Nonparametric estimation of distributions with given marginals via Bernstein–Kantorovich polynomials: $L_1$ and pointwise convergence theory

Alessio Sancetta*

Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, UK

Received 21 October 2005
Available online 7 March 2007

Abstract

The copula density is estimated using Bernstein–Kantorovich polynomials. The estimator is the usual one based on the smoothed histogram. Strong consistency is obtained in $L_1$ and pointwise almost everywhere, allowing for dependent data. For $L_1$ convergence, no condition is imposed on the copula density, while for pointwise convergence, the condition imposed on the true copula density appears to be minimal.

© 2007 Elsevier Inc. All rights reserved.

AMS 2000 subject classification: 62G07; 62G05

Keywords: Bernstein polynomial; Kantorovich polynomial; Copula; Nonparametric estimation

1. Introduction and notation

In this paper we consider the problem of density estimation for multivariate distributions with given marginals. By Sklar’s theorem (e.g. [17]), any multivariate distribution can be written as a function, called the copula, whose arguments are the marginal distributions. Suppose $F(z)$, $z \in \mathbb{R}^K$, is a joint distribution function with marginals $F_1, \ldots, F_K$, then

$$F(x) = C(F_1(z_1), \ldots, F_K(z_K)),$$

where $C: [0, 1]^K \to [0, 1]$ is the copula function. This representation is unique at all points of continuity of the marginal distributions. Assume $C$ is absolutely continuous with respect to the Lebesgue measure, so that the copula density, say $f$, is well defined. We shall be concerned with
a smooth estimator of \( f \), say \( f_n \), which converges in \( L_1 \) and pointwise to \( f \) with probability one. In particular, this density estimator will be defined in terms of Bernstein–Kantorovich polynomials so that \( f_n \) converges pointwise (almost everywhere) when \( f (\ln^+ f)^{K-1} \in L_1 ([0,1]^K) \) and in \( L_1 \) for any copula density (i.e. \( f \in L_1 ([0,1]^K) \)), and the estimator is derived from a sample of possibly dependent observations. We shall assume identically distributed random variables, but, as remarked below, this is not required under certain alternative conditions. The contribution of this paper is to obtain a strong universal consistency of \( f_n \) and to obtain pointwise convergence under fairly weak conditions. By universal consistency it is meant that consistency holds for any copula density with no smoothness conditions. Estimation of the copula by Bernstein–Kantorovich polynomials has not attracted much attention (e.g. [16]), however, univariate density estimation by Bernstein–Kantorovich polynomials has been extensively studied. For \( L_1 \) convergence, it can be anticipated that the present results hold under weaker conditions on the order of polynomial than in previous results (if we could extend those to the \( K \) dimensional case).

To put the results of this paper into perspective, note that Kantorovich and Bernstein polynomials are just linear combinations of binomial probabilities and their difference lies in the way the coefficients of the polynomial are derived (details are given in Section 2). For statistical estimation, these coefficients are the same whether we use Kantorovich or Bernstein polynomials, so we use the generic term Bernstein–Kantorovich polynomials. Kantorovich polynomial approximations are defined for \( L_1 ([0,1]^K) \) functions satisfying a Marcinkiewicz–Zygmund condition. Bernstein approximations are only defined for continuous functions on \([0,1]^K\). For statistical estimation in \( L_1 \), this difference is not relevant, as the set of continuous functions with compact support is densely embedded in \( L_1 \). However, for pointwise convergence, it is convenient to directly use the approximation properties of Kantorovich polynomials when we deal with the bias of the estimator.

Univariate density estimation by Bernstein (Kantorovich) polynomials has been suggested by Vitale [22]. Generalization and further study of the convergence properties in the univariate case has been carried out by many authors (e.g. [1–3,8–11,18]), while the multidimensional case has attracted less attention (e.g., [16,20]). Most of these authors consider the mean square error convergence, while the asymptotic distribution of the estimator is studied by a few (e.g. [1,11,18], for univariate densities, and [16] for weak convergence of the Bernstein copula density estimator). The uniform and/or \( L_1 \) convergence is considered by Babu et al. [1], Bouezmarni and Rolin [2] Bouezmarni and Scaillet [3]. The results in the present paper are compared to these last three studies in an effort to obtain universal consistency of the copula density estimator in \( L_1 \) and pointwise, though unlike these studies, our interest lies on the copula density, and hence in the extension to the multivariate case. The reader interested in other issues of inferential nature (e.g. convergence in distribution) is referred to the relevant cited references.

