Submitted to the Annals of Statistics

ESTIMATION OF THE NUMBER OF COMPONENTS OF NON-PARAMETRIC MULTIVARIATE FINITE MIXTURE MODELS

By Caleb Kwon [*] , and Eric Mbakop [†]

Harvard University[*] and University of Calgary[†]

We propose a novel estimator for the number of mixture components (denoted by M) in a non-parametric finite mixture model. The setting that we consider is one where the analyst has repeated observations of $K \geq 2$ variables that are conditionally independent given a finitely supported latent variable with M support points. Under a mild assumption on the joint distribution of the observed and latent variables, we show that an integral operator T that is identified from the data has rank equal to M. We use this observation, in conjunction with the fact that singular values of operators are stable under perturbations, to propose an estimator of M which essentially consists of a thresholding rule that counts the number of singular values of a consistent estimator of T that are greater than a data-driven threshold. We prove that our estimator of M is consistent, and establish some non-asymptotic results which provide finite sample performance guarantees for our estimator. We present a Monte Carlo study which shows that our estimator performs well for samples of moderate size.

1. Introduction. Finite mixture models provide a flexible means to model unobserved heterogeneity, and their usage spans across several disciplines including social sciences, medicine, biology and engineering. We refer the reader to Compiani and Kitamura [7] and McLachlan and Peel [23] for a discussion of their usage in economics and other disciplines.

This paper derives a novel estimator for the number of mixture components in a non-parametric finite mixture model. We consider a setting where the analyst observes an i.i.d sample of $K \geq 2$ variables $(X_1, X_2, \cdots, X_K)$ that are assumed to be independent (but not necessarily identically distributed) given some finitely supported latent variable Θ ($\Theta \in \{1, \cdots, M\}$), i.e,

MSC 2010 subject classifications: Primary 62G05; secondary 62G15, 62H30, 47A55, 47G10, 47N30

Keywords and phrases: Finite mixture model, latent model, nonparametric mixture, conditional independence, multivariate data

$$F(x) = F(x_1, \cdots, x_K) = \sum_{m=1}^{M} P(\Theta = m) \prod_{k=1}^{K} F_k^m(x_k), \tag{1.1}$$

where $F(x_1, \cdots, x_K)$ denotes the distribution of $X = (X_1, X_2, \cdots, X_K)$ (which is identified from the data), and each *mixture component* $\prod_{K=1}^{K} F_k^m(x_k)$, for $m \in \{1, \cdots, M\}$, represents the distribution of X conditional on $\{\Theta = m\}$ (the latter being equal to the product of the marginals under the conditional independence assumption). Here we do not impose any parametric assumption on the distribution of the mixture components. It was shown in Allman, Matias, and Rhodes [8] (Theorem 8 and 9) that if $K \geq 3$ and the component distributions $\{F_k^m\}_{m=1}^M$ are linearly independent (for each $k \in \{1, \cdots, K\}$), then the representation 1.1 is unique up to swaps of the labels of the mixture components. Hence the joint distribution of (X, Θ) is identified (up to label swapping) from that of X (see also Hall and Zhou [9], Hettmansperger and Thomas [11], and Hall et al. [10]). Moreover, when $K \geq 2$ and the component distributions are linearly independent, Kasahara and Shimotsu [18] show that the number of mixture components M is identified. In this paper we provide a new proof of the latter fact. We show that an integral operator T that is identified from the distribution of X has finite rank equal to M, and we use this observation to construct a consistent estimator of M. Indeed, we prove that a thresholding rule which essentially counts the number of singular values of of a consistent estimator $\hat{T}$ of T (in the operator norm) greater than a sample size dependent threshold, yields a consistent estimator of M. For implementation of our estimator, we provide simple numerical procedures to compute the singular values of $\hat{T}$ and the threshold rule.

An example of a setting (in economics) where the mixture representation of equation 1.1 arises, is the study of first and second-price auctions with private values and unobserved heterogeneity. In Hu, McAdams, and Shum [15] (for instance) the authors consider an auction model where bidders' valuations for the auctioned object are independent given an unobserved heterogeneity Θ. There, Θ represents characteristics of the auctioned object that are commonly observed by the bidders (and affect their valuations), but which are not observed by the analyst. The conditional independence of bidders' valuation given Θ implies that the bids (which by assumption are observed by the analyst) are also independent given Θ and thus satisfy equation 1.1, where X now represents the vector of observed bids. The goal is to recover the joint distribution of bids and unobserved heterogeneity (all the terms on the right-hand side of equation 1.1) from the distribution of the ob-

served bids (the term of the left-hand side of 1.1). Once the joint distribution of bids and unobserved heterogeneity is identified, standard results from the auction literature (see Hu, McAdams, and Shum [15]) can be used to identify the joint distribution of valuations and unobserved heterogeneity, from which the analyst can then perform counter-factual analysis under different auction environments (see also Hu [14], Kasahara and Shimotsu [17], Hu, McAdams and Shum [15], An, Hu and Shum [2], Hu and Shum [16], Aguirregabiria and Mira [1], and Xiao [26] for other instances in economics where modelling assumptions give rise to the mixture structure of 1.1). Although Hu, McAdams, and Shum [15] show that the number of mixture components M is identified in their model, they do not provide a way to estimate it and simply assume it to be known when they estimate the mixture model (Bonhomme, Jochmans, and Robin [5], Bonhomme, Jochmans, and Robin [6], Levine, Hunter, and Chauveau [21] and Benaglia, Chauveau, and Hunter [3] also provide estimators of the mixture model 1.1 under the assumption that M is known). However, incorrectly specifying the number of mixture components can lead to incorrect inference of the model's parameters. Under the identifying assumption of Hu, McAdams, and Shum [15], our procedure provides a consistent estimator of the number of mixture components M, and can thus be viewed as a first step toward estimating the mixture model 1.1.

A paper closely related to ours is Kasahara and Shimotsu [18] which studies the identification and estimation of M (or lower bounds on M) in Equation 1.1, and as in this paper, does not impose any parametric restrictions on the distribution of (X, Θ). There, it is shown that when $K = 2$ (for instance), some matrices P_Δ − each one associated to a rectangular partition Δ of the support of $X = (X_1, X_2)$ − are identified from the distribution of X and have rank at most M (see Section 2.3). Moreover, under the linear independence assumption, Kasahara and Shimotsu [18] show that there exist some *good* partitions Δ for which the associated matrices P_Δ have rank equal to M. However, those *good* partitions Δ for which the matrices P_Δ have rank equal to M depend on the distribution of X, and in general (for an arbitrary partition Δ) the rank of P_Δ is only a lower bound on M. The approach of Kasahara and Shimotsu [18] consists in estimating the rank of P_Δ for a partition Δ chosen at the discretion of the analyst. We show below (Section 2.3) that our approach is very much related to theirs. Indeed, when the components of X are continuous, the matrix P_Δ can be seen as a restriction of our operator T to the finite dimensional subspace of piecewise constant functions on the partition Δ (see Proposition 2.4 below).

Our estimator offers many advantages over that of Kasahara and Shi-

motsu [18]. First, under the linear independence assumption, our estimator always consistently estimates the number of mixture components, whereas that of Kasahara and Shimotsu [18] is in general only consistent to a lower bound on M. Hence, to our knowledge, our paper is the first one in the literature to provide a consistent estimator of M under the linear independence assumption. Secondly, when the linear independence assumption does not hold, our estimator is consistent to a lower bound on the number of mixture components which is always at least as large as the lower bound estimated by the method of Kasahara and Shimotsu [18]. Thirdly, we establish non-asymptotic results which provide finite sample performance guarantees for our estimator. In contrast, all the results of Kasahara and Shimotsu [18] are asymptotic in nature, and they do not provide results to assess the finite sample performance of their procedure. Fourthly, unlike the procedure of Kasahara and Shimotsu [18], our procedure does not require the analyst to have knowledge of a good upper bound M_0 on M. We show in a simulation study that for moderate sample sizes, the performance of our procedure is comparable to theirs when M_0 is *slightly* larger than M, and that having M_0 much larger or much smaller than M can lead to a significant reduction in the performance of their procedure. This makes our procedure relatively more appealing in empirical settings where (bounds on) M can plausibly take a wide range of values.

The rest of the paper is organized as follows. In Section 2 we introduce the model and provide our main identification results which relate the number of mixture components M to the rank of an integral operator T, and in Section 2.3 we discuss the connection between our approach and that of Kasahara and Shimotsu [18]. Using our identification argument, we provide in Section 3 an estimator for M, and establish some of its statistical properties. Section 4 presents our Monte Carlo study, and all proofs are provided in Section 5

Notation Given a continuous linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable Hilbert spaces, we will use $\|\cdot\|$ to denote the operator norm defined by $\|T\| := \sup_{\{\omega\in\mathcal{H}_1,\ \|\omega\|_{\mathcal{H}_1}=1\}} \|T(\omega)\|_{\mathcal{H}_2}$, where $\|\cdot\|_{\mathcal{H}_1}$ and $\|\cdot\|_{\mathcal{H}_2}$ denote the norms associated with the inner product on $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. For $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$, $g \otimes f$ denotes their tensor product, which is the rank-one operator defined by $g \otimes f : \mathcal{H}_1 \to \mathcal{H}_2$ with $g \otimes f(\omega) = g\langle f, \omega\rangle_1$, where $\omega \in \mathcal{H}_1$ and $\langle\cdot,\cdot\rangle_1$ denotes the inner product on $\mathcal{H}_1$. When T is compact, we use $\sigma_1(T) \geq \sigma_2(T) \geq \cdots$ to denote the singular values of T in decreasing order (repeated according to their multiplicities). When $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, we use $\|T\|_{HS}$ to denote the Hilbert-Schmidt norm of T defined by $\|T\|_{HS}^2 := \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left(\langle e_j, T(e_i)\rangle_{\mathcal{H}}\right)^2$ where $\{e_i\}_{i=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$ (the sum is independent of the choice of the basis). be treated as one.

2. Model and identification. We consider a K-variate ($K \geq 2$) finite mixture model where the observed random vectors $X_1, \cdots, X_K$ are conditionally independent given some latent variable Θ, as described by equation 1.1. We refer to each X_k ($k = 1, \cdots, K$) as a *component of* X, and the X_k's can be either discrete or continuous. Our goal is to estimate the number of mixture components M in equation 1.1 from an i.i.d sample of X. It can be shown (see Allman, Matias, and Rhodes [8]) that in general, there are distributions F that admit at least two mixture representations as in equation 1.1, with different numbers of mixture components. However, from Proposition 3 of Kasahara and Shimotsu [18] (see also Theorem 8 and 9 of Allman, Matias, and Rhodes [8]), when $K \geq 2$ the number of mixture components is identified from the distribution of X (i.e., all possible representation of the type 1.1 will have the same number of mixture components) if the conditional distributions of the components of X given Θ satisfy a *full rank/linear independence* condition. As the latter is a key assumption needed to identify M, we state it below as a main assumption. We discuss after stating some of our results how the conclusions change when the full rank condition fails.

ASSUMPTION 2.1. **(Full rank /Linear independence)** *There are at least two components X_i and X_j of X ($i, j \in \{1, \cdots, K\}$) for which the corresponding families of conditional distributions $\{F_i^m\}_{m=1}^M$ and $\{F_j^m\}_{m=1}^M$ that appear in equation 1.1 are linearly independent.*

Assumption 2.1 is mild; it is shown in Mbakop [22] (Proposition 7.4) that it holds generically (see also Proposition 2 in Kasahara and Shimotsu [18]). It requires the distribution of at least two components of X varies sufficiently across the M groups. In fact, in the case of a two components mixtures, requiring linear independence of two distributions is equivalent to requiring that they are different (not equal everywhere).

