 0 & -(x-1)^2 & -x(x-1)^2 \end{bmatrix}. \tag{A.7}$$

The Green function is

$$G(t,s) = s(s-1)\left[(s-1)\log(1-s) - s\log(s)\right] + s^2(s-1)\log(t) - s(s-1)^2\log(1-t), \quad s <= t. \tag{A.8}$$

Then, the RK of $\mathcal{H}_1$ is

$$\begin{aligned}
R_1(x,z) =& \int_0^{x\wedge z} G(x,s)G(z,s)ds \\
=&(\log(z)\log(1-x)+\log(x)\log(1-z))I(x\wedge z;3,3,0,0) \\
&+\log(1-x)\log(1-z)I(x\wedge z;2,4,0,0)+\log(x)\log(z)I(x\wedge z;4,2,0,0) \\
&+I(x\wedge z;2,4,0,2)+(\log(x)+\log(z))I(x\wedge z;3,3,0,1) \\
&-(\log(1-x)+\log(1-z))I(x\wedge z;2,4,0,1) \\
&+I(x\wedge z;4,2,2,0)-(\log(x)+\log(z))I(x\wedge z;4,2,1,0) \\
&-(\log(1-x)+\log(1-z))I(x\wedge z;3,3,1,0) \\
&+2I(x\wedge z;3,3,1,1),
\end{aligned} \tag{A.9}$$

where

$$I(y;m_1,m_2,m_3,m_4)=\int_0^y x^{m_1}(1-x)^{m_2}\log(x)^{m_3}\log(1-x)^{m_4}dx. \tag{A.10}$$

A.4 Derivation of a Reproducing Kernel for the GGIG Family

To save space we show a brief derivation of the RK for the GGIG family with $p=1$ only. Note that $L = D^4 + 6x^{-1}D^3 + 6x^{-2}D^2$ and $\mathcal{H}_0 = \text{span}\{1,\log(x),x,x^{-1}\}$. The

Wronskian matrix associated with $\mathcal{H}_0$ is

$$W(x) = \begin{bmatrix} 1 & x & x^{-1} & \log(x) \\ 0 & 1 & -x^{-2} & x^{-1} \\ 0 & 0 & 2x^{-3} & -x^{-2} \\ 0 & 0 & -6x^{-4} & 2x^{-3} \end{bmatrix}, \tag{A.11}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & -x & -x^2 + 3x^2\log(x) & x^3\log(x) \\ 0 & 1 & 2x & .5x^2 \\ 0 & 0 & -x^3 & -.5x^4 \\ 0 & 0 & -3x^2 & -x^3 \end{bmatrix}. \tag{A.12}$$

The Green function is

$$G(t,s) = -\frac{s^4}{2t} + \frac{s^2 t}{2} + s^3\log(s) - s^3\log(t), \quad s \le t. \tag{A.13}$$

Thus, the RK of $\mathcal{H}_1$

$$\begin{aligned} R_1(x,z) &= \int_0^T G(x,s)G(z,s)ds \\ &= \frac{1}{36xz}(x\wedge z)^9 - \frac{1}{16}(x\wedge z)^8\log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) \\ &\quad +\frac{1}{16}(x\wedge z)^8\left(\frac{1}{8x}+\frac{1}{8z}+\frac{1}{z}\log(x)+\frac{1}{x}\log(z)\right) \\ &\quad -\frac{1}{7}(x\wedge z)^7\log(x\wedge z)\left(\frac{2}{7}+\log(x)+\log(z)\right)+\frac{1}{7}(x\wedge z)^7\log(x\wedge z)^2 \\ &\quad +\frac{1}{7}(x\wedge z)^7\left(\frac{2}{49}-\frac{z}{4x}-\frac{x}{4z}+\frac{1}{7}\log(x)+\frac{1}{7}\log(z)+\log(x)\log(z)\right) \\ &\quad +\frac{1}{12}(x+z)(x\wedge z)^6\log(x\wedge z)-\frac{1}{12}(\frac{x}{6}+x\log(z)+\frac{z}{6}+z\log(x))(x\wedge z)^6+\frac{1}{20}(x\wedge z)^5 xz. \end{aligned}$$

