HOEFFDING'S INEQUALITY, MINIMUM DISTANCE ESTIMATION, KOLMOGOROV ENTROPY, RATES OF CONVERGENCE AND MATCHING

1 Probability Inequalities

• **Markov Inequality:** If X is a positive random variable (r.v.), $EX < \infty, \epsilon > 0$,

$$P(X > \epsilon) \leq \frac{EX}{\epsilon}.$$

Proof: (for continuous r.v.'s) Let f_X be the density of X.

$$EX = \int_0^\infty x f_X(x)dx \geq \int_\epsilon^\infty \frac{x}{\epsilon} \cdot \epsilon f_X(x)dx \geq \epsilon \int_\epsilon^\infty 1 \cdot f_X(x)dx = \epsilon P(X \geq \epsilon).$$

• **Chebychev Inequality:** Let X be r.v. with $EX^2 < \infty$, then

$$P[|X - EX| > \epsilon] \leq \frac{Var(X)}{\epsilon^2}.$$

• **Cauchy-Schwartz inequality:** If U and V are r.vs, $EU^2 < \infty, EV^2 < \infty$,

a)

$$|EUV| \leq [E(U^2)]^{1/2}[E(V^2)]^{1/2}, \tag{1}$$

b) for $U = |X|, V = |Y|$,

$$E|X| \cdot |Y| \leq [E(X^2)]^{1/2}[E(Y^2)]^{1/2}.$$

Proof: $0 \leq E(U - aV)^2 = E(U^2) + a^2E(V^2) - 2aEUV$ which is minimized at $a = \frac{EUV}{E(V^2)}$

$$\to 0 \leq E(U^2) + \frac{(EUV)^2}{E(V^2)} - 2\frac{(EUV)^2}{E(V^2)} = E(U)^2 - \frac{(EUV)^2}{E(V^2)} \to |EUV| \leq [E(U^2)]^{1/2}[E(V^2)]^{1/2}.$$

Definition 1.1 *Let $f(x)$ be a real valued function defined on the interval $I = [a, b]$. f is convex if for every $x_1, x_2 \in [a, b]$ and $0 \leq \lambda \leq 1$,*

$$f[\lambda x_1 + (1 - \lambda)x_2] \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{2}$$

Proposition 1.1 *(Jensen's Inequality) Let X be a r.v. with domain the real line and with expected value EX. Let f be a convex function with domain the range of the values of X. Then,*

$$f(EX) \leq Ef(X). \tag{3}$$

Note: If you want to see the Proof for Jensen's inequality, please let me know.

2 Hoeffding's Inequality

Recall from Probability Chebychev's inequality: Let $X_1, \ldots, X_n$ be $i.i.d.$ r.vs with mean μ and variance σ^2, $\bar{X}_n$ denotes the average of the X's. Then, for every $\epsilon > 0$,

$$P[|\bar{X}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Observe that the upper probability bound converges to zero as $n \uparrow \infty$ at rate $\frac{1}{n}$.

We would prefer an upper bound that tends in probability to zero at faster rate.

A sharper inequality is Hoeffding's inequality, with the upper bound decreasing exponentially to zero. A lemma will be used to prove it.

Lemma 2.1 *Let X be a r.v.* **with mean 0,** $a \leq X \leq b$, $a < 0 < b$. *Then, for any* $t > 0$,

$$M_X(t) = Ee^{tX} \leq e^{t^2(b-a)^2/8}. \tag{4}$$

Proof: Since $t > 0$, the function e^{tx} is convex. Consider $x \in [a, b]$, then

$$x = \lambda b + (1 - \lambda)a$$

with

$$\lambda = \frac{x-a}{b-a}, 1 - \lambda = \frac{b-x}{b-a}.$$

Then, by convexity of $g(x) = e^{tx}$ when $t > 0$,

$$e^{tx} = e^{\lambda tb + (1-\lambda)ta} \leq \lambda e^{tb} + (1-\lambda)e^{ta} = \frac{x-a}{b-a}e^{tb} + \frac{b-x}{b-a}e^{ta},$$

and since by assumption $EX = 0$,

$$\implies Ee^{tX} = \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \text{ and taking ln in both sides}$$

$$\ln M_X(t) = \ln(\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}) = ta + \ln(\frac{b}{b-a} - \frac{a}{b-a}e^{t(b-a)}).$$