2.1. *The K=2 case.* For simplicity of exposition, we will first consider the case where $K = 2$, and consider the general case further below. We further assume that the components of X are continuously distributed, and that X has a density with respect to the Lebesgue measure. The case with discrete components is somewhat simpler, and will be discussed further below (see Remark 2.3 and 2.4).

Let D_1 (resp. D_2) denote the dimension of X_1 (resp. X_2), and set $D = D_1 + D_2$, i.e., we have $X_1 \in \mathbb{R}^{D_1}$, $X_2 \in \mathbb{R}^{D_2}$, and $X \in \mathbb{R}^D$. We assume that the random vector X has a density with respect to the Lebesgue measure on $\mathbb{R}^D$, denoted f, which is square integrable. In what follows, we assume that $D_1 = D_2 = 1$. The higher dimensional case can be handled similarly.

Note that the density f is identified from the data and can be estimated consistently (at some rate) under additional smoothness assumptions. Let $L^2(\mathbb{R})$ denote the Hilbert space of square integrable functions on $\mathbb{R}$, and let the integral operator T, $T : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, be defined by

$$[T(w)](x_2) = \int_{\mathcal{S}_1} w(x_1) f(x_1, x_2) dx_1, \tag{2.1}$$

for any $w \in L^2(\mathbb{R})$. Note that the operator T is identified from the data (since it is entirely determined by the density f which is identified from the data), and equation 1.1 implies that T has the following representation:

$$T = \sum_{m=1}^{M} \pi_m f_2^m \otimes f_1^m \tag{2.2}$$

where f_i^m ($i \in \{1, 2\}$ and $m \in \{1, \cdots, M\}$) denotes the conditional density of X_i given $\Theta = m$, and $\pi_m = P(\Theta = m)$. The following proposition shows that in general, the operator T has rank (defined as the dimension of the range of T) less than or equal to M. Moreover, when Assumption 2.1 holds, the operator T has rank (exactly) equal to M, and the number of mixture components is identified. The identification of the number of mixture components under Assumption 2.1 was already established in Kasahara and Shimotsu [18] (see Proposition 3 (a)); besides providing an alternative proof of the identification of M, the proposition is useful as it relates M to the rank of the operator T, a fact which we exploit to estimate M (or a lower bound on M). The content of the proposition is similar in spirit to that of Lemma 10 of Elizabeth, Matias, and Rhode [8], and a proof is provided in Section 5

PROPOSITION 2.1. *Suppose that the distribution of $X = (X_1, X_2)$ satisfies a mixture representation of the form given by equation 1.1. Then we have $rank(T) \leq M$. Moreover, if Assumption 2.1 holds, then $rank(T) = M$.*

As a consequence of Proposition 2.1, the operator T is compact, and it admits a singular value decomposition (see Theorem 15.16 in Kress [20]) of the form:

$$T = \sum_{m=1}^{rank(T)} \sigma_m v_m \otimes u_m. \tag{2.3}$$

Here $\{u_m\}_{m=1}^{rank(T)}$ forms an orthonormal basis for the orthogonal complement (with respect to the inner product on $L^2(\mathbb{R})$) to the null space of T,

$\{v_m\}_{m=1}^{rank(T)}$ forms an orthonormal basis for the range of T, and $\{\sigma_m\}_{m=1}^{rank(T)}$ denote the singular values of T which are strictly positive. We exploit this singular value decomposition further below to construct an estimator for M or a lower bound on M, depending on whether or not we maintain Assumption 2.1.

We now introduce a family of operators $\{T_h\}_{(h\geq 0)}$, $T_h : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, which can be thought of as *regularizations* of the operator T, and which are defined by:

$$[T_h(w)](x_2) = \int_{\mathcal{S}_1} w(x_1) f_h(x_1, x_2) dx_1, \tag{2.4}$$

for any $w \in L^2(\mathbb{R})$, and with the function f_h denoting the convolution of the density f with a "product kernel":

$$f_h(x_1, x_2) = \int_{\mathbb{R}^2} f(u, v) K_h(x_1 - u) K_h(x_2 - v) du dv. \tag{2.5}$$

Here $K_h(\cdot) = (1/h)K(\cdot/h)$, where K is some density function (or kernel function in general) on $\mathbb{R}$ – the density of the standard normal for instance (the dependence of T_h on the choice of the *regularizing* density K is left implicit for notational simplicity). As we show in Proposition 2.2 below, $rank(T_h) = rank(T)$ (for all $h > 0$) when the Fourier transform of K vanishes at most on a set of Lebesgue measure zero, and the estimation of $rank(T)$ is equivalent to the estimation of $rank(T_h)$ for any $h > 0$. As we show below in Section 3, the main advantage of the operators T_h over the operator T, is that they admit consistent *unbiased* estimators, and concentration inequalities can be used to derive bounds on their estimation error.

PROPOSITION 2.2. *Let the integral operators T and T_h be defined as in equation 2.1 and 2.4, and let the kernel function K, which appears in the definition of the operator T_h, be any function that is square-integrable with a Fourier Transform that vanishes on a set of measure at most zero. Then $rank(T_h) = rank(T)$ for any $h > 0$, and each operator T_h admits a singular value decomposition*

$$T_h = \sum_{m=1}^{rank(T)} \sigma_m^h \; v_m^h \otimes u_m^h \tag{2.6}$$

with all the singular values $\{\sigma_m^h\}_{m=1}^{rank(T)}$ strictly positive.

REMARK 2.1. *In general, there is no simple expression which relates the mixture representation of equation 2.2 to the singular value decomposition of equation 2.3. However, for some mixture models, both representations coincide and the singular value decomposition is given by the mixture representation. Consider for instance the bi-variate mixture model* $X_1 = \Theta + U$, $X_2 = \Theta + V$, *where* $\{U, V, \Theta\}$ *are independent,* $U \sim V \sim uniform([0,1])$ *and* $Support(\Theta) = \{0,1,2\}$. *For this particular example the mixture representation is given by*

$$T = \pi_0 f_2^0 \otimes f_1^0 + \pi_1 f_2^1 \otimes f_1^1 + \pi_2 f_2^2 \otimes f_1^2 \tag{2.7}$$

where the densities f_1^m *and* f_2^m *are equal to the density of a* $uniform([m, m+1])$, *and* $\pi_m = P(\Theta = m)$. *Since for* $i \in \{1,2\}$ *the densities* $\{f_i^m\}_{m=0}^2$ *have disjoint support, we have* $\int f_i^m(x) f_i^{m'}(x) dx = \delta_{mm'}$ ($\delta_{mm'} = 0$ *if* $m \neq m'$ *and* $\delta_{mm'} = 1$ *otherwise), and the functions* $\{f_i^m\}_{m=0}^2$ *are mutually orthogonal with unit* (L^2) *norm. We thus conclude that the singular value decomposition of the operator* T *is given by*

$$T = \pi_0 f_2^0 \otimes f_1^0 + \pi_1 f_2^1 \otimes f_1^1 + \pi_2 f_2^2 \otimes f_1^2$$

and the singular values $\{\sigma_m\}$ *are given by the proportion of types* $\{\pi_m\}$. *Note that if for each* $i \in \{1,2\}$, *the densities* $\{f_i^m\}_{m=1}^M$ *have disjoint supports but are not necessarily uniformly distributed, then a slight modification of the above argument shows that the singular value decomposition of the operator* T *is now given by*

$$T = \sigma_0 \tilde{f}_2^0 \otimes \tilde{f}_1^0 + \sigma_1 \tilde{f}_2^1 \otimes \tilde{f}_1^1 + \sigma_2 \tilde{f}_2^2 \otimes \tilde{f}_1^2$$

where $\tilde{f}_i^m = f_i^m / \|f_i^m\|$ *and the singular values* σ_m *are given by* $\sigma_m = \pi_m \|f_1^m\| \|f_2^m\|$, *with* $\|f\|$ *denoting the* L^2 *norm of* f. *In our Monte Carlo study (Section 4), we will consider designs given by uniform mixtures of the type given by equation 2.7. We show in Section 3 that the performance of our procedure depends on the magnitude of the singular values of* T, *and the uniform designs of the type given by equation 2.7 will have the advantage that their singular values are known exactly. For the other designs that we consider, we will only know that a singular value decomposition exists, but we will not know the exact magnitudes of the singular values. However, we will be able to obtain estimates of the magnitudes of the singular values through simulations.*

We now provide some heuristics for our estimation procedure. The full details are given below in Section 3. Let $\{X_i\}_{i=1}^N$ be an i.i.d sample of X

(for notational simplicity, we use X_i to denote either the i^{th} observation of the vector X in the sample $\{X_i\}_{i=1}^N$, or the i^{th} component of the vector X; although this may raise some confusion, we think that the correct interpretation of X_i will be clear from the context), and let $\hat{T}_h$ be a consistent estimator of T_h in the Hilbert-Schmidt norm (hence in the operator norm) constructed from the sample $\{X_i\}_{i=1}^N$. Our estimation of $rank(T)$ (equivalently $rank(T_h)$) hinges on the observation that the singular values of T_h are *stable*. Indeed, by Weyl's inequality for singular values (See Horn and Johnson [13] −Inequality 3.3.19 p.178− or H. Weyl (1912)), if $\sigma_1(T) \geq \sigma_2(T) \geq \cdots$ denote the singular values of a compact operator T in non-increasing order (repeated according to their multiplicities), we have

$$|\sigma_i(T) - \sigma_i(T')| \leq \|T - T'\| \tag{2.8}$$

for any compact operators T and T', for any $i \geq 1$. Furthermore, by the Hoffman-Wielandt inequality (see Horn and Johnson [13] − inequality 3.3.32 p.186, which is valid in our setting since all the operators that we consider in this paper are of finite rank) we have

$$\sum_{i\geq 1} |\sigma_i(T) - \sigma_i(T')|^2 \leq \|T - T'\|_{HS}^2. \tag{2.9}$$

As a consequence of inequality 2.8, if $\hat{\tau}_h(N) = o_p(1)$ is such that $P(\|T_h - \hat{T}_h\| > \hat{\tau}_h(N)) \to 0$, then a consistent estimator of $rank(T)$ is given by the number of singular values of $\hat{T}_h$ that are larger than $\hat{\tau}_h(N)$, i.e,

$$\widehat{M} = \#\{i | \sigma_i(\hat{T}_h) \geq \hat{\tau}_h(N)\}. \tag{2.10}$$

Moreover, as a consequence of inequality 2.9, if the threshold $\hat{\tau}_h(N) = o_p(1)$ is now chosen such that $P(\|T_h - \hat{T}_h\|_{HS} > \hat{\tau}_h(N)) \to 0$, then an alternative consistent estimator of $rank(T)$ is given by

$$\widehat{M} = \#\{j \mid \left(\sum_{i\geq j} \sigma_i(\hat{T}_h)^2\right)^{1/2} \geq \hat{\tau}_h(N)\}. \tag{2.11}$$

Indeed, letting $R := rank(T)$ (implying $\sigma_j(T_h) = 0$ for all $j > R$), inequality 2.9 implies that for all $j > R$ we have $\left(\sum_{i\geq j} \sigma_i(\hat{T}_h)^2\right)^{1/2} \leq \|\hat{T}_h - T_h\|_{HS} \leq \hat{\tau}_h(N)$ (with high probability), and that for all $j \leq R$ $\left(\sum_{i\geq j} \sigma_i(\hat{T}_h)^2\right)^{1/2} \to \left(\sum_{i\geq j} \sigma_i(T_h)^2\right)^{1/2}$ which is strictly positive, and thus much larger than (with

high probability) the threshold $\hat{\tau}_h(N)$ $(= o_p(1))$. As the Hilbert-Schmidt norm is a Hilbertian norm, it is easier (for us) to control the estimation error of T_h in the Hilbert-Schmidt norm ($\|\hat{T}_h - T_h\|_{HS}$) than in the operator norm ($\|\hat{T}_h - T_h\|$). Hence, the estimator of $rank(T)$ that we consider in this paper is the one resulting from the Hoffman-Wielandt inequality, equation 2.11, and we leave the investigation of estimators of the type given by equation 2.10 for future research. In Section 3, we provide a consistent estimator $\hat{T}_h$ of T_h, and a data-driven threshold $\hat{\tau}_h(N)$, for the estimator 2.11, which converges in probability to zero (as the sample size $N \to \infty$) and is an upper bound on the estimation error $\|\hat{T}_h - T_h\|_{HS}$ with probability approaching 1 (as $N \to \infty$). We also provide a simple numerical procedure to compute the singular values of $\hat{T}_h$.