A.5 Derivation of a Reproducing Kernel for the Inverse Gamma Distribution

Here we show a brief derivation of the RK for the inverse Gamma distribution. Note that $L = x^2D^3+4xD^2+2D$ and the null space is $\mathcal{H}_0 = \text{span}\{1, \log(x), \frac{1}{x}\}$. The Wronskian matrix associated with $\mathcal{H}_0$ is

$$W(x) = \begin{bmatrix} 1 & 1/x & \log(x) \\ 0 & -\frac{1}{x^2} & 1/x \\ 0 & \frac{2}{x^3} & -\frac{1}{x^2} \end{bmatrix},$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & -x - 2x\log(x) & -x^2 + x^2\log(x) \\ 0 & x^2 & x^3 \\ 0 & 2x & x^2 \end{bmatrix}.$$

Green function is

$$G(t,s) = -s^2 - s^2\log(s) + \frac{1}{t}s^3 + s^2 log(t) \quad (s <= t).$$

Thus, the RK of $\mathcal{H}_1$ is

$$
\begin{aligned}
R_1(x,z) = &\int_0^T G(x,s)G(z,s)ds \\
=&\frac{1}{7xz}(x\wedge z)^7 - \frac{1}{6}(x\wedge z)^6\log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) - \\
&\frac{1}{6}(x\wedge z)^6\left(\frac{5}{6x}+\frac{5}{6z}-\frac{1}{z}\log(x)-\frac{1}{x}\log(z)\right) + \\
&\frac{1}{5}(x\wedge z)^5\log(x\wedge z)\left(\frac{8}{5}-\log(x)-\log(z)\right)+\frac{1}{5}(x\wedge z)^5\log(x\wedge z)^2 + \\
&\frac{1}{5}(x\wedge z)^5\left(\frac{17}{25}-\frac{4}{5}\log(x)-\frac{4}{5}\log(z)+\log(x)\log(z)\right).
\end{aligned}
$$

Bibliography

[1] K. Pearson, *On the systematic fitting of curves to observations and measurements*, *Biometrika* **1** (1902), no. 3 265–303.

[2] K. Pearson, *On the systematic fitting of curves to observations and measurments: Part ii*, *Biometrika* **2** (1902), no. 1 1–23.

[3] M. G. Kendall *et. al.*, *The advanced theory of statistics.*, *The advanced theory of statistics.* (1946), no. 2nd Ed.

[4] R. Fisher, *On an absolute criterion for fitting frequency curves*, *Statistical Science* **12** (1997), no. 1 39–41.

[5] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[6] A. J. Izenman, *Review papers: Recent developments in nonparametric density estimation*, *Journal of the American Statistical Association* **86** (1991), no. 413 205–224.

[7] M. Rudemo, *Empirical choice of histograms and kernel density estimators*, *Scand. J. Statist.* (1982) 65–78.

[8] A. W. Bowman, *An alternative method of cross-validation for the smoothing of density estimates*, *Biometrika* **71** (1984) 353–360.

[9] P. Hall, *Large sample oprimality of least squares cross-validation in density estimation*, *Ann. Statist.* **11** (1983) 1156–1174.

[10] C. J. Stone, *An asymptotically optimal window selection rule for kernel density estimates.*, *Ann. Statist.* **12** (1984) 1285–1297.

[11] C. Gu, *Smoothing Spline ANOVA Models, 2nd ed.* Springer-Verlag, New York, 2013.

[12] C. Gu and J. Wang, *Penalized likelihood density estimation: Direct cross-validation and scalable approximation*, *Statistica Sinica* (2003) 811–826.

[13] I. Olkin and C. H. Spiegelman, *A semiparametric approach to density estimation*, *Journal of the American Statistical Association* **82** (1987), no. 399 858–865.

[14] N. L. Hjort and I. K. Glad, *Nonparametric density estimation with a parametric start*, *The Annals of Statistics* (1995) 882–904.

[15] B. Efron, R. Tibshirani, *et. al.*, *Using specially designed exponential families for density estimation*, *The Annals of Statistics* **24** (1996), no. 6 2431–2461.

[16] P. J. Lenk, *Bayesian semiparametric density estimation and model verification using a logistic–gaussian process*, *Journal of Computational and Graphical Statistics* **12** (2003), no. 3 548–565.