Let $u = t(b-a) \to ta = \frac{a}{b-a}u$, then

$$\ln M_X(t) = \frac{a}{b-a}u + \ln(\frac{b}{b-a} - \frac{a}{b-a}e^{u}) = f(u),$$

observe that $f(0) = 0$,

$$f'(u) = \frac{a}{b-a} + \frac{1}{\frac{b}{b-a} - \frac{a}{b-a}e^u} \cdot \frac{-a}{b-a}e^u = \frac{a}{b-a} - \frac{ae^u}{b-ae^u} \to f'(0) = 0,$$

$$f''(u) = -\frac{ae^u(b-ae^u) + a^2e^{2u}}{(b-ae^u)^2} = -\frac{abe^u}{(b-ae^u)^2}.$$

To show:

$$f''(u) \le \frac{1}{4} \iff -4abe^u \le b^2 - 2abe^u + a^2e^{2u} \iff 0 \le b^2 + 2abe^u + a^2e^{2u} = (b+ae^u)^2,$$

which holds. It the follows that,

$$\ln M_X(t) = f(u) = f(0) + f'(0)u + f''(u_0)\frac{u^2}{2} \le \frac{u^2}{8} = \frac{t^2(b-a)^2}{8} \to M_X(t) \le e^{\frac{t^2(b-a)^2}{8}}.$$

(1963)

Proposition 2.1 *(**Hoeffding's inequality**-One of several versions) Let $X_1, \ldots, X_n$ be independent, <u>centered</u> random variables, $EX_i = 0, a_i \le X_i \le b_i, a_i < 0 < b_i, i = 1, \ldots, n, S_n = \sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,*

$$P(S_n > \epsilon) \le e^{-2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{5}$$

Using (5) for $-X_1, \ldots, -X_n$ it follows that

$$P(-S_n > \epsilon) = P(S_n < -\epsilon) \le e^{-2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2} \tag{6}$$

and

$$P(|S_n| > \epsilon) \le 2 \cdot e^{-2\epsilon^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{7}$$

Proof: For any $t > 0$, use Markov Inequality for e^{tS_n},

$$P(S_n > \epsilon) = P(e^{tS_n} > e^{t\epsilon}) \leq e^{-\epsilon t} E e^{tS_n} = e^{-\epsilon t} \Pi_{i=1}^n M_{X_i}(t)$$

From Lemma 2.1, $M_{X_i}(t) \leq e^{t^2(b_i - a_i)^2/8}, i = 1, \ldots, n$, and the quadratic $t^2 \frac{\sum_{i=1}^n (b-a)^2}{8} - \epsilon t$ in the probability bound is minimized at $t = 4\epsilon / \sum_{i=1}^n (b_i - a_i)^2$, thus

$$P(S_n > \epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}.$$

Remark 2.1 **a)** For a sample of i.i.d. $Bernoulli(p)$ random variables, $X_1, \ldots, X_n$ compare the Chebychev and Hoeffding bounds for $P(|\bar{X}_n - p| > \epsilon)$ for your choice of $\epsilon, n; \bar{X}_n = n^{-1}(X_1 + \ldots + X_n),\ X_i = 1,$ with probability p and $X_i = 0$ otherwise, $i = 1, \ldots, n$.
b) Let A be a measurable set in R, i.e. for which we can calculate the probability $P(A)$ and $X_1, \ldots, X_n$ are *i.i.d.* P. Let $I_A(X_i) = 1$, when X_i take value in A and otherwise $I_A(X_i) = 0$. Then, $I_A(X_1), \ldots, I_A(X_n)$ are *i.i.d.* Bernoulli random variables with probability $P(A)$ of taking the value 1. Obtain Hoeffding's bound for $P[|\frac{1}{n}\sum_{i=1}^n I_A(X_i) - P(A)| > k_n]$.

3 Distances and deviations between probability measures/densities

Let P, Q measures on a space $\mathcal{X}$ with a σ-field $\mathcal{A}$. Assume the measures have densities p and q respectively, with respect to dominating measure $\mu : \frac{dP}{d\mu} = p, \frac{dQ}{d\mu} = q$. You can think of μ as Lebesgue measure, i.e. $\mu(dx) = dx$.

- L_1**-distance (or Total Variation distance) between** P, Q :

$$||P - Q||_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|. \tag{8}$$

• Show that $\int_{\mathcal{X}} |p(x) - q(x)| \mu(dx) = ||P - Q||_1$ denoted also $||p - q||_1$.
• Draw the graph of two normal densities, e.g. $\mathcal{N}(2, 1), \mathcal{N}(4, 1)$ on the real line and see graphically what their L_1-distance is; use $||p - q||_1$.