REMARK 2.2. *As we recall in Section 2.3, the method of Kasahara and Shimotsu [18] also relates the number of mixture components M to the rank of some operators. Indeed, they show that some matrices P_Δ (defined in equation 2.13 below) have rank at most M, and their estimation procedure is based on estimating the rank of an empirical analogue of P_Δ.*

REMARK 2.3. *A natural extension of the definition of the operator T in 2.1 to the case with discrete components can be obtained by replacing f in equation 2.1 by the probability mass function. When both components of X are discrete (for instance) the operator T reduces to a matrix, and the estimation of M under Assumption 2.1 reduces to the estimation of the rank of a matrix. In the latter setting, the problem becomes essentially finite dimensional, and the method of Kasahara and Shimotsu [18] (like our method) will provide a consistent estimator of $rank(T)$. In fact, as we show below (Proposition 2.4), the operator T in the discrete case is equal to the matrix P_Δ, with Δ given by the finest partition of the support of X. However, when a component of X is continuous, the operator T is a proper infinite dimensional operator. In contrast to the approach of Kasahara and Shimotsu [18] that estimates the rank of a restriction of the operator T to a fixed finite dimensional subspace (see Proposition 2.4) (with the rank of the restriction of T possibly smaller than that of T), the approach of the present paper is fully non-parametric and estimates directly the rank of T.*

REMARK 2.4. *The requirement that $\{F_i^m\}_{m=1}^M$ are linearly independent in Assumption 2.1 puts a restriction on the size of the support of the component X_i if it is discrete: it implies that X_i must have at least M support points.*

2.2. *The general case ($K \geq 2$).* We now consider the case where the observed multivariate vector X has more than two components that are conditionally independent, i.e, $X = (X_1, \cdots, X_K)$ with $K \geq 2$, and equation 1.1 holds. For each $i \in \{1, \cdots, K\}$, let D_i denote the dimension of of the i^{th} component of X, i.e., $X_i \in \mathbb{R}^{D_i}$, let $\mathcal{S}_i \subset \mathbb{R}^{D_i}$ denote the support of X_i, and let $L^2(\mathcal{S}_i)$ denote the space of square integrable functions on $\mathcal{S}_i$. We assume that all the continuous components of X have a joint density with respect to the Lebesgue measure (on the Euclidean space of corresponding dimension). For each $1 \leq i < j \leq K$, let $f_{i,j}$ denote the density (or probability mass function in the discrete case) of the pair (X_i, X_j), and let the (associated) integral operator $T_{i,j} : L^2(\mathcal{S}_i) \to L^2(\mathcal{S}_j)$, which to a square integrable function $w \in L^2(\mathcal{S}_i)$ of the i^{th} component of X assigns the square integrable function $T_{i,j}(w)$ of the j^{th} component of X defined by

$$[T_{i,j}(w)](x_j) = \int_{\mathcal{S}_1} w(x_i) f_{i,j}(x_i, x_j) dx_j. \tag{2.12}$$

The following proposition is a straightforward generalization (or corollary) of Proposition 2.1

PROPOSITION 2.3. *Suppose that the distribution of $X = (X_1, \cdots, X_K)$ ($K \geq 2$) satisfies a mixture representation of the form 1.1. Then for any $1 \leq i < j \leq K$, we have $rank(T_{i,j}) \leq M$. Moreover, if Assumption 2.1 holds, then $\max_{1 \leq i < j \leq K} rank(T_{i,j}) = M$, with the maximal rank being achieved by operators $T_{i,j}$ such that each set of distributions $\{F_i^m\}_{m=1}^M$ and $\{F_j^m\}_{m=1}^M$ is linearly independent.*

2.3. *Connection to the approach of Kasahara and Shimotsu.* In this section, we first give a brief description of the approach proposed by Kasahara and Shimotsu [18], and then discuss how their procedure is related to ours. As done in Kasahara and Shimotsu [18], we focus on the case where $K = 2$. The case where $K > 2$ can be reduced to the case where $K = 2$ by considering an aggregation of the components of X. When $K = 3$ for instance, and the components of $X = (X_1, X_2, X_3)$ are conditionally independent given some latent variable Θ, we can define the variable $X_2' = (X_2, X_3)$ and consider the bivariate random vector $X = (X_1, X_2')$ (whose components are also conditionally independent given Θ) which can be analyzed using the tools developed for the case where $K = 2$.

Let $X = (X_1, X_2)$ denote a bivariate random vector, with X_i supported on $\mathcal{S}_i$ ($i = 1, 2$). Let $\Delta = \Delta^1 \times \Delta^2$ be a rectangular partition of the support of X, with $\Delta^i := \{\delta_1^i, \cdots, \delta_{|\Delta^i|}^i\}$ forming a partition of $\mathcal{S}_i$. Given the partition

Δ, let $P_\Delta \in \mathbb{R}^{|\Delta^1|\times|\Delta^2|}$, denote the matrix with $(i,j)_{th}$ element given by

$$[P_\Delta]_{i,j} = P(X_1 \in \delta_i^1, X_2 \in \delta_j^2). \tag{2.13}$$

The method of Kasahara and Shimotsu [18] hinges on the observation that under the mixture representation of equation 1.1, the matrices P_Δ (for any partition Δ) have rank at most M. Indeed, the conditional independence assumption implies that

$$P(X_1 \in \delta_i^1, X_2 \in \delta_j^2) = \sum_{m=1}^{M} P(\Theta = m)P(X_1 \in \delta_i^1|\Theta = m)P(X_2 \in \delta_j^2|\Theta = m),$$

and the matrix P_Δ can be written as the sum of M rank 1 matrices as follows

$$P_\Delta = \sum_{m=1}^{M} \pi_m P_1^m \otimes P_2^m \tag{2.14}$$

where $\pi_m = P(\Theta = m)$, P_1^m (with a similar definition for P_2^m) is a vector in $\mathbb{R}^{|\Delta^1|}$ with i^{th} element given by $[P_1^m]_i = P(X_1 \in \delta_i^1|\Theta = m)$, and the tensor product $u\otimes v$ here has the simpler interpretation of the vector outer product, i.e, $u\otimes v = uv^T$. As the matrices P_Δ (one for each partition Δ) can be represented as the sum of M rank-one matrices, they each have rank at most M. Therefore, any consistent estimator of the rank of P_Δ (for a given partition Δ) will also be a consistent estimator of a lower bound on M. The approach of Kasahara and Shimotsu [18] essentially consists in constructing such consistent estimators for $rank(P_\Delta)$. In addition, Kasahara and Shimotsu [18] show that under assumption 2.1, there exists at least one partition Δ for which P_Δ has rank M, thus showing that M is identified under Assumption 2.1 (note that such a partition Δ necessarily satisfies $\min\{|\Delta^1|, |\Delta^2|\} \geq M$). However, the *identifying partitions* Δ for which $rank(P_\Delta) = M$ can only be determined from the distribution of X, and Kasahara and Shimotsu [18] do not provide a method for choosing/estimating such identifying partitions in finite sample. As a consequence, their approach is in general only consistent to a lower bound on M, and is consistent for M only in those cases when the partition Δ chosen by the analyst happens to satisfy $rank(P_\Delta) = M$.

We now establish the connection between the two approaches. The following proposition shows that the matrices P_Δ are simply the restrictions of the integral operator T (equation 2.1) to finite dimensional subspaces. A proof is provided in Section 5. Before stating the proposition, we first introduce some notation. Given a partition $\Delta = \Delta^1 \times \Delta^2$, let $\mathcal{M}_{\Delta^i} \subset L^2(\mathcal{S}_i)$, for

$i \in \{1,2\}$, denote subspaces of piecewise constant functions on the elements of the partition Δ^i, defined by

$$\mathcal{M}_{\Delta^i} := \{\omega \in L^2(\mathcal{S}_i) \mid \omega = \sum_{j=1}^{|\Delta^i|} a_j \mathbb{I}_{\delta_j^i}, \text{ with the } a_i's \in \mathbb{R}\}.$$

Note that the subspace $\mathcal{M}_{\Delta^i}$, for $i \in \{1,2\}$, has (finite) dimension equal to $|\Delta^i|$. For $i = 1$ or 2, let $\Gamma_{\Delta^i} : \mathbb{R}^{|\Delta^i|} \to \mathcal{M}_{\Delta^i}$, be defined by $\Gamma_{\Delta^i}(a) = \sum_{j=1}^{|\Delta^i|} a_j \mathbb{I}_{\delta_j^i}$ (for $a \in \mathbb{R}^{|\Delta^i|}$), and let its adjoint $\Gamma^*_{\Delta^i} : L^2(\mathcal{S}_i) \to \mathbb{R}^{|\Delta^i|}$ be the operator which to each element $\omega \in L^2(\mathcal{S}_i)$ assigns the vector $\Gamma^*_{\Delta^i}(\omega) \in \mathbb{R}^{|\Delta^i|}$ with j^{th} component given by $[\Gamma^*_{\Delta^i}(\omega)]_j = \int_{\delta_j^i} \omega(x_i) dx_i$.

PROPOSITION 2.4. *Suppose that $K = 2$ and that the conditional independence (equation 1.1) representation holds. For each partition $\Delta = \Delta^1 \times \Delta^2$, we have*

$$P_\Delta^T = \Gamma^*_{\Delta^2} \circ T \circ \Gamma_{\Delta^1} \tag{2.15}$$

where $\circ$ denotes operator composition. As a consequence, for all Δ we have

$$rank(P_\Delta) \leq rank(T). \tag{2.16}$$

Moreover, there exists at least one partition Δ such that $rank(P_\Delta) = rank(T)$.

REMARK 2.5. *Note that Assumption 2.1 is not needed to establish Proposition 2.4. As a consequence, when Assumption 2.1 does not hold, Proposition 2.1 and 2.4 imply that our approach (which estimates the rank of T) will be consistent to a lower bound on M that is in general at least as large as the lower bound estimated by the procedure of Kasahara and Shimotsu [18]. Moreover, when linear independence holds, our approach will always be consistent for M, whereas that of Kasahara and Shimotsu [18] will in general only be consistent to a lower bound on M. For instance, if the partition Δ is such that $\max\{|\Delta^1|, |\Delta^2|\} < M$, then any consistent estimator of the rank of P_Δ will be asymptotically strictly less than M (with probability approaching 1).*

3. Estimation. In the setting of Section 2.1 ($K = 2$), we propose in this section an estimator of $rank(T)$ based on an i.i.d sample $\{X_i\}_{i=1}^N$ of X, and discuss further below (see Remark 3.4) how to extend the results to the general setting ($K > 2$). The main result of this section is Theorem 3.1 which provides a consistent estimator of $rank(T)$ of the type given

by equation 2.11 with a data-driven threshold, as well as non-asymptotic performance guarantees. The main tools used to derive the results of this section are perturbation theory results (Hoffman-Wielandt inequality-2.9) and concentration inequalities for sums of independent Hilbert space valued random elements (Lemma 2 in Smale and Zhou [25]). Our approach is similar to that taken in Koltchinskii and Gine [19], Zwald and Blanchard [27], Blanchard, Bousquet, and Zwald [4] and Rosasco, Belkin, and De Vito [24], who also combine perturbation theory results and concentration inequalities to study spectral properties of estimates of integral operators.