To the author’s knowledge, Babu et al. [1] provide the best up to date results on uniform convergence for univariate density estimators by Bernstein polynomials clearly requiring the underlying density to be continuous. Bouezmarni and Scaillet [3] is a recent reference for universal consistency in \( L_1 \) in the univariate case. On top of the extension to the high dimensional case, we use weaker conditions on the order of polynomial for the \( L_1 \) convergence. Since it is not uncommon for the copula density to have a singularity at some of the edges of \([0,1]^K\) (e.g. the Gaussian copula, the Kimeldorf Sampson copula, etc.) continuity everywhere in \([0,1]^K\) is too strong an assumption for copula densities. For this reason, attention is given to almost everywhere pointwise convergence rather than uniform convergence. By Ergoff’s Theorem, this can be turned into almost uniform convergence (e.g. [12]) in order to avoid continuity conditions. Moreover, we
consider convergence in $L_1$ with probability one, while Bouezmarni and Scaillet [3] consider weak consistency under stronger conditions on the order of polynomial. The merits of $L_1$ convergence have been described by Devroye and Györfi [5,6].

The reason for using Bernstein–Kantorovich polynomials to estimate the copula is that by simple restrictions on the coefficients, we can define a new family of copulae in terms of these polynomials, in which case the estimation is fully parametric [16]. Since Bernstein–Kantorovich polynomials possess interesting approximation properties, the estimation can also be seen as a method of sieves where the order of the polynomial goes to infinity. We shall not remark further on this in this paper, and shall only consider nonparametric estimation of joint densities with given marginals.

The plan for the paper is as follows. In this section we introduce some notation used in the paper. In Section 2 we introduce the copula density estimator and state the result on convergence while the proof is deferred to Section 3.

1.1. Notation

We shall define the following quantities:

$$\mathbb{M}_m^K := \{0, 1, \ldots, m_1\} \times \cdots \times \{0, 1, \ldots, m_K\}, \text{ where } m_k \in \mathbb{N}_+, k = 1, \ldots, K,$$

$$m := (m_1, \ldots, m_K)$$

and for $i \in \mathbb{M}_m^K$

$$p(i, m) := \left[\frac{i_1}{m_1 + 1}, \frac{i_1 + 1}{m_1 + 1}\right] \times \cdots \times \left[\frac{i_K}{m_K + 1}, \frac{i_K + 1}{m_K + 1}\right],$$

$$I_{i,m} := I_{\{U \in p(i,m)\}}, \text{ where } I_{\{\ldots\}} \text{ is the indicator function},$$

$$P_{i(k),m} (u_k) := \left(\frac{m_k}{i_k}ight)^{i(k)} (1 - u_k)^{m(k) - i(k)}, \quad u_k \in [0, 1],$$

$$P^K_{i,m} (u) := \prod_{k=1}^K P_{i(k),m} (u_k), \quad u \in [0, 1]^K.$$ 

We often write $i_k = i(k)$ for legibility reasons. We shall also define $|p(i, m)|$ to be the Lebesgue measure of the parallelepiped $p(i, m)$. (Note that $|p(i, m)|$ is the same for each $i$.) Moreover, $\mathbb{P}_n$ is the empirical measure that assigns mass $1/n$ to each observation, e.g. for some measurable function $g$ and random variables $U_1, \ldots, U_n$, $\mathbb{P}_n g(U) = n^{-1} \sum_{s=1}^n g(U_s)$. For two $K$ dimensional vectors $a$ and $b$, $ab$ is understood as their componentwise product (Hadamard product), $a + 1$ as $a$ plus 1 to each elements of $a$, and inequality relations are again understood componentwise, i.e. $a \leq b$ if $a_k \leq b_k$ for $k = 1, \ldots, K$, where $a = (a_1, \ldots, a_K)$ and similarly for $b$. The sets $C([0, 1]^K)$ and $L_1([0, 1]^K)$ are the spaces of continuous functions and $L_1$ integrable functions with support in $[0, 1]^K$. Finally, almost everywhere and almost surely are abbreviated to a.e. and a.s.

2. Copula density estimation via Kantorovich polynomials

The marginal distributions are known, so we shall suppose that we actually observe a sample from $(U_s)_{s \in \mathbb{Z}}$ which is a sequence of $[0, 1]^K$ uniform random variables. If we know the marginals, there is no loss of generality in considering the uniform random variables obtained by a measurable
transformation of the original sequence of random variables, say \((Z_s)_{s \in \mathbb{Z}}\). In Section 2.1.1 we show that we can always assume this, once the marginals are known.

Define the Kantorovich operator \(K_m\) such that for \(f \in L_1([0, 1]^K)\),

\[
(K_m f) (u) = \sum_{i \in \mathbb{M}_m^K} |p(i, m)|^{-1} \left( \int_{p(i, m)} f(v) \, dv \right) P_{i, m}^K (u)
\]

([4], and [19], for the \(K > 1\) dimensional case). If we replace \(|p(i, m)|^{-1} \left( \int_{p(i, m)} f(v) \, dv \right)\) by \(f(i/m)\) we obtain Bernstein polynomials (division of vectors is understood to be elementwise).

For continuous functions the two quantities are close to each other as \(\min_{k \in \{1, \ldots, K\}} m_k \to \infty\).