[17] Y. Yang, *Penalized semiparametric density estimation*, *Statistics and Computing* **19** (2009), no. 4 355.

[18] A. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione, giornale dell*, *Instituto Italiano degli Attuari* (1933) 83–91.

[19] H. W. Lilliefors, *On the kolmogorov-smirnov test for normality with mean and variance unknown*, *Journal of the American Statistical Association* **62** (1967), no. 318 399–402.

[20] K. Pearson, *X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50** (1900), no. 302 157–175.

[21] A. K. Dey and D. Kundu, *Discriminating between the log-normal and log-logistic distributions*, *Communications in Statistics-Theory and Methods* **39** (2009), no. 2 280–292.

[22] R. S. S. Ying Li and Y. Sun, *Goodness-of-fit tests of a parametric density functions: Monte carlo simulation studies*, *Journal of Statistical Research* **39** (2005), no. 02 103–125.

[23] Y. Fan, *Testing the goodness of fit of a parametric density function by kernel method*, *Econometric Theory* **10** (1994), no. 02 316–356.

[24] F. J. Rubio, M. F. Steel, *et. al.*, *Inference in two-piece location-scale models with jeffreys priors*, *Bayesian Analysis* **9** (2014), no. 1 1–22.

[25] S. Cai, J. Chen, and J. V. Zidek, *Hypothesis testing in the presence of multiple samples under density ratio models*, *arXiv preprint arXiv:1309.4740* (2013).

[26] G. Li and J. Qin, *Semiparametric likelihood-based inference for biased and truncated data when the total sample size is known*, *Journal of the Royal Statistical Society. Series B, Statistical Methodology* (1998) 243–254.

[27] T. De Wet, *Goodnes-of-fit tests for location and scale families based on a weighted l2-wasserstein distance measure*, *Test* **11** (2002), no. 1 89–107.

[28] Y. Wang, *Smoothing Spline Methods and Application*. Chapman and Hall, 2011.

[29] B. W. Silverman, *Spline smoothing: the equivalent variable kernel method*, *Annals of Statistics* **12** (1984) 898–916.

[30] K. V. Mardia, *Measures of multivariate skewness and kurtosis with applications*, *Biometrika* **57** (1970), no. 3 519–530.

[31] K. V. Mardia, *Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies*, *Sankhyā: The Indian Journal of Statistics, Series B* (1974).

[32] N. Henze and B. Zirkler, *A class of invariant consistent tests for multivariate normality*, *Communications in Statistics-Theory and Methods* **19** (1990), no. 10 3595–3617.

[33] J. Royston, *An extension of shapiro and wilk's w test for normality to large samples*, *Applied Statistics* (1982).

[34] J. Royston, *Some techniques for assessing multivarate normality based on the shapiro-wilk w*, *Applied Statistics* (1983).

[35] P. Royston, *Approximating the shapiro-wilk w-test for non-normality*, *Statistics and Computing* **2** (1992), no. 3 117–119.

[36] G. J. Székely and M. L. Rizzo, *A new test for multivariate normality*, *Journal of Multivariate Analysis* **93** (2005), no. 1 58–80.

[37] J. Kellner and A. Celisse, *New normality test in high dimension with kernel methods*, *arXiv preprint arXiv:1404.3188* (2014).

[38] S. Korkmaz, D. Goksuluk, and G. Zararsiz, *Mvn: an r package for assessing multivariate normality*, *The R Journal* **6** (2014), no. 2 151–162.

[39] R. L. Horswell, *A Monte Carlo comparison of tests for multivariate normality based on multivariate skewness and kurtosis*. 1990.

[40] K. V. Mardia and K. Foster, *Omnibus tests of multinormality based on skewness and kurtosis*, *Communications in Statistics-theory and methods* **12** (1983), no. 2 207–221.

[41] E. B. Wilson and M. M. Hilferty, *The distribution of chi-square*, *Proceedings of the National Academy of Sciences* **17** (1931), no. 12 684–688.

[42] R. L. Horswell and S. W. Looney, *A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis*, *Journal of Statistical Computation and Simulation* **42** (1992), no. 1-2 21–38.

[43] S. S. Shapiro and M. B. Wilk, *An analysis of variance test for normality (complete samples)*, *Biometrika* **52** (1965), no. 3/4 591–611.