The estimator $\widehat{M}$ that we propose is based on a consistent estimator $\hat{T}_h$ of T_h. From Proposition 2.2, the operator T and the operators T_h ($h > 0$) have the same rank. However, as we show below (Proposition 3.1 and Proposition 3.2), one main advantage of using the operators T_h's (instead of the operator T) to estimate $rank(T)$ is that the operators T_h can be estimated *without bias* and concentration inequalities readily yield simple parametric ($\sqrt{N}$ rate) data-driven bounds on their estimation errors $\|\hat{T}_h - T_h\|_{HS}$. By contrast, the estimation of the operator T necessarily involves a bias term, which may converge to zero at a very slow non-parametric rate, unless the density f is sufficiently smooth. Moreover, the presence of a bias term makes it difficult to obtain good bounds on $\|\hat{T} - T\|_{HS}$, as bounds on the approximation error $\|T - E\hat{T}\|_{HS}$ necessarily depend on smoothness properties of the density f which may be unknown to the analyst.

We now provide a consistent estimator $\hat{T}_h$ of T_h, and derive further below (Proposition 3.1) a data-driven bound $\hat{\tau}_h(N)$ on the estimation error $\|\hat{T}_h - T_h\|_{HS}$. Note that the function f_h defined in equation 2.5 can be rewritten as

$$f_h(x_1, x_2) = EK_h(x_1 - X_1)K_h(x_2 - X_2). \tag{3.1}$$

Given an i.i.d sample $\{X_i\}_{i=1}^N$, a natural estimator for the operator T_h is given by:

$$[\hat{T}_h(w)](x_2) = \int_{\mathbb{R}} w(x_1)\hat{f}_h(x_1, x_2)dx_1, \tag{3.2}$$

for any $w \in L^2(\mathbb{R})$, with the function $\hat{f}_h$ given by the sample analogue of equation 3.1, i.e,

$$\hat{f}_h(x_1, x_2) = \frac{1}{N}\sum_{i=1}^{N} K_h(x_1 - X_{1i})K_h(x_2 - X_{2i}). \tag{3.3}$$

Since $E\hat{f}_h(x_1, x_2) = f_h(x_1, x_2)$, we have $E\hat{T}_h = T_h$ (see Blanchard, Bousquet and Zwald [4] for the definition of the expectation of a Hilbert space valued

random variable, and note that the random variables $\hat{T}_h$ take their values in the space of Hilbert-Schmidt operators). For each $x = (x_1, x_2) \in \mathbb{R}^2$ and $h > 0$, let $T_{h,x} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ denote the rank-one operator defined by

$$T_{h,x} = K_h(x_2 - \cdot) \otimes K_h(x_1 - \cdot). \tag{3.4}$$

The following proposition provides a non-asymptotic data-driven bound on the estimation error $\|\hat{T}_h - T_h\|_{HS} = \|\hat{T}_h - E\hat{T}_h\|_{HS}$, with $\hat{T}_h$ defined as in equation 3.2. As noted above, the main tools that we use to derive bounds on the estimation error are concentration inequalities. The proof of the proposition is provided in Section 5.

PROPOSITION 3.1. *Let X' be an independent copy of X. For all $0 < \delta < 1$ and for all $N \geq 2$, the following inequality holds with probability greater than $1 - \delta$*

$$\|\hat{T}_h - T_h\|_{HS} \leq \frac{2L_h ln(2/\delta))}{N} + \sqrt{\frac{ln(2/\delta) E\|T_{h,X} - T_{h,X'}\|^2_{HS}}{N}} \tag{3.5}$$

where $L_h =: \sup_{x,x' \in \mathbb{R}^2} \|T_{h,x} - T_{h,x'}\|_{HS}$. Moreover, if $0 < \delta < 1/2$, then the following inequality holds with probability greater than $1 - 2\delta$

$$\|\hat{T}_h - T_h\|_{HS} \leq \frac{2L_h ln(2/\delta))}{N} + \sqrt{\frac{ln(2/\delta)}{N}\left(\frac{1}{N(N-1)}\sum_{i \neq j}\|T_{h,X_i} - T_{h,X_j}\|^2_{HS} + L_h^2\sqrt{\frac{ln(1/\delta)}{N}}\right)} \tag{3.6}$$

REMARK 3.1. *From the proof of Proposition 3.1, the supremum in the definition of L_h can be replaced by the supremum over the support of X (instead of all of $\mathbb{R}^2$). However, we have opted for the supremum over all of $\mathbb{R}^2$ to make L_h distribution free (not dependent on the distribution of X). Note that the bound on the right-hand side of inequality 3.6 can be computed from the data. Indeed, the quantities L_h and $\|T_{h,X_i} - T_{h,X_j}\|^2_{HS}$ can be computed explicitly (or bounded) as they only depend on the kernel K and the bandwidth h, which are both chosen by the analyst. The right-hand size of inequality 3.5, on the other hand, depends on the expectation $E\|T_{h,X} - T_{h,X'}\|^2_{HS}$ and cannot be computed from the data. Although Theorem 3.1 below is established with the threshold given by the right-hand side of inequality 3.6 (equation 3.7 below), when we implement the method in Section 4, the threshold that we use will be based on the right-hand side of inequality 3.5 (where we will replace $E\|T_{h,X} - T_{h,X'}\|^2_{HS}$ by its sample analogue $\frac{1}{N(N-1)}\sum_{i \neq j}\|T_{h,X_i} - T_{h,X_j}\|^2_{HS}$).*

Let $\hat{\tau}_h(N,\delta)$ be defined by

(3.7)
$$\hat{\tau}_h(N,\delta) := \frac{2L_h ln(2/\delta))}{N} + \sqrt{\frac{ln(2/\delta)}{N}\left(\frac{1}{N(N-1)}\sum_{i\neq j}\|T_{h,X_i} - T_{h,X_j}\|_{HS}^2 + L_h^2\sqrt{\frac{ln(1/\delta)}{N}}\right)}$$

and for each $j \in \{1,\cdots,N\}$, define

$$r_j(\hat{T}_h) := \sqrt{\sum_{i\geq j}\sigma_i(\hat{T}_h)^2}. \tag{3.8}$$

Note that $\hat{\tau}_h(N,\delta) = o_P(1)$. The following theorem is the main result of this section, and is a direct consequence of Proposition 3.1 and Hoffman-Wielandt inequality (equation 2.9).

THEOREM 3.1. *Suppose that the distribution of X satisfies the mixture representation of equation 1.1, and for $h > 0$, let $\hat{T}_h$ be defined by equation 3.2. Consider the estimator of $rank(T)$ given by:*

$$\widehat{M} = \#\{j \mid r_j(\hat{T}_h) \geq \hat{\tau}_h(N,\delta)\}, \tag{3.9}$$

where $\hat{\tau}_h(N,\delta)$ and $r_j(\hat{T}_h)$ are defined as in equation 3.7 and 3.8. Then, for any $0 < \delta < 1/2$, we have

$$P(\widehat{M} \leq rank(T)) \geq 1 - 2\delta, \tag{3.10}$$

(3.11)
$$P(\{\sigma_{rank(T)}(T_h) > 2\hat{\tau}_h(N,\delta)\}\cap\{\|\hat{T}_h - T_h\|_{HS} \leq \hat{\tau}_h(N,\delta)\}) \leq P(\widehat{M} = rank(T)),$$

and

$$P(\{\sigma_{rank(T)}(T_h) + \|\hat{T}_h - T_h\|_{HS} < \hat{\tau}_h(N,\delta)\}) \leq P(\widehat{M} < rank(T)), \tag{3.12}$$

where $\sigma_{rank(T)}(T_h)$ denotes the smallest nonzero singular value of T_h. As a consequence, if $\delta = \delta(N) \to 0$ and $ln(1/\delta(N)) = o(N)$, then $P(\widehat{M} = rank(T)) \to 1$. Moreover, if Assumption 2.1 is satisfied, then $rank(T) = M$ and $\widehat{M}$ is a consistent estimator of M.

REMARK 3.2. *Inequality 3.10 shows that our choice of threshold $\hat{\tau}_h(N,\delta)$ guarantees that $\widehat{M}$ is a lower bound on $rank(T)$ (and hence on M) with*

probability at least $1-2\delta$ *for any* N*. Moreover, Inequality 3.11 shows that* $\widehat{M}$ *is a non-trivial lower bound on* $rank(T)$ *(the trivial lower bound* $\widehat{M} \equiv 1$ *satisfies inequality 3.10), and* $\widehat{M} = rank(T)$ *with high probability whenever the threshold* $\hat{\tau}_h(N,\delta)$ *is much smaller than the the smallest non zero singular value of* T_h *with high probability (which holds true for large* N *as* $\hat{\tau}_h(N,\delta) = o_P(1)$*). This implies in particular that* $\widehat{M}$ *will perform well (will be equal to* $rank(T)$ *with high probability) for designs where the smallest nonzero singular value of* T_h *is well-separated from zero (relative to the sample size). This is confirmed by our simulation studies; see Figure 1 (a) and (b), which correspond to design 2 in Section 4, where the largest nonzero singular value (third in this case) is well away from zero, and note the good performance of our method on this design in the simulation study. By contrast, inequality 3.12 shows that* $\widehat{M}$ *will underestimate* $rank(T)$ *with high probability if the smallest non-zero singular value of the operator* T_h *is close to zero and much smaller than the bound* $\hat{\tau}_h(N,\delta)$ *on the estimation error; see Figure 1 (c) and (d), which correspond to design 1 in Section 4, where (as shown by the figures) the smallest nonzero singular value (third in this case) is close to zero, and note the poor performance of our method on design 1 in the simulation study.*

REMARK 3.3. *The results of Theorem 3.1 are valid for any choice of bandwidth* $h > 0$*, and as noted in Remark 3.2, inequality 3.11 implies that our procedure will correctly estimate* $rank(T)$ *with high probability whenever the smallest non-zero singular value of* T_h *(*$\sigma_{rank(T)}(T_h)$*) is much larger than the bound on the estimation error* $\|T_h - \hat{T}_h\|_{HS}$ *(given by* $\hat{\tau}_h(N,\delta)$*) with high probability. It can also be shown that the smallest nonzero singular values of* T_h *converge to that of* T *as* $h \to 0$ *(Proposition 3.2 below in conjunction with inequality 2.8) , and that the smallest nonzero singular value of* T_h *tends to zero as* $h \to \infty$*. In contrast, for fixed* N*, the bound* $\hat{\tau}_h(N,\delta)$ *on the estimation error* $\|T_h - \hat{T}_h\|_{HS}$ *tends to zero as* $h \to \infty$*, and tends to infinity as* $h \to 0$*. Therefore, for a fixed sample size* N*, values of* h *that are either very large or very small may lead to thresholds* $\hat{\tau}_h(N,\delta)$ *that are much larger than* $\sigma_{rank(T)}(T_h)$*, and inequality 3.12 implies that our procedure will underestimate* $rank(T)$ *for such choices of* h*. We leave the determination of "good" data-driven choices of the bandwidth* h*, as well as the choice of the kernel* K*, for future research. In our simulation studies below (Section 4), we implement* $\widehat{M}$ *with a bandwidth* h *given by Silverman's rule (*$h \sim N^{-1/6}$ *when* $X \in \mathbb{R}^2$*).*

We now suggest an estimator of the operator T, which is shown to be

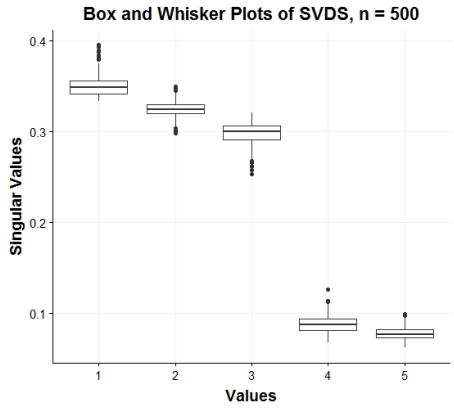
consistent in Proposition 3.2 below. A consequence of Proposition 3.2 (in conjunction with inequality 2.8) is that the ordered singular values of the operator T_h converge to those of the operator T as $h \to 0$. Unlike Proposition 3.1 above, some additional regularity conditions are needed on the density f for our estimator of T to be consistent; in particular, we assume in Proposition 3.2 that f is continuous. It is not difficult to modify the proof of the proposition to obtain a convergence rate for $\|\hat{T} - T\|$ when f satisfies additional regularity conditions (twice differentiable for instance).