It is natural to replace \(\int_{p(i, m)} f(v) \, dv\) by its empirical counterpart \(P_{nIi, m}^K\), and to consider the following estimator:

\[
f_n(u) := \sum_{i \in \mathbb{M}_m^K} |p(i, m)|^{-1} P_{nIi, m}(U) P_{i, m}^K (u).
\]

(2)

Clearly, we would obtain (2) starting from Bernstein polynomials using the histogram at \(i/m\) as an estimator for \(f(i/m)\). We now turn to the technical conditions required to state the convergence result of the paper.

**Condition 1.** \((U_s)_{s \in \mathbb{Z}}\) is a sequence of stationary \([0, 1]^K\) uniform random variables such that \(U_s\) has copula density \(f\). Moreover, \((U_s)_{s \in \mathbb{Z}}\) is a sequence of uniform mixing random variables such that

\[
\left( 1 + \sum_{j=1}^{n-1} \varphi_j \right)^2 \sim n^{1-\varepsilon}
\]

for some \(\varepsilon \in (0, 1]\), where

\[
\varphi_j := \sup_{l \in [n-j-1]} \sup_{A \in \mathcal{F}_r^j, B \in \mathcal{F}_n^j} \left| \Pr (B \mid A) - \Pr (B) \right|
\]

and \(\mathcal{F}_r^j\) is the sigma algebra generated by \(U_r, \ldots, U_t\).

**Condition 2.** As \(n \to \infty\), the following are satisfied:

(i)

\[
\min_{k \in \{1, \ldots, K\}} m_k \to \infty,
\]

(ii)

\[
\max_{k \in \{1, \ldots, K\}} m_k = o \left( \exp \left\{ \frac{\min_{k \in \{1, \ldots, K\}} m_k}{K} \right\} \right).
\]

**Condition 3.** The copula density \(f\) is such that \(f (\ln^+ f)^{K-1}\) is integrable.

**Remark 4.** For any measurable map \(Q\) such that \(Q(Z_s) = (Q_1(Z_{s1}), \ldots, Q_K(Z_{sK}))\), the mixing coefficients of \((Q(Z_s))_{s \in \mathbb{Z}}\) and \((Z_s)_{s \in \mathbb{Z}}\) are the same ((\(Z_s)_{s \in \mathbb{Z}}\) is the sequence from which we obtain the sequence of uniform random variables \((U_s)_{s \in \mathbb{Z}}\), hence it is more convenient to impose the condition on the uniform variables \(U_s\)). To see this, set \(Q_k\) equal to the distribution function of \(Z_{sk}\) when this is continuous (see Section 2.1.1 for more on this). The above mixing condition holds for a variety of processes, e.g. some linear time series processes (Doukhan [23])
for details and Doukhan and Louhichi [24] for discussion on the limitations of this condition. Weaker conditions might work but would require a more technical proof and stronger restrictions on the order of polynomial.

**Remark 5.** Condition 1 can be further weakened to some heterogeneous sequences. In fact, suppose \((U_s)_{s \in \mathbb{N}}\) has an asymptotically mean stationary measure \(\bar{P}\) [25]. Suppose \(\mathbb{E}(P_n I_{\{U \in A\}}) \to \mathbb{P}I_{\{U \in A\}}\) for every Borel set \(A\). Then, we can consider \(f\) to be the copula density corresponding to \(\bar{P}\). The result of the paper applies to this case as well, under minor modifications. Details are left to the interested reader.

**Remark 6.** The second part of Condition 2 only says that we cannot allow for the rate \(m_k \to \infty\) to vary too much across \(k\)’s.

**Remark 7.** Condition 3 implies the following Marcinkiewicz–Zygmund condition ([13, Theorem 4]),
\[
\lim_{\delta(A_u) \to 0} \frac{1}{|A_u|} \int_{A_u} f(v) \, dv = f(u), \text{ a.e.}
\]
where \(A_u\) is a set containing the point \(u \in [0, 1]^K\), \(\delta(A_u)\) is the diameter of this set, and \(|A_u|\) is the Lebesgue measure of \(A_u\). Indeed all that is required for pointwise convergence is the above strong differentiability of the integral of \(f\).

Hence we have the following.

**Theorem 8.** Suppose Conditions 1 and 2 hold. Then,
\[
\int_{[0,1]^K} |f_n(u) - f(u)| \, du \xrightarrow{a.s.} 0
\]
if \(n^\varepsilon |p(i,m)|^{(1-\varepsilon)} \to \infty\) for some \(\varepsilon \in [0, 1/2)\) (with \(\varepsilon\) as in Condition 1).

**Remark 9.** For the sake of comparison, consider the case of summable uniform mixing coefficients, i.e. \(\varepsilon = 1\) (actually all the cited references considered the iid case). Bouezmarni and Scaillet [3] consider the \(L_1([0, 1])\) convergence in probability of density functions estimated by Bernstein polynomials under stronger conditions on the bin size. In the present case, their result would translate to \(n |p(i,m)|^{2} \to \infty\), which is much stronger. Moreover, under this restriction, they cannot achieve a.s. convergence, but only convergence in probability.