[44] P. Royston, *Algorithm as 181: The w test for normality. appi*, *Stat* **31** (1982) 176–180.

[45] J. Royston, *Correction: Algorithm as 181: The w test for normality*, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **32** (1983), no. 2 224–224.

[46] N. Henze, *On mardia?s kurtosis test for multivariate normality*, *Communications in statistics-theory and methods* **23** (1994), no. 4 1031–1045.

[47] G. Wahba, *Spline Models for Observational Data*. SIAM, Philadelphia, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

[48] N. Heckman and J. O. Ramsay, *Penalized regression with model-based penalties*, *Canadian Journal of Statistics* **28** (2000) 241–258.

[49] T. W. Anderson and D. A. Darling, *A test of goodness of fit*, *Journal of the American Statistical Association* **49** (1954), no. 268 765–769.

[50] M. A. Stephens, *EDF statistics for goodness of fit and some comparisons*, *Journal of the American Statistical Association* **69** (1974), no. 347 730–737.

[51] M. A. Stephens, *Tests Based on EDF Statistics*, vol. 68 of *Statistics: Textbooks and Monographs*, ch. 4, pp. 97–191. Marcel Dekker, Inc., New York, 1986.

[52] K. Pearson, *Contributions to the mathematical theory of evolution*, *Philosophical Transactions of the Royal Society of London. A* **185** (1894) 71–110.

[53] M. Shakil, B. G. Kibria, and J. N. Singh, *A new family of distributions based on the generalized pearson differential equation with some applications*, *Austrian Journal of Statistics* **39** (2016), no. 3 259–278.

[54] Y. V. Romantsova, *On an asymptotic goodness-of-fit test for a two-parameter gamma-distribution*, *Journal of Mathematical Sciences* **81** (1996), no. 4 2759–2765.

[55] J. Del Castillo and P. Puig, *Testing departures from gamma, rayleigh and truncated normal distributions*, *Annals of the Institute of Statistical Mathematics* **49** (1997), no. 2 255–269.

[56] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

[57] C. Gu, *Smoothing spline anova models: R package gss*, *Journal of Statistical Software* **58** (6, 2014).

[58] T. Heskes, *Bias/variance decompositions for likelihood-based estimators*, *Neural Computation* **10** (1998), no. 6 1425–1433.

[59] L. Bordes, S. Mottelet, P. Vandekerkhove, *et. al.*, *Semiparametric estimation of a two-component mixture model*, *The Annals of Statistics* **34** (2006), no. 3 1204–1232.

[60] Y. Ma, W. Yao, *et. al.*, *Flexible estimation of a semiparametric two-component mixture model with one parametric component*, *Electronic Journal of Statistics* **9** (2015), no. 1 444–474.

[61] J. A. Nelder and R. Mead, *A simplex method for function minimization*, *The computer journal* **7** (1965), no. 4 308–313.

[62] C. J. Potgieter and F. Lombard, *Nonparametric estimation of location and scale parameters*, *Computational Statistics & Data Analysis* **56** (2012), no. 12 4327–4337.

[63] G. Cheng and Z. Shang, *Joint asymptotics for semi-nonparametric regression models with partially linear structure*, *The Annals of Statistics* **43** (2015), no. 3 1351–1390.

[64] A. Meir and E. Keeler, *A theorem on contraction mappings*, *Journal of Mathematical Analysis and Applications* **28** (1969), no. 2 326–329.

[65] D. W. Scott, *The curse of dimensionality and dimension reduction*, *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition* (1992) 217–240.

[66] S. J. Sheather and M. C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation*, *Journal of the Royal Statistical Society. Series B (Methodological)* (1991) 683–690.

[67] J. Copas and M. Fryer, *Density estimation and suicide risks in psychiatric treatment*, *Journal of the Royal Statistical Society. Series A (General)* (1980) 167–176.

[68] M. P. Wand, J. S. Marron, and D. Ruppert, *Transformations in density estimation*, *Journal of the American Statistical Association* **86** (1991), no. 414 343–353.

[69] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression*. No. 12. Cambridge university press, 2003.

[70] W. Härdle, *Smoothing techniques: with implementation in S.* Springer Science & Business Media, 2012.