PROPOSITION 3.2. *Suppose that the density f is continuous and compactly supported, and suppose that the kernel K used in equation 3.3 is such that $\int K = 1$. Let the estimator of T be given by $\hat{T}_h$ (defined in equation 3.2) with $h = h(N) \to 0$ such that $Nh^2 \to \infty$. Then we have*

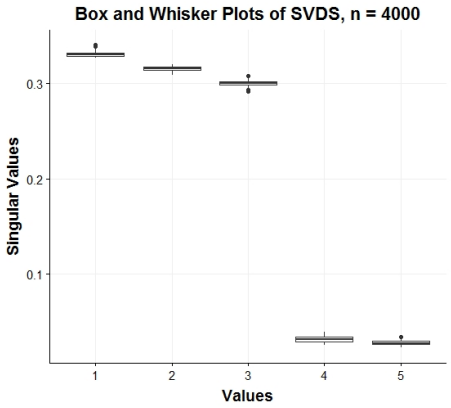
$$E\|\hat{T}_h - T\| \leq \sqrt{E\|\hat{T}_h - T\|_{HS}^2} = o(1). \tag{3.13}$$

The proof of Proposition 3.2 involves the decomposition of the error $\|\hat{T}_h - T\|_{HS}$ into an approximation bias that controls the difference $T - T_h$, and an estimation error that controls the difference $\hat{T}_h - T_h$. The condition $h = h(N) \to 0$ is needed to make the approximation bias converge to zero, and the condition $Nh^2 \to \infty$ is needed to make the estimation error converge to zero. Figure 1 (Box (a) and (b)) provide an illustration of Proposition 3.2; it shows the five largest singular values of the operator $\hat{T}_h$, for $h = .05$. Figures 1 ((a) and (b)) correspond to design 2 in Section 4, where the data is generated from a mixture of three uniforms with equal weights: $\pi_0 = \pi_1 = \pi_3 = 1/3$. As noted in Remark 2.1 (equation 2.7), the nonzero singular values of the operator T for this design coincide with the mixing proportions, and we have $\sigma_1(T) = \sigma_2(T) = \sigma_3(T) = 1/3$. Note that the 3 largest singular values of the estimator $\hat{T}_h$ plotted in Figure 1 (Box (a) and (b)) are all close to $1/3$.

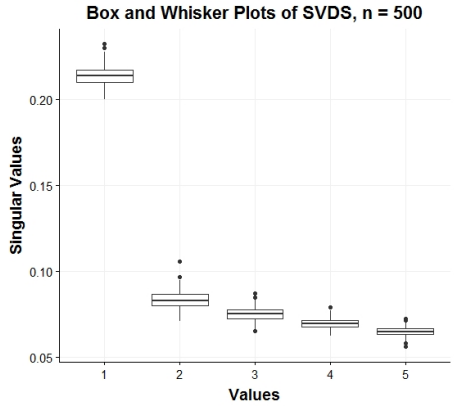
REMARK 3.4. *When $K = 2$ and one of the components X_1 and X_2 has dimension greater than one, say $X_1 = (X_{11}, X_{22})$ with X_{11} and X_{12} of dimension one, then the construction of this section can be applied to the operator associated with the pairs (X_{11}, X_2) and (X_{22}, X_2), and we can take as an estimator the maximum of the estimates of the ranks of operators associated with (X_{11}, X_2) and (X_{22}, X_2). The same procedure can be applied to the case $K \geq 2$*



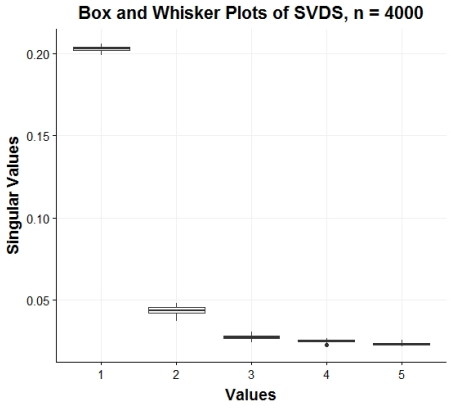
(a) Mixture of 3 uniform distributions ($n = 500$)



(b) Mixture of 3 uniform distributions ($n = 4000$)



(c) Mixture of 3 normal distributions ($n = 500$)



(d) Mixture of 3 normal distributions ($n = 4000$)

Fig 1: Box and Whisker plot of the largest five singular values of $\hat{T}$ computed from Equation 3.14, with $h = 0.05$. Box (a) and (b) corresponds to data generated from a mixture of 3 uniform distributions (design 2 in Section 4), and Box (c) and (d) to a mixture of 3 Normal distributions (design 1 in Section 4). Note that the three largest singular values of the uniform design are all close to 1/3 (see Remark 2.1).

3.1. *Computation of singular values.* To evaluate $\widehat{M}$ in Theorem 3.1, it is necessary to provide a procedure for computing the singular values of $\hat{T}_h$. Let $\hat{T}_h$ be as in equation 3.2. As we show in Section 5, the singular values of $\hat{T}_h$ are equal to the singular values of the matrix $\hat{A}_h$ defined by

$$\hat{A}_h = \frac{\hat{W}_{2h}^{1/2}\hat{W}_{1h}^{1/2}}{N}, \tag{3.14}$$

with the matrices $\hat{W}_{1h}$ and $\hat{W}_{2h} \in \mathbb{R}^{N\times N}$ given by

$$[\hat{W}_{1h}]_{i,j} = \phi_h(X_{1i}, X_{1j}) \quad \text{and} \quad [\hat{W}_{2h}]_{i,j} = \phi_h(X_{2i}, X_{2j}) \tag{3.15}$$

for $1 \le i, j \le N$, and where the function ϕ_h, with domain $\mathbb{R}^2$, is defined by

$$\phi_h(a,b) = \int K_h(a-u)K_h(b-u)du, \tag{3.16}$$

which can be computed in closed form for many choices of the kernel K: for instance, $\phi_h(a,b) = (2h\sqrt{\pi})^{-1}\exp\left(-\frac{(a-b)^2}{4h^2}\right)$ if the kernel K is Gaussian ($K(x) = \frac{\exp -x^2/2}{\sqrt{2\pi}}$), and $\phi_h(a,b) = \mathbf{1}\{|a-b| \le 2h\}\frac{2h-|a-b|}{4h^2}$ if the kernel K is uniform ($K(x) = (1/2)\mathbf{1}\{|x| \le 1\}$). We state the foregoing observations as a corollary.

COROLLARY 3.1. *The estimator $\widehat{M}$ of Lemma 3.1 with $\hat{T}$ given by 3.2, is equivalently given by*

$$\widehat{M} = \#\{j \mid r_j(\hat{A}_h) \ge \hat{\tau}_h(N,\delta)\}, \tag{3.17}$$

where the matrix $\hat{A}_h$ is as defined in equation 3.14, and $r_j(\cdot)$ is defined as in equation 3.8.

3.2. *Computation of the threshold rule.* In this section, we provide a numerical procedure to compute the threshold $\hat{\tau}_h(N,\delta)$. When we implement our method in Section 4 below, we will use the threshold suggested by the right-hand side of inequality 3.5, as opposed to the one suggested by the right-hand side of inequality 3.6, i.e, we will use

(3.18)
$$\hat{\tau}_h(N,\delta) := \frac{2\hat{L}_h ln(2/\delta))}{N} + \sqrt{\frac{ln(2/\delta)}{N}\left(\frac{1}{N(N-1)}\sum_{i\neq j}\|T_{h,X_i} - T_{h,X_j}\|_{HS}^2\right)}.$$

where $\hat{L}_h$ is a sample analogue of L_h, i.e, $\hat{L}_h =: \sup_{i\neq j}\|T_{h,X_i} - T_{h,X_j}\|_{HS}$. Note that the latter threshold is essentially obtained from the one given by

equation 3.7 by dropping the lowest order term (a term of order $N^{-3/4}$), and although the latter change is not justified by our results, it has no (relevant) effect on the performance of our procedure for large N. For small sample sizes however (N less than 500), we have observed from our simulation studies that using the threshold suggested by equation 3.7 makes our method very conservative ($\widehat{M} < M$ with high probability), and the threshold 3.18 leads to a more reasonable performance.

To implement the threshold suggested by equation 3.18, it suffices to provide a procedure to compute $\|T_{h,x} - T_{h,x'}\|_{HS}^2$ for any $x = (x_1, x_2)$ and $x' = (x_1', x_2')$ in $\mathbb{R}^2$ (note that $\hat{L}_h = \sqrt{\sup_{i \neq j} \|T_{h,X_i} - T_{h,X_j}\|_{HS}^2}$). As the Hilbert-Schmidt norm is an inner product norm, we have

$$\|T_{h,x} - T_{h,x'}\|_{HS}^2 = \|T_{h,x}\|_{HS}^2 + \|T_{h,x'}\|_{HS}^2 - 2\langle T_{h,x}, T_{h,x'}\rangle_{HS},$$

where $\langle \cdot, \cdot \rangle_{HS}$ denotes the Hilbert-Schmidt inner product. A straightforward computation (using the definition of the Hilbert-Schmidt inner product) yields

(3.19)
$$\begin{aligned} &\|T_{h,x} - T_{h,x'}\|_{HS}^2 \\ &= \phi_h(x_1, x_1)\phi_h(x_2, x_2) + \phi_h(x_1', x_1')\phi_h(x_2', x_2') - 2\phi_h(x_1, x_1')\phi_h(x_2, x_2'), \end{aligned}$$

with the function ϕ_h defined by equation 3.16, and the threshold $\hat{\tau}_h(N, \delta)$ of equation 3.19 can be easily computed from 3.18.