**Theorem 10.** Suppose Conditions 1–3 hold. Then,
\[
|f_n(u) - f(u)| \xrightarrow{a.s.} 0, \text{ a.e. in } [0, 1]^K
\]
if \(n^\varepsilon |p(i,m)|^{(2-\varepsilon)} \to \infty\) for some \(\varepsilon \in [0, 1/2)\).

As a Corollary to the previous result we can turn the a.e. pointwise convergence to almost uniform (a.u.) convergence, i.e. uniform convergence over some set \(S \subset [0, 1]^K\) such that \(|S| > 1 - \beta\) for arbitrarily small \(\beta > 0\) (e.g. [12] for precise details).
Corollary 11. Suppose Conditions 1–3 hold. Then, there exists a set $S$ as defined above, such that

$$\sup_{u \in S} |f_n(u) - f(u)| \xrightarrow{a.s.} 0$$

if $n^\varepsilon |p(i, m)|^{(2-\varepsilon)} \to \infty$ for some $\varepsilon \in [0, 1/2)$.

Remark 12. Note that a.u. convergence is slightly stronger than uniform a.e. convergence, as for the later, convergence holds uniformly over sets $S' \subset [0, 1]^K$ such that $|S'| = 1$.

Remark 13. If $f \in C([0, 1]^K)$ we can replace a.u. convergence with uniform convergence and compare with existing results. Hence, assume this here, though for the purpose of copula estimation, this case is uninteresting (as most copula densities are not continuous everywhere in $[0, 1]^K$), but it allows comparison to existing results. Consider the case $\varepsilon = 1$. Bouezmarni and Rolin [2] derive a.s. uniform convergence under conditions that in higher dimensions are stronger than the present ones. Babu et al. [1] obtain a result under conditions that in $k$ dimensions corresponds to $n |p(i, m)| / \ln n \to \infty$. Under independence, this is possible using Bernstein inequality.

2.1. Further remarks

We briefly comment on some issues of practical and theoretical relevance.

2.1.1. Uniform representation for random variables

Suppose $Z = (Z_1, \ldots, Z_K)$ is a vector of random variables with joint distribution function $F$, as in (1). If each $F_k$ in (1) is continuous, the copula is unique, otherwise this is not true (e.g. [17, Corollary of Theorem 1]). If $F_k$ is not continuous, we define

$$\tilde{F}_k(z, v) := \Pr(Z_k < z) + v \Pr(Z_k = z).$$

If $V_k$ and $U_k$ are uniform random variables with values in $[0, 1]$ and $V_k$ is independent of $Z_k$, then, $\tilde{F}_k(Z_k, V_k) \equiv U_k$ (i.e. weakly) and $Z_k \xrightarrow{a.s.} \tilde{F}_k^{-1}(U_k)$ (e.g. [15]). Therefore, the copula can be understood to be the joint distribution of the continuous $[0, 1]$ uniform random variables $\left\{ U_k := \tilde{F}_k(Z_k, V_k), k = 1, \ldots, K \right\}$ (where the $V_k$’s are independent of the $Z_k$’s). Then, there is a unique copula corresponding to these continuous uniform marginals and knowledge of the marginals is enough to find a $[0, 1]^K$ uniform transformation for the data. Hence, there is no loss of generality in assuming that we directly observe a sample from $(U_s)_{s \in \mathbb{Z}}$ where $U_s = (U_{s1}, \ldots, U_{sK})$. In this case, the copula is just the joint distribution of the random vector $U_s$.

2.1.2. Is the estimator in (2) a copula density?

The previous results state that (2) converges to a copula density. However, it is of interest to know if (2) is a copula density for finite $n$ and $m$. Note that (2) is always positive. This can be verified using the properties of Bernstein operators which are positive operators (e.g. [4]). Hence we only need to check that, for $k = 1, \ldots, K$,

$$\int_{[0,1]^{K-1}} f_n(u) \, du_{-k} = 1, \quad \text{where } u_{-k} = (u_1, \ldots, u_{k-1}, u_{k+1}, \ldots, u_K),$$
which stands for integration over all variables but the $k$th one, as the copula has uniform marginals. This is only approximately true. It is not difficult to show that

$$\left| \int_{[0,1]} f_n(u) du_{-k} - 1 \right| \leq (m_k + 1) \mathbb{P}_n I \left\{ U_k \in \left[ \frac{i_k}{m_k + 1}, \frac{i_k + 1}{m_k + 1} \right] \right\} - 1,$$

where $\mathbb{P}_n I \left\{ U_k \in \left[ \frac{i_k}{m_k + 1}, \frac{i_k + 1}{m_k + 1} \right] \right\}$ is the empirical estimator for the mass of a $[0,1]$ uniform random variable $U_k$ over an interval of length $(m_k + 1)^{-1}$. The empirical estimator is equal to

$$(m_k + 1)^{-1}$$

only asymptotically as $n \to \infty$ (see [16], for some details). Hence, (2) is an exact copula density ($\forall m$) only when $n \to \infty$.

2.1.3. Representation for efficient computation

Computation of (2) can be complicated because it requires the histogram estimator $\mathbb{P}_n I_{i,m} (U)$ followed by the application of a $K$ dimensional summation weighted by the binomial mass function. This might not always be convenient for practical purposes. Implementation in statistical software like S-Plus becomes inefficient when $K > 2$. The estimator can be rewritten in a way that is better suited for statistical applications. Define $g_m : [0,1]^K \to \mathbb{M}_m^K$ to be a quantizer, i.e.

$$g_m(u) = i \text{ if } u \in \prod_{k=1}^{K} \left[ \frac{i_k}{m_k + 1}, \frac{i_k + 1}{m_k + 1} \right],$$

which can also be written as $\lfloor (m + 1) u \rfloor$, which is the integer part (componentwise) of $(m + 1) u$. Then, (2) can be rewritten as

$$f_n(u) = \mathbb{P}_n P^K_{g_m(U),m}(u)$$

and computation of $f_n(u)$ is immediate as for usual kernel density estimators.

2.1.4. Empirical marginal estimation

This paper assumes that the marginals are known, but in practice this may not be the case. It is of interest to understand if consistent estimation of the marginals invalidates the convergence results of the estimator. The results are certainly true under strong convergence of the marginals’ estimator when the marginals are continuous. To see this, define $Q(z) := (F_1(z_1), \ldots, F_K(z_K))$, where $F_k$ is the marginal of $Z_{1:k}$. Define $Q_n(z)$ as $Q(z)$ but using strongly consistent marginals’ estimators. Moreover, $Q_n^*$ and $Q_n^{*\times}$ will denote their inverse. Then, a crucial step is to show that $\mathbb{P}_n I \left\{ Q_n(Z) \leq u \right\} \overset{a.s.}{\to} \mathbb{P}_n I \left\{ Q(Z) \leq u \right\}$. To this end

$$\mathbb{P}_n I \left\{ Q_n(Z) \leq u \right\} = \mathbb{P}_n I \left\{ Z \leq Q_n^*(u) \right\} = \mathbb{P}_n I \left\{ Q(Z) \leq Q \left( Q_n^*(u) \right) \right\} = \mathbb{P}_n I \left\{ U \leq Q \left( Q_n^*(u) \right) \right\}.$$
Table 1
Minimum $L_1$ distance.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Kantorovich</th>
<th>Histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 200 400 800</td>
<td>100 200 400 800</td>
</tr>
<tr>
<td>Kimeldor Sampson copula</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low dependence</td>
<td>0.21 0.17 0.14 0.12</td>
<td>0.55 0.49 0.46 0.44</td>
</tr>
<tr>
<td>High dependence</td>
<td>0.39 0.32 0.28 0.24</td>
<td>1.06 1.04 1.03 1.03</td>
</tr>
<tr>
<td>Gaussian copula</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low dependence</td>
<td>0.17 0.14 0.12 0.10</td>
<td>0.51 0.47 0.43 0.42</td>
</tr>
<tr>
<td>High dependence</td>
<td>0.29 0.23 0.19 0.16</td>
<td>1.00 0.98 0.97 0.96</td>
</tr>
</tbody>
</table>

holds without continuity of the marginals (e.g. [7]). A complete study of this issue under minimal condition will be the subject of future research.

2.1.5. Experimental performance of the estimator

Copula density estimation using Bernstein–Kantorovich polynomials appears to perform well. The experimental performance of density estimators using Bernstein polynomials has been considered by several authors (e.g. [1,16], in the case of the copula density). As an illustrative example, consider a sample of $n = 100, 200, 400, 800$ observations from a Kimeldorf Sampson (KS) copula and a Gaussian (G) copula with parameters $\rho_1 \in \{0.76, 3.19\}$ and $\rho_2 \in \{0.416, 0.813\}$ (e.g. Joe [26]). The parameter $\rho_1$ refers to the KS copula and the other one to the G copula and in each case we consider two parameters values such that the corresponding Spearman’s rho for each copula is 0.4 and 0.8, which will be referred to as the Low dependence and High dependence cases in the results reported below. The higher the dependence, the more challenging is the estimation. The true copula density is estimated by (2) and the histogram estimator. The histogram is a natural choice especially if we wish to impose the boundary conditions required for the estimator to be a true copula density. Moreover, it is universally consistent in $L_1$ (e.g. [6]). The estimator in (2) and the histogram are computed over the partition $(p(i, m))_{i \in \mathbb{M}_m}$ with $m_1 = m_2 = 4 : 40 (4)$ and with abuse of notation we suppress the subscripts, i.e. $m = m_1$. The expected $L_1$ distance for the two estimators is computed. The $L_1$ integral is computed using Monte Carlo integration over 1000 simulated uniform $[0,1]^K$ random points. For the expectation, we use averaging over 100 independent simulated samples. Table 1 only reports the minimum expected $L_1$ distance with respect to $m$ for each estimator. In this simulation, it was not uncommon for the minimizer to be on the boundary of the specified range: the optimal $m$ clearly depends on the sample size and the dependence parameter of the copula we are trying to estimate. Fig. 1 reports the expected $L_1$ distance for the 10 different choices of $m$ when the true copula is the KS copula and $n = 200$. The performance of the Bernstein–Kantorovich estimator is superior to the histogram. The choice of $m$ appears not to be so crucial for the performance as for the histogram. These results are in line with the ones reported by Sancetta and Satchell [16] for the $L_2$ distance.