4. Monte Carlo Experiments. In this section, we assess the performance of our estimator $\widehat{M}$ on four designs. The performance of $\widehat{M}$ is then compared to the four procedures suggested by Kasahara and Shimotsu [18] (SHT, AIC, BIC and HQ). The designs that we consider have $M = 3$ and $M = 5$ mixture components, and for each design we simulate 500 samples of sizes $N = 500$ and $N = 2000$. To compute $\widehat{M}$ for each synthetic sample, we construct the matrix $\hat{A}_h$ defined in Equation 3.14, and compute its singular values. We use the Gaussian kernel, i.e. $K(x) = \frac{\exp -x^2/2}{\sqrt{2\pi}}$, and the bandwidth h is chosen according to Silverman's rule. Finally, we use the threshold rule $\hat{\tau}_h(N, \delta)$ given by equation 3.18, with $\delta = 0.05$ for all of our simulations. We consider the following four designs when generating samples of X. The first design is from Kasahara and Shimotsu [18], and the other three designs are chosen to highlight different aspects of the data generating process that affect the performance of our procedure.

1. Design 1 (mixture of 3 normal distributions): $(X_1, X_2) \sim \sum_{m=1}^{3} \frac{1}{3}\mathcal{N}_2(\mu^m, I_2)$ with $\mu^1 = (0, 0)'$, $\mu^2 = (1, 2)'$, $\mu^3 = (2, 1)'$, and I_2 is the 2 by 2 identity matrix.

2. Design 2 (mixture of 3 uniform distributions): $X_1 \sim \sum_{m=1}^{3} \frac{1}{3}\mathcal{U}(k_{1,a}^m, k_{1,b}^m)$ and
$X_2 \sim \sum_{m=1}^{3} \frac{1}{3}\mathcal{U}(k_{2,a}^m, k_{2,b}^m)$ with $(k_{1,a}^1, k_{1,b}^1) = (k_{2,a}^1, k_{2,b}^1) = (0,1)$, $(k_{1,a}^2, k_{1,b}^2) = (k_{2,a}^2, k_{2,b}^2) = (1,2)$ and $(k_{1,a}^3, k_{1,b}^3) = (k_{2,a}^3, k_{2,b}^3) = (2,3)$.
3. Design 3 (mixture of 3 normal distributions): $(X_1, X_2) \sim \sum_{m=1}^{3} \frac{1}{3}\mathcal{N}_2(\mu^m, I_2)$ with $\mu^1 = (0,0)'$, $\mu^2 = (3,3)'$, $\mu^3 = (-3,-3)'$, and I_2 is the 2 by 2 identity matrix.
4. Design 4 (mixture of 5 uniform distributions): $X_1 \sim \sum_{m=1}^{5} \frac{1}{5}\mathcal{U}(k_{1,a}^m, k_{1,b}^m)$ and
$X_2 \sim \sum_{m=1}^{5} \frac{1}{5}\mathcal{U}(k_{2,a}^m, k_{2,b}^m)$ with $(k_{1,a}^1, k_{1,b}^1) = (k_{2,a}^1, k_{2,b}^1) = (0,1)$, $(k_{1,a}^2, k_{1,b}^2) = (k_{2,a}^2, k_{2,b}^2) = (1,2)$, $(k_{1,a}^3, k_{1,b}^3) = (k_{2,a}^3, k_{2,b}^3) = (2,3)$, $(k_{1,a}^4, k_{1,b}^4) = (k_{2,a}^4, k_{2,b}^4) = (3,4)$, and $(k_{1,a}^5, k_{1,b}^5) = (k_{2,a}^5, k_{2,b}^5) = (4,5)$.

The outcome of the simulations are presented in the tables below (one table for each design). The implementation of the method of Kasahara and Shimotsu [18] requires us to choose a value for the parameter M_0. We recall that the parameter M_0 in their procedure represents a guess by the analyst of an upper bound on M, and they recommend using a partition $\Delta = \Delta^1 \times \Delta^2$ of size M_0 ($|\Delta^1| = |\Delta^2| = M_0$) when implementing their procedure. We consider the choices $M_0 = 4$ and $M_0 = 8$. The partitions Δ are then constructed by partitioning the supports of X_1 and X_2 into M_0 equiprobable (with respect to the empirical distribution) intervals as suggested by Kasahara and Shimotsu [18].

Table 1: Simulation outcomes for Design 1

		$N = 500$				$N = 2000$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SVT		1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
$M_0 = 4$	SHT	0.021	0.891	0.082	0.006	0.000	0.566	0.414	0.020
	AIC	0.004	0.757	0.215	0.024	0.000	0.317	0.609	0.074
	BIC	0.464	0.533	0.003	0.000	0.000	0.989	0.011	0.000
	HQ	0.092	0.876	0.031	0.001	0.000	0.766	0.226	0.008
$M_0 = 8$	SHT	0.094	0.874	0.032	0.000	0.000	0.690	0.306	0.004
	AIC	0.022	0.830	0.148	0.000	0.000	0.384	0.542	0.074
	BIC	0.704	0.296	0.000	0.000	0.000	1.000	0.000	0.000
	HQ	0.212	0.788	0.000	0.000	0.000	0.954	0.046	0.000

Table 2: Simulation outcomes for Design 2

		$N = 500$				$N = 2000$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SVT		0.000	0.002	0.996	0.000	0.000	0.000	1.000	0.000
$M_0 = 4$	SHT	0.425	0.000	0.575	0.000	0.520	0.000	0.480	0.000
	AIC	0.454	0.000	0.544	0.002	0.452	0.000	0.492	0.056
	BIC	0.410	0.000	0.590	0.000	0.497	0.000	0.458	0.045
	HQ	0.422	0.000	0.578	0.000	0.520	0.000	0.462	0.018
$M_0 = 8$	SHT	0.382	0.013	0.072	0.523	0.478	0.244	0.002	0.276
	AIC	0.362	0.018	0.028	0.592	0.466	0.204	0.000	0.330
	BIC	0.339	0.028	0.140	0.493	0.472	0.224	0.004	0.300
	HQ	0.352	0.018	0.076	0.554	0.476	0.282	0.000	0.242

Table 3: Simulation outcomes for Design 3

		$N = 500$				$N = 2000$			
		$M = 1$	$M = 2$	$M = 3$	$M \geq 4$	$M = 1$	$M = 2$	$M = 3$	$M \geq 4$
SVT		0.072	0.928	0.000	0.000	0.000	0.000	1.000	0.000
$M_0 = 4$	SHT	0.000	0.000	0.980	0.020	0.000	0.000	0.950	0.050
	AIC	0.000	0.000	0.886	0.114	0.000	0.000	0.882	0.118
	BIC	0.000	0.000	1.000	0.000	0.000	0.000	0.992	0.008
	HQ	0.000	0.000	0.978	0.022	0.000	0.000	0.958	0.042
$M_0 = 8$	SHT	0.000	0.000	0.940	0.060	0.000	0.000	0.930	0.070
	AIC	0.000	0.000	0.824	0.176	0.000	0.000	0.806	0.194
	BIC	0.000	0.000	0.992	0.008	0.000	0.000	0.998	0.002
	HQ	0.000	0.000	0.964	0.036	0.000	0.000	0.968	0.032

Table 4: Simulation outcomes for Design 4

		$N = 500$						$N = 2000$					
		$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M \geq 5$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M \geq 5$
SVT		0.000	0.000	0.992	0.008	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
$M_0 = 4$	SHT	0.466	0.000	0.000	0.534	0.000	0.000	0.484	0.000	0.000	0.516	0.000	0.000
	AIC	0.475	0.000	0.000	0.525	0.000	0.000	0.478	0.000	0.000	0.522	0.000	0.000
	BIC	0.481	0.000	0.000	0.519	0.000	0.000	0.470	0.000	0.000	0.530	0.000	0.000
	HQ	0.480	0.000	0.000	0.520	0.000	0.000	0.482	0.000	0.000	0.518	0.000	0.000
$M_0 = 8$	SHT	0.458	0.208	0.013	0.170	0.083	0.068	0.504	0.274	0.000	0.072	0.097	0.053
	AIC	0.430	0.212	0.014	0.213	0.075	0.056	0.458	0.328	0.000	0.078	0.089	0.047
	BIC	0.415	0.225	0.019	0.227	0.067	0.047	0.449	0.325	0.000	0.109	0.086	0.031
	HQ	0.422	0.216	0.022	0.216	0.081	0.043	0.461	0.324	0.000	0.093	0.094	0.028

As a general remark (valid across all designs), and in accordance with inequality 3.10, our method never overestimates M, i.e, $P(\widehat{M} > M) = 0$. However, our method often substantially underestimates M. This is particularly true for Design 1, where even for a sample size of $N = 2000$, our method selects 1 component in all of our Monte Carlo samples. Indeed, even for $N = 4000$ (not reported in the tables), our method selects at most 2 components in all the Monte Carlo samples, and $P(\widehat{M} = 2) \approx .84$. As noted in Remark 3.2, we expect our approach to yield very conservative estimates of M if the singular values of the operator T are very close to zero (relative to the sample size). From Figure 1 (Box (d)), we see that the (estimated) second largest singular value of T in Design 1 is very small (approximately equal to 0.05), which is smaller in magnitude than our bound $\hat{\tau}_h(N, \delta)$ on the estimation error (the average value of the threshold $\hat{\tau}_h(N, \delta)$ in our simulation when $N = 2000$ is 0.0647 with a standard deviation of 0.0029). By contrast, the methods of Kasahara and Shimotsu [18] perform better on Design 1, and AIC selects the correct number of components ($M = 3$) 61% of the time when $N = 2000$.

In Design 2, all nonzero singular values of T are equal to $1/3$ (see Remark 2.2), hence much larger in magnitude than those of design 1. And as can be expected from inequality 3.11, our estimator performs quite well; $\widehat{M}$ always selects 3 components when $N = 2000$, and the selection frequency for $M = 3$ is close to 1 when $N = 500$. By contrast, all the methods of Kasahara and Shimotsu [18] perform poorly on this design, with their best method (BIC) selecting $M = 3$ with a frequency of approximately 50% when $N = 2000$ and $M_0 = 4$. Moreover, all of their estimation procedures tend to substantially overestimate the true number of components when $M_0 = 8$, with AIC

selecting $M \geq 4$ approximately 30% of the time when $N = 2000$. From this design and Design 4 below, we observe that the methods of Kasahara and Shimotsu [18] seem to perform poorly when the support X is "irregular" and the matrix P_Δ is sparse (has many zeros).

Design 3 combines the desirable aspects of Design 1 and 2: the variable X has full support as in Design 1, and the nonzero singular values of the operator T have moderate size as in Design 2 (from simulations $\sigma_3(T) \approx 0.1$). Our method as well as the procedures of Kasahara and Shimotsu [18] perform well on this design, with their methods performing better for smaller sample sizes (BIC selects 3 components in all of our simulations). However, the performance of their procedures decrease when the number of partitions is increased ($M_0 = 8$), and AIC tends to overestimates the number M even when $N = 2000$ (by as much as 20% of the time when $M_0 = 8$) (as noted in Kasahara and Shimotsu [18], the method AIC is not necessarily consistent, and it will tend to overestimate the rank of P_Δ when N is large).

Design 4 is a variation of Design 2 (also a mixture of uniforms), where $M = 5$ and the nonzero singular values of T are smaller (all five nonzero singular values of T are equal to $1/5$). As the nonzero singular values of T are smaller in comparison to those of design 2, the performance of our method deteriorates relative to design 2. Indeed, our method underestimates M with higher frequency and we have $P(\widehat{M} = 3) \approx 1$ (recall that $M = 5$ in Design 4). However, when $N = 2000$ our method selects the true number of components in all of the Monte Carlo samples. As in Design 2, the methods of Kasahara and Shimotsu [18] do not perform well on this design. We recall here that given an upper bound M_0 on M, the procedures of Kasahara and Shimotsu [18] yield an estimate of a lower bound on M that is at most equal to M_0. We see here that when the upper bound is incorrectly specified $M_0 = 4$, all of their procedures select $M = 4$ approximately 50% of the time when $N = 2000$. When $M_0 = 8$, all of their procedures select the true number of components in approximately 10% of the simulations when $N = 2000$. As noted above, the poor performance of their procedures is probably due to the fact that the support of X is highly "irregular" and that the matrices P_Δ are sparse.

5. Proofs.

PROOF. **(Proof of Proposition 2.1)** By equation 1.1 T has the representation $T = \sum_{m=1}^{M} \pi_m f_2^m \otimes f_1^m$. Let $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) denote the subspace of $L^2(\mathcal{S}_1)$ (resp $L^2(\mathcal{S}_2)$) spanned by the functions $\{f_1^m\}_{m=1}^M$ (resp. $\{f_2^m\}_{m=1}^M$). Under Assumption 2.1, the subspaces $\mathcal{M}_1$ and $\mathcal{M}_2$ have dimension M.

Let $\langle \cdot, \cdot \rangle_1$ denote the inner product on $L^2(\mathcal{S}_1)$. For $\omega \in L^2(\mathcal{S}_1)$, we have

$$T(\omega) = \sum_{m=1}^{M} \pi_m f_2^m \langle f_1^m, \omega \rangle_1$$

which is an element of $\mathcal{M}_2$, and the range of the operator T is thus a subspace of $\mathcal{M}_2$ which has dimension at most M (dimension equal to M when 2.1 holds). To show that the range of T has dimension M under Assumption 2.1, it thus suffices to show that each f_2^m belongs to the range of T. Let ω_m be equal to the residual of the projection of f_1^m on the subspace of $\mathcal{M}_1$ spanned by the functions $\{f_1^{m'}\}_{m' \neq m}$ normalized to have norm 1 (with respect to the norm on $L^2(\mathcal{S}_1)$). The latter operation is well defined by the linear independence of the functions $\{f_1^m\}_{m=1}^M$. Then $< \omega_m, f_1^{m'} >= \delta_{mm'}$ (the Kronecker delta), and we have $T(\omega_m) = f_2^m$. We thus conclude that range of T spans $\mathcal{M}_2$ and it has dimension M. □

PROOF. **(Proof of Proposition 2.2)** From equations 2.3, 2.4 and 2.5, we get

$$\begin{aligned}
[T_h(\omega)](x_2) &= \int_{\mathbb{R}} \omega(x_1) \int_{\mathbb{R}^2} f(u,v) K_h(x_1 - u) K_h(x_2 - v) du dv dx_1 \\
&= \int_{\mathbb{R}} \omega(x_1) \int_{\mathbb{R}^2} \sum_{m=1}^{rank(T)} \sigma_m v_m(v) u_m(u) K_h(x_1 - u) K_h(x_2 - v) du dv dx_1 \\
&= \sum_{m=1}^{rank(T)} \sigma_m \; v_m \star K_h(x_2) \int_{\mathbb{R}} \omega(x_1) u_m \star K_h(x_1) dx_1,
\end{aligned}$$

and we conclude that

$$T_h = \sum_{m=1}^{rank(T)} \sigma_m \; v_m \star K_h \otimes u_m \star K_h. \tag{5.1}$$

Here $u_m \star K_h$ (similarly for $v_m \star K_h$) denote the convolution u_m and K_h defined by

$$u_m \star K_h(x_1) = \int_{\mathbb{R}} u_m(u) K_h(x_1 - u) du.$$

Given $\omega \in L^2(\mathbb{R})$, let $\mathcal{F}[\omega]$ denote its Fourier transform. We have $\mathcal{F}[u_m \star K_h] = \mathcal{F}[u_m]\mathcal{F}[K_h]$ and $\mathcal{F}[v_m \star K_h] = \mathcal{F}[v_m]\mathcal{F}[K_h]$, and the linearity and invertibility of the Fourier transform imply that $\{u_m \star K_h\}_{m=1}^{rank(T)}$ is linearly independent if and only if $\{\mathcal{F}[u_m \star K_h]\}_{m=1}^{rank(T)}$ is linearly independent.

Since $\mathcal{F}[K_h]$ is non-vanishing by assumption, the linear independence of $\{\mathcal{F}[u_m \star K_h]\}_{m=1}^{rank(T)}$ is equivalent to that of $\{\mathcal{F}[u_m]\}_{m=1}^{rank(T)}$. By linearity and invertibility of the Fourier transform, the functions $\{\mathcal{F}[u_m]\}_{m=1}^{rank(T)}$ are linearly independent since the functions $\{[u_m\}_{m=1}^{rank(T)}$ are linearly independent (they are orthonormal). We thus conclude that $\{u_m \star K_h\}_{m=1}^{rank(T)}$ and $\{v_m \star K_h\}_{m=1}^{rank(T)}$ are both sets of linearly independent functions. An argument similar to that used in the proof of Proposition 2.1 then yields that the operator T_h given by equation 5.1 has rank equal to $rank(T)$.

□

Proof. **(Proof of Proposition 2.4)** We first establish identity 2.15. Let $a \in \mathbb{R}^{|\Delta^1|}$, $b \in \mathbb{R}^{|\Delta^2|}$, and let $\langle \cdot, \cdot \rangle_2$ denote the inner product on $L^2(\mathcal{S}_2)$. We have

$$\begin{aligned} b^T \Gamma^*_{\Delta^2} \circ T \circ \Gamma_{\Delta^1}(a) &= \langle \Gamma_{\Delta^2}(b), T \circ \Gamma_{\Delta^1}(a) \rangle_2 \\ &= \sum_{i=1}^{|\Delta^1|} \sum_{j=1}^{|\Delta^2|} a_i b_j \int_{\delta_j^2} \int_{\delta_i^1} f(x_1, x_2) dx_1 dx_2 \\ &= \sum_{i=1}^{|\Delta^1|} \sum_{j=1}^{|\Delta^2|} a_i b_j [P_\Delta]_{i,j} \\ &= a^T P_\Delta b, \end{aligned}$$

which establishes identity 2.15, and inequality 2.16 is a direct consequence. We now prove that inequality 2.16 is an equality for some partitions Δ. The singular value decomposition 2.3 of the integral operator T implies that P_Δ has the following representation (contrast to equation 2.14)

$$P_\Delta^T = \sum_{m=1}^{rank(T)} \sigma_m Q_2^m \otimes Q_1^m \tag{5.2}$$

where Q_1^m (with a similar expression for Q_2^m) is a vector in $\mathbb{R}^{|\Delta^1|}$, with i^{th} element given by $[Q_1^m]_i = \int_{\delta_i^1} u_m(x_1) dx_1$ (note that identity 5.2 yields an alternative proof of inequality 2.16). Here the functions u_m are the eigenfunctions of T^*T that appear in the singular value decomposition 2.14. Since the functions $\{u_m\}_{m=1}^{rank(T)}$ (resp. $\{v_m\}_{m=1}^{rank(T)}$) are orthonormal, they are necessarily linearly independent. Hence there exist partitions Δ^1 (resp. Δ^2) of the support of X_1 (resp. X_2) such that the vectors $\{Q_1^m\}_{m=1}^{rank(T)}$ (resp.$\{Q_2^m\}_{m=1}^{rank(T)}$)

are linearly independent (see the proof of Proposition 3−part (a)−in Kasahara and Shimotsu [18]); it then follows by an argument similar to that used in the proof of 2.1 that $rank(P_\Delta) = rank(T)$ for such a partition Δ. □

PROOF. **(Proof of Proposition 3.1)** Let ξ_i be defined by $\xi_i = T_{h,X_i} - ET_{h,X}$, and note that if X' is an independent copy of X, then we have $\|\xi_i\|_{HS} \leq E\|T_{h,X} - T_{h,X'}\|_{HS} \leq L_h$ (with T_{h,X_i} defined as in equation 3.4). Using Lemma 2 in Smale and Zhou [25], where the Hilbert space H is that of the Hilbert-Schmidt operators on $L^2(\mathbb{R})$, and the Hilbert space valued random element is given by ξ_i, we get

$$\|\hat{T}_h - T_h\|_{HS} = \|(1/N)\sum_{i=1}^{N}\xi_i\|_{HS} \leq \frac{2L_h ln(2/\delta))}{N} + \sqrt{\frac{2ln(2/\delta)E\|\xi_i\|_{HS}^2}{N}}$$

with probability greater than $1-\delta$. To obtain inequality 3.5 from the preceding inequality, note that if $X' = (X_1', X_2')$ is an independent copy of $X = (X_1, X_2)$, we have

$$\begin{aligned}
&E\|T_{h,X} - T_{h,X'}\|_{HS}^2 \\
&= 2E\int_{\mathbb{R}^2}\left(K_h(X_1 - x_1)K_h(X_2 - x_2) - E\{K_h(X_1 - x_1)K_h(X_2 - x_2)\}\right)^2 dx_1 dx_2 \\
&= 2E\|T_{h,X} - ET_{h,X}\|_{HS}^2 = 2E\|\xi_i\|_{HS}^2.
\end{aligned}$$

To obtain inequality 3.6 from 3.5 we use Hoeffding's concentration inequality (for U-statistics Hoeffding [12]), which yields:

$$E\|T_{h,X} - T_{h,X'}\|_{HS}^2 \leq \frac{1}{N(N-1)}\sum_{i\neq j}\|T_{h,X_i} - T_{h,X_j}\|_{HS}^2 + L_h^2\sqrt{\frac{ln(1/\delta)}{N}}$$

with probability greater than $1-\delta$.

□

PROOF. **(Proof of Proposition 3.2)** By the definition of the operator norm, we have:

$$\|\hat{T} - T\| = \sup_{\left\{\|w\|_{L^2(\mathcal{S}_1)}\leq 1\right\}}\left[\int_{\mathcal{S}_2}\left(\int_{\mathcal{S}_1} w(x_1)(f(x_1,x_2) - \hat{f}(x_1,x_2))dx_1\right)^2 dx_2\right]^{1/2}.$$

Using Minkowski's integral inequality, the left-hand side is bounded by:

$$\|\hat{T} - T\| \leq \sup_{\left\{\|w\|_{L^2(\mathcal{S}_1)}\leq 1\right\}}\int_{\mathcal{S}_1}|w(x_1)|\left(\int_{\mathcal{S}_2}(f(x_1,x_2) - \hat{f}(x_1,x_2))^2 dx_2\right)^{1/2} dx_1.$$

Cauchy-Schwartz inequality then yields

$$\|\hat{T} - T\|^2 \leq \int_{\mathcal{S}_1} \int_{\mathcal{S}_2} (f(x_1, x_2) - \hat{f}(x_1, x_2))^2 dx_2 dx_1,$$

and we get

$$E\|\hat{T} - T\|^2 \leq \int_{\mathcal{S}_1} \int_{\mathcal{S}_2} E(f(x_1, x_2) - \hat{f}(x_1, x_2))^2 dx_2 dx_1 = E\|\hat{T} - T\|_{HS}^2. \tag{5.3}$$

The middle term of the latter inequality represents the integrated mean-squared error (IMSE) of the estimator of the density $\hat{f}$, which by standard arguments decomposes into a bias and variance term, and the conditions $h \to 0$ and $Nh^2 \to \infty$ imply respectively that the bias and the variance terms converge to zero. The conclusion of Proposition 3.2 then follows from Jensen's inequality. □