3. Proof of Theorem 8

The proof is split into a few intermediate results, but first we need an important result for binomial random variables.
Lemma 14. Suppose $X_1, \ldots, X_K$ are independent binomial random variables with parameters $(m_k, u_k)$ $k = 1, \ldots, K$, where $m_k$ and $u_k$ stand, respectively, for the number of trials and the probability of success in each trial. For $\alpha \in [0, 1/2)$,

$$\Pr \{|X_k/m_k - u_k| \leq m_k^{-\alpha}, k = 1, \ldots, K\} \geq 1 - 2 \sum_{k=1}^{K} \exp \left\{-2m_k^{1-2\alpha}\right\}.$$
Proof. By Bonferroni’s inequality,
\[
\Pr \left\{ \left| \frac{X_k}{m_k} - u_k \right| \leq m_k^{-\alpha}, \ k = 1, \ldots, K \right\} \geq 1 - \sum_{k=1}^{K} \Pr \left\{ \left| \frac{X_k}{m_k} - u_k \right| > m_k^{-\alpha} \right\}.
\]
Then, apply Hoeffding’s inequality (e.g. [6]) to the terms in the sum. □

Hence we can show that Bernstein polynomials can be truncated.

Lemma 15. Suppose \( f : [0, 1]^K \rightarrow \mathbb{R} \) is a bounded function with support \([0, 1]^K\). Then, for \( \alpha \in [0, 1/2) \),
\[
\sum_{i \in \mathcal{M}_m^K} f \left( \frac{i}{m} \right) P_{i,m}^K (u) \leq \sum_{\left| \frac{i}{m} - u \right| \leq m^{-\alpha}} f \left( \frac{i}{m} \right) P_{i,m}^K (u) + 2K \max_{k \in \{1, \ldots, K\}} \exp \left\{ -2m_k^{1-2\alpha} \right\} \max_{i \in \mathcal{M}_m^K} |f \left( \frac{i}{m} \right)|,
\]
where \( \left| \frac{i}{m} - u \right| \leq m^{-\alpha} \) means \( \max_{k \in \{1, \ldots, K\}} \left( \left| \frac{i_k}{m_k} - u_k \right| - m_k^{-\alpha} \right) \leq 0 \).

Proof. By positivity of \( P_{i,m}^K (u) \),
\[
\sum_{i \in \mathcal{M}_m^K} f \left( \frac{i}{m} \right) P_{i,m}^K (u) = \left( \sum_{\left| \frac{i}{m} - u \right| \leq m^{-\alpha}} f \left( \frac{i}{m} \right) P_{i,m}^K (u) + \sum_{\left| \frac{i}{m} - u \right| > m^{-\alpha}} f \left( \frac{i}{m} \right) P_{i,m}^K (u) \right)
\leq \sum_{\left| \frac{i}{m} - u \right| \leq m^{-\alpha}} f \left( \frac{i}{m} \right) P_{i,m}^K (u) + \max_{i \in \mathcal{M}_m^K} |f \left( \frac{i}{m} \right)| \sum_{\left| \frac{i}{m} - u \right| > m^{-\alpha}} P_{i,m}^K (u)
= I + II,
\]
where \( \left| \frac{i}{m} - u \right| > m^{-\alpha} \) means \( \max_{k \in \{1, \ldots, K\}} \left( \left| \frac{i_k}{m_k} - u_k \right| - m_k^{-\alpha} \right) > 0 \). Suppose \( X := (X_1, \ldots, X_K) \), where \( X_k \) is a binomial random variable for \( m_k \) trials with probability of success \( u_k \). Then, by Lemma 14,
\[
\sum_{\left| \frac{i}{m} - u \right| > m^{-\alpha}} P_{i,m}^K (u) = 1 - \Pr \left\{ \left| \frac{X_k}{m_k} - u_k \right| \leq m_k^{-\alpha}, \ k = 1, \ldots, K \right\}
\leq 2K \max_{k \in \{1, \ldots, K\}} \exp \left\{ -2m_k^{1-2\alpha} \right\}.
\]
Hence,
\[
II \leq 2K \max_{k \in \{1, \ldots, K\}} \exp \left\{ -2m_k^{1-2\alpha} \right\} \max_{i \in \mathcal{M}_m^K} |f \left( \frac{i}{m} \right)|,
\]
so that I plus the last display gives the result. □

By the previous result, we can show that the estimator converges to its expectation in \( L_1 \).
Lemma 16. Under Conditions 1 and 2,

\[ \int_{[0,1]^K} |(1 - \mathbb{E}) f_n(u)| \, du \overset{a.s.}{\to} 0 \]

if \( n^\varepsilon |p(i,m)(1-\xi) \to \infty \) for some \( \varepsilon \in [0, 1/2) \).