PROOF. **(Proof of Lemma 3.1)** Note that under the mixture representation 1.1, the singular values of T satisfy: $\sigma_R(T) > 0$ and $\sigma_{R+1}(T) = 0$, where R denotes the rank of T. Also, by inequality 2.9 and the triangle inequality, for all $j \in \{1, \cdots, N\}$, we have

$$|r_j(\hat{T}_h) - r_j(T_h)| \leq \|T_h - \hat{T}_h\|_{HS}, \tag{5.4}$$

where $r_j(\hat{T}_h)$ is defined as in equation 3.8. Given the result in Proposition 3.1, to establish inequality 3.10, it suffices to show that $\{\|\hat{T}_h - T_h\|_{HS} \leq \hat{\tau}_h(N, \delta)\} \subset \{r_{R+1}(\hat{T}_h) < \hat{\tau}_h(N, \delta)\}$. The latter is a direct consequence of inequality 5.4, as $r_{R+1}(\hat{T}_h) \leq r_{R+1}(T_h) + \|T_h - \hat{T}_h\|_{HS}$ and $r_{R+1}(T_h) = 0$. To establish inequality 3.11, it suffices to show that

$$\begin{aligned} &\{\sigma_R(T_h) > 2\hat{\tau}_h(N, \delta)\} \cap \{\|\hat{T}_h - T_h\|_{HS} \leq \hat{\tau}_h(N, \delta)\} \\ &\quad \subset \{r_R(\hat{T}_h) \geq \hat{\tau}_h(N, \delta\} \cap \{r_{R+1}(\hat{T}_h) < \hat{\tau}_h(N, \delta)\}. \end{aligned} \tag{5.5}$$

From inequality 5.4, we have $r_R(\hat{T}_h) \geq r_R(T_h) - \|T_h - \hat{T}_h\|_{HS} = \sigma_R(T_h) - \|T_h - \hat{T}_h\|_{HS} \geq \hat{\tau}_h(N, \delta)$ on the event $\{\sigma_R(T_h) > 2\hat{\tau}_h(N, \delta)\} \cap \{\|\hat{T}_h - T_h\_HS| \leq \hat{\tau}_h(N, \delta)\}$. In addition, as in the proof of inequality 3.10, $r_{R+1}(\hat{T}_h) < \hat{\tau}_h(N, \delta)$ on the event $\{\|\hat{T}_h - T_h\|_{HS} \leq \hat{\tau}_h(N, \delta)\}$. Therefore, the inclusion 5.5 holds, and inequality 3.11 follows. Finally, to establish inequality 3.12, it suffices to verify the inclusion

$$\{\sigma_R(T_h) + \|\hat{T}_h - T_h\|_{HS} < \hat{\tau}_h(N, \delta)\} \subset \{r_R(\hat{T}_h) < \hat{\tau}_h(N, \delta\}\}, \tag{5.6}$$

which follows from inequality 5.4, as $r_R(\hat{T}_h) \leq r_R(T_h) + \|T_h - \hat{T}_h\|_{HS} = \sigma_R(T_h) + \|T_h - \hat{T}_h\|_{HS}$.

□

PROOF. **(Proof of Corollary 3.1)** Given the random sample $\{X_i\}_{i=1}^N$, define the random vector spaces $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$ by

$$\hat{\mathcal{H}}_1 = span\{K_h(X_{1i}-\cdot)|i=1,\cdots,N\} \text{ and } \hat{\mathcal{H}}_2 = span\{K_h(X_{2i}-\cdot)|i=1,\cdots,N\}.$$

Note that the operator $\hat{T}_h$ has range in $\hat{\mathcal{H}}_2$. Indeed, for $w \in L^2(\mathcal{S}_1)$, we have

$$[\hat{T}_h(w)](x_2) = (1/N)\sum_{i=1}^N K_h(X_{2i}-y)\int_{\mathcal{S}_1} w(x)K_h(X_{1i}-x_1)dx_1. \tag{5.7}$$

Moreover, since the kernel K has compact support, the vector spaces $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$ have dimension equal to N, as long as the X_{1i}'s and the X_{2i}'s are all distinct, and the latter occurs with probability one (it can be easily shown that the functions $\{K_h(X_{1i}-\cdot)\}_{i=1}^N$ are linearly independent if the X_{1i}'s are distinct). Let $\Gamma_1 : \mathbb{R}^N \to \hat{\mathcal{H}}_1$ and $\Gamma_2 : \mathbb{R}^N \to \hat{\mathcal{H}}_2$ be defined by:

$$\Gamma_1(a) = \sum_{i=1}^N a_i K_h(X_{1i}-\cdot)$$

and

$$\Gamma_2(a) = \sum_{i=1}^N a_i K_h(X_{2i}-\cdot),$$

where $a \in \mathbb{R}^N$. Note that

$$\|\Gamma_1(a)\|^2_{L^2(\mathcal{S}_1)} = a^T\hat{W}_{1h}a \text{ and } \|\Gamma_2(a)\|^2_{L^2(\mathcal{S}_2)} = a^T\hat{W}_{2h}a \tag{5.8}$$

where the matrices $\hat{W}_{1h}$ and $\hat{W}_{2h}$ are as defined in equation 3.15. Since the matrices $\hat{W}_{1h}$ and $\hat{W}_{2h}$ are symmetric and positive definite (see equation 3.15, and recall that the functions $\{K_h(X_{1i}-\cdot)\}_{i=1}^N$ are linearly independent with probability one), their powers $(\hat{W}_{1h})^d$ and $(\hat{W}_{2h})^d$, for any $d \in \mathbb{R}$, are well defined. Let $R : \mathbb{R}^N \to \hat{\mathcal{H}}_1$ and $S : \hat{\mathcal{H}}_2 \to \mathbb{R}^N$ be defined by

$$Ra = \Gamma_1(\hat{W}_{1h}^{-1/2}a) \text{ and } S(\Gamma_2(a)) = \hat{W}_{2h}^{1/2}a.$$

It follows from equation 5.8 that the operators S and R are isometries, i.e, $\|Ra\|_{L^2(\mathcal{S}_1)} = \|a\|$ and $\|S(\Gamma_2(a))\| = \|\Gamma_2(a)\|_{L^2(\mathcal{S}_2)}$. Also, using the representation of equation 5.7, it can be shown that

$$\hat{T}_h(\Gamma_1(a)) = (1/N)\Gamma_2(W_1a) \tag{5.9}$$

Let $\langle\cdot,\cdot\rangle$ denote the inner product on $\mathbb{R}^N$. We show below that the operator $\tilde{T}_h = S\hat{T}_hR : \mathbb{R}^N \to \mathbb{R}^N$ has the same singular values as $\hat{T}_h$. Moreover, the

matrix representation of the operator $\tilde{T}_h$ is given by $\hat{A}_h$ in equation 3.14. Indeed, for $a, b \in \mathbb{R}^N$, identity 5.9 yields

$$\begin{aligned}\langle b, \tilde{T}_h a\rangle &= \langle b, S\hat{T}_h R a\rangle \\ &= \langle b, S\hat{T}_h \Gamma_1 (\hat{W}_{1h}^{-1/2} a)\rangle \\ &= \langle b, (1/N) S\Gamma_2 (\hat{W}_{1h}^{1/2} a)\rangle \\ &= \langle b, (1/N)\hat{W}_{2h}^{1/2}\hat{W}_{1h}^{1/2} a\rangle = \langle b, \hat{A}_h a\rangle.\end{aligned}$$

It now remains to show that $\tilde{T}_h$ and $\hat{T}_h$ have the same singular values. This follows by noting that given a singular value decomposition

$$\hat{T}_h = \sum_{i=1}^{N} \sigma_i(\hat{T}_h)\hat{v}_i \otimes \hat{u}_i$$

of $\hat{T}_h$, the operator $\tilde{T}_h$ has the representation

$$\tilde{T}_h = \sum_{i=1}^{N} \sigma_i(\hat{T}_h) S\hat{v}_i \otimes R^*\hat{u}_i, \tag{5.10}$$

where R^* denotes the adjoint of R. Since the sets $\{S\hat{v}_i\}_{i=1}^N$ and $\{R^*\hat{u}_i\}_{i=1}^N$ are orthonormal (R and S are isometries), 5.10 represents a singular value decomposition of $\tilde{T}_h$, and we conclude that $\tilde{T}_h$ and $\hat{T}_h$ have the same singular values. □

Acknowledgements. We thank Ivan Canay, Denis Chetverikov and Joel Horowitz for their helpful comments and suggestions.

References.

[1] Aguirregabiria, V. and Mira, P. (2019). Identification of Games of Incomplete Informationwith Multiple Equilibria and Unobserved Heterogeneity. *Quantitative Economics.* **Forthcoming**.

[2] An, Y., Hu, Y. and Shum, M. (2010). Estimating first-price auctions with an unknown number of bidders: A misclassification approach. *Journal of Econometrics* **157** 328-341.

[3] Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics* **18** 505-526.

[4] Blanchard, G., Bousquet, O. and Zwald, L. (2007). Statistical Properties of Kernel Principal Component Analysis. *Machine Learning* **66** 259-294.

[5] Bonhomme, S., Jochmans, K. and Robin, J.-M. (2014). Nonparametric Estimation of Finite Mixtures from Repeated Measurements. *Journal of the Royal Statistical Society Series B* **78** 211-229.

[6] Bonhomme, S., Jochmans, K. and Robin, J.-M. (2016). Estimating Multivariate Latent-Structure Models. *Annals of Statistics* **44** 540-563.
[7] Compiani, G. and Kitamura, Y. (2016). Using mixtures in econometric models: a brief review and some new results. *Econometrics Journal* **19** 95-127.
[8] Elizabeth, A., Matias, C. and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* **37** 3099-3132.
[9] Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics* **31** 201-224.
[10] Hall, P., Neeman, A., Pakyari, R. and Elmore, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika* **92** 667-678.
[11] Hettmansperger, T. P. and Thomas, H. (2000). Almost Nonparametric Inference for Repeated Measures in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* **62** 811-825.
[12] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58** 13-30.
[13] Horn, R. and Johnson, C. (1991). *Topics in Matrix Analysis.* Cambridge University Press, Cambridge.
[14] Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics* **144** 27-61.
[15] Hu, Y., McAdams, D. and Shum, M. (2013). Identification of first-price auctions with non-separable unobserved heterogeneity. *Journal of Econometrics* **174** 186-193.
[16] Hu, Y. and Shum, M. (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics* **171** 32-44.
[17] Kasahara, H. and Shimotsu, K. (2009). Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices. *Econometrica* **77** 135-175.
[18] Kasahara, H. and Shimotsu, K. (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society Series B* **76** 97-111.
[19] Koltchinskii, V. and Gine, E. (2000). Random Matrix Approximation of Spectra of Integral Operators. *Bernoulli* **6** 113-167.
[20] Kress, R. (1999). *Linear Integral Equations.* Springer, New York.
[21] Levine, M., Hunter, D. R. and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403-416.
[22] Mbakop, E. (2017). Identification of Auctions with Incomplete Bid Data in the Presence of Unobserved heterogeneity Working Papers, University of Calgary.
[23] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.
[24] Rosasco, L., Belkin, M. and De Vito, E. (2010). On Learning with Integral Operators. *Journal of Machine Learning Research* **11** 905-934.
[25] Smale, S. and Zhou, D.-X. (2007). Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation* **26** 153-172.
[26] Xiao, R. (2018). Identification and estimation of incomplete information games with multiple equilibria. *Journal of Econometrics* **203** 328-343.
[27] Zwald, L. and Blanchard, G. (2005). On the Convergence of Eigenspaces in Kernel Principal Component Analysis. *In NIPS.*

Caleb Kwon
Harvard Business School
20 N Harvard St.
Boston, MA 02613
USA
E-mail: calebkwon@g.harvard.edu
URL: http://calebkwon.com

Eric Mbakop
Department of Economics
University of Calgary
2500 University Dr. N.W.
Calgary, Alberta T2N 1N4
Canada
E-mail: eric.mbakop@ucalgary.edu