Proof. Using Lemma 15 with \( f(i/m) = |p(i,m)|^{(1-\xi)} (\mathbb{P}_n - \mathbb{E}) I_{i,m}(U) \),

\[
\int_{[0,1]^K} |(1 - \mathbb{E}) f_n(u)| \, du \\
= \int_{[0,1]^K} \left| \sum_{i \in M^K_{m}} |p(i,m)|^{(1-\xi)} (\mathbb{P}_n - \mathbb{E}) I_{i,m}(U) P^K_{i,m}(u) \right| \, du \\
\leq \int_{[0,1]^K} \left| \sum_{\frac{i}{m} - u \leq m^{-\varepsilon}} |p(i,m)|^{(1-\xi)} (\mathbb{P}_n - \mathbb{E}) I_{i,m}(U) P^K_{i,m}(u) \right| \, du \\
+ 2K \max_{k \in \{1, \ldots, K\}} \exp \left\{ -2m_k^{1-2\varepsilon} \right\} \max_{i \in M^K_{m}} |p(i,m)|^{(1-\xi)} (\mathbb{P}_n - \mathbb{E}) I_{i,m}(U) \\
= I + II.
\]

By direct calculation using the definition of beta function,

\[
\int_{[0,1]} P_{i(k),m(k)}(u_k) \, du_k = (m_k + 1)^{-1},
\]

so that

\[
\int_{[0,1]^K} P_{i,m}(u) \, du = |p(i,m)|.
\]

Hence, by convexity of norms and the last display,

\[
I \leq \sum_{\frac{i}{m} - u \leq m^{-\varepsilon}} |p(i,m)|^{(1-\xi)} (\mathbb{P}_n - \mathbb{E}) I_{i,m}(U) \int_{[0,1]^K} P^K_{i,m}(u) \, du \\
= \sum_{\frac{i}{m} - u \leq m^{-\varepsilon}} |(\mathbb{P}_n - \mathbb{E}) I_{i,m}(U)|.
\]

Define \( A \) to be the union of the sets \( p(i,m) \) such that \( |i - mu| \leq m^{1-\varepsilon} \). Therefore, by Scheffé’s lemma (e.g. [6, Theorem 5.4]), and Rio [14, Corollary 1] (giving a Hoeffding’s inequality for
dependent sequences), for any \( x > 0 \),

\[
\Pr \left( \sum_{|\frac{i}{m} - u| \leq m^{-\epsilon}} \left| (\mathbb{P}_n - \mathbb{E}) I_{i,m} (U) \right| > x \right)
\]

\[
= \Pr \left( 2 \sup_{A \in \mathcal{A}} |\mathbb{P}_n (U \in A) - \mathbb{P} (U \in A)| > x \right)
\]

[Scheffé’s Lemma]

\[
\leq 2 |p(i,m)|^{-(1-\alpha)} \sup_{A \in \mathcal{A}} \Pr \left( |\mathbb{P}_n (U \in A) - \mathbb{P} (U \in A)| > x / 2 \right)
\]

[because the power set of \( \mathcal{A} \) is \( 2 \sum_{i \in \mathcal{A}} |p(i,m)|^{-(1-\alpha)} \)]

\[
\leq 2 |p(i,m)|^{-(1-\alpha)} 2 \exp \left\{ -\kappa n \left( \frac{x}{2} \right)^2 \right\}
\]

[Rio [14, Corollary 1], with \( \kappa \) defined below]

\[
= 2 \exp \left\{ -n \kappa \left( \frac{x^2}{4} - \frac{|p(i,m)|^{-(1-\alpha)}}{\kappa n} \ln 2 \right) \right\},
\]

where

\[
\kappa := 2 \left( 1 + 2 \sum_{i=1}^{n-1} \phi_i \right)^{-2}
\]

(3)

and by Condition 1, \( \kappa \asymp n^{-(1-\alpha)} \). Hence, the above display is summable when \( |p(i,m)|^{-(1-\alpha)} n^{-\epsilon} = o(1) \) so that the Borell–Cantelli lemma implies \( 1 \overset{a.s.}\rightarrow 0 \). Clearly, \( II \overset{a.s.}\rightarrow 0 \) under Condition 2. \( \square \)

Similarly, we show convergence of the estimator to its expectation under the uniform norm.

Lemma 17. Under Conditions 1 and 2,

\[
\sup_{u \in [0,1]^K} \left| (1 - \mathbb{E}) f_n (u) \right| \overset{a.s.}\rightarrow 0.
\]

if \( n^{\epsilon} |p(i,m)|^{(2-\alpha)} \rightarrow \infty \) for some \( \alpha \in [0, 1/2) \).

Proof. Again, using Lemma 15 with \( f(i/m) = |p(i,m)|^{-1} (\mathbb{P}_n - \mathbb{E}) I_{i,m} (U) \),

\[
\sup_{u \in [0,1]^K} \left| (1 - \mathbb{E}) f_n (u) \right|
\]

\[
= \sup_{u \in [0,1]^K} \left| \sum_{i \in [m \mathbb{K}_m]} |p(i,m)|^{-1} (\mathbb{P}_n - \mathbb{E}) I_{i,m} (U) P_{i,m}^K (u) \right|
\]

\[
\leq \sup_{u \in [0,1]^K} \left| \sum_{|\frac{i}{m} - u| \leq m^{-\epsilon}} |p(i,m)|^{-1} (\mathbb{P}_n - \mathbb{E}) I_{i,m} (U) P_{i,m}^K (u) \right|
\]
and by the previous lemma it is enough to deal with \( I \) only. To avoid trivialities in the notation, assume \( m_k/2 \in \mathbb{N} \). Then,

\[
\begin{align*}
\Pi(k),m(k) (u_k) &\leq \Pi m/2, m(k) (1/2)
\end{align*}
\]

because the binomial mass function attains its maximum when the probabilities of success and failure are the same and when we consider \( m_k/2 \) successes over \( m_k \) trials (i.e. the most likely event). By Stirling’s approximation

\[
\sqrt{m} \Pi m/2, m(k) (1/2) \to \sqrt{2/\pi},
\]

so that

\[
P_i,m (u) du \leq c |p (i, m)|^{1/2}
\]

for some finite constant \( c \), implying

\[
I \leq c \sup_{u \in [0,1]^K} \sum_{m/2-u \leq m^{-2}} |p (i, m)|^{-1/2} (\mathbb{P}_n - \mathbb{E}) I_i,m (U).
\]

Define \( \mathcal{A}' \) to be the union of the sets \( (p (i, m))_i \in \mathbb{N}_m^K \) such that \( \sup_{u \in [0,1]^K} |i - mu| \leq m^{1-z} \) and note that \( \mathcal{A}' \) has cardinality \( |p (i, m)|^{-(1-z)} \) so that the power set of \( \mathcal{A}' \) is \( 2^{|p(i,m)|^{-(1-z)}} \). By the arguments in the proof of Lemma 16,

\[
\begin{align*}
\Pr \left( \sup_{u \in [0,1]^K} \sum_{m/2-u \leq m^{-2}} c |p (i, m)|^{-1/2} (\mathbb{P}_n - \mathbb{E}) I_i,m (U) > x \right) &\leq 2 |p(i,m)|^{-(1-z)} 2 \exp \left\{ -\kappa n \left[ x / \left( 2c |p (i, m)|^{-1/2} \right) \right]^2 \right\} \\
&= 2 \exp \left\{ -n |p (i, m)| \kappa \left( x^2 / 4c^2 - |p (i, m)|^{(2-z) / n} \ln 2 \right) \right\},
\end{align*}
\]

which is summable if \( |p (i, m)|^{-(2-z) n^{-\epsilon}} = o (1) \) because by Condition 1, \( \kappa \asymp n^{-(1-\epsilon)} \). Hence, \( I \overset{a.s.}{\to} 0 \) by the Borell–Cantelli Lemma.

We can now prove Theorem 8.

**Proof of Theorem 8.** By the triangle inequality,

\[
\int_{[0,1]^K} |f_n (u) - f (u)| du \leq \int_{[0,1]^K} |(1 - \mathbb{E}) f_n (u)| du + \int_{[0,1]^K} |\mathbb{E} f_n (u) - f (u)| du
\]

where \( I \) is the variation term, while \( II \) is the bias. By Lemma 16, \( I \overset{a.s.}{\to} 0 \). Finally, since \( C ([0, 1]^K) \) is densely embedded in \( L_1 ([0, 1]^K) \) (e.g. [6, Lemma 5.1]), it is sufficient to consider \( f (u) \in \)
C ([0, 1]^K) implying \( f (\ln^+ f)^{K-1} \in L_1 ([0, 1]^K) \). By the almost everywhere convergence of Kantorovich polynomials [19, Corollary]

\[
\|Ef_n(u) - f(u)\| \to 0, \text{ a.e. in } [0, 1]^K,
\]

so that by continuity of \( Ef_n(u) \), the dominated convergence theorem gives \( II \to 0 \). \( \square \)

**Proof of Theorem 10.** The proof is similar to that for Theorem 8, using Lemma 17 instead of Lemma 16. Note that the Corollary in Szili [19] gives the pointwise a.e. convergence of Kantorovich approximations of functions that satisfy the Marcinkiewicz–Zygmund condition of Remark 7. Hence, we can deduce the result. \( \square \)

**Proof of Corollary 11.** Lemma 17 gives the uniform convergence of the variation term, hence we only need to show the uniform convergence of the bias term. Ergoff’s Theorem (e.g. [12]) allows us to replace a.e. pointwise convergence of Kantorovich polynomials with a.u. convergence. This remark together with Lemma 17 gives the result. \( \square \)

**Acknowledgment**

I am very grateful for the comments of the reviewers that led to considerable improvement in the content and clarity of the presentation. Supported by the ESRC Research Award 000-23-0400.

**References**