• **Hellinger distance** $h(P,Q)$ **between** P, Q :

$$h^2(P,Q) = h^2(p,q) = \int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 \mu(dx) = 2[1 - \int_{\mathcal{X}} \sqrt{p(x)}\sqrt{q(x)}\mu(dx)] = 2[1 - \rho(p,q)], \tag{9}$$

$$\rho(p,q) = \int_{\mathcal{X}} \sqrt{p(x)}\sqrt{q(x)}\mu(dx).$$

$\rho(p,q)$ was called by Le Cam *the affinity of* P, Q and it holds from (9) and via Cauchy-Schwartz inequality

$$0 \le \rho(p,q) = 1 - \frac{1}{2}h^2(P,Q) \le 1. \tag{10}$$

(Indeed: $\int_{\mathcal{X}} \sqrt{p(x)}\sqrt{q(x)}\mu(dx) = \int_{\mathcal{X}} q(x)\sqrt{\frac{p(x)}{q(x)}}\mu(dx) \le [\int_{\mathcal{X}} q(x)\frac{p(x)}{q(x)}\mu(dx)]^{1/2} = 1.$)
• It follows that $0 \le h(P,Q) \le \sqrt{2}$.

• For $\mathcal{N}(\theta_1, 1), \mathcal{N}(\theta_2, 1),\ \theta_1 < \theta_2$, calculate their Hellinger distance and their L_1-distance.

Remark 3.1 *Express the L_1-distance like the last equality in (9). What will be the corresponding affinity in L_1-distance?*

Exercise:

$$a) \qquad ||P - Q||_1 = 2[P(x : p(x) > q(x)) - Q(x : p(x) > q(x))]. \tag{11}$$

In the proof you may use the integral version in (8).

$$b) \qquad ||P-Q||_1 = \int_{\mathcal{X}} |p(x) - q(x)|dx = 2[1 - \int_{\mathcal{X}} p(x) \wedge q(x)dx] = 2[\int_{\mathcal{X}} p(x) \vee q(x)dx - 1] \tag{12}$$

Proof of *b)*: Use the notation $p > q$ for the set $\{x : p(x) > q(x)\}$.

$$2 = \int_{p>q} p(x)dx + \int_{p<q} p(x)dx + \int_{q>p} q(x)dx + \int_{q<p} q(x)dx \tag{13}$$

and observing for the second and the fourth integrals in (13)

$$\int_{p<q} p(x)dx + \int_{q<p} q(x)dx = \int_{\mathcal{X}} p(x) \wedge q(x)dx$$

it follows for the sum of first and third integrals

$$\rightarrow \int_{p>q} p(x)dx + \int_{q>p} q(x)dx = 2 - \int_{\mathcal{X}} p(x) \wedge q(x)dx$$

$$\to P(p>q)-Q(p>q)=1-\int_{\mathcal{X}} p(x)\wedge q(x)dx \to ||P-Q||_1 = 2[1-\int_{\mathcal{X}} p(x)\wedge q(x)dx].$$

Note that the first equality above is easily seen drawing the graphs of densities p, q.

Also, (13) can be rewritten

$$2=\int_{\mathcal{X}} p(x)\vee q(x)dx+\int_{\mathcal{X}} p(x)\wedge q(x)dx \to 2=\int p(x)\vee q(x)dx+(1-\frac{1}{2}||P-Q||_1)$$

$$\to \int p(x)\vee q(x)dx = 1+\frac{1}{2}||P-Q||_1.$$

- Kullback-Leibler non-distance (WHY?) between P, Q :

$$d_{KL}(P,Q)=d_{KL}(p,q)=\int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\mu(dx)$$

Observe: $d_{KL}(P,Q)=\int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\mu(dx)=-\int_{\mathcal{X}} p(x)\log\frac{q(x)}{p(x)}\mu(dx)=E_P[\,\mathbf{-\log}\frac{q(X)}{p(X)}]$
$\geq\ -\mathbf{log}E_P\frac{q(X)}{p(X)}=-\log(1)=0.$

- L^r**-distances for densities,** $r\geq 1$: $||p-q||_r=[\int_{\mathcal{X}} p(x)-q(x)|^r dx]^{1/r}$.

- **Kolmogorov distance,** d_K, **between c.d.fs** For c.d.fs F, G in R^d,

$$d_k(F,G)=\sup_{x\in R^d}|F(x)-G(x)|\,;$$

it is also called Kolmogorov-Smirnov distance.

Inequalities for distances

- $h^2(P,Q)\leq ||P-Q||_1\leq h(P,Q)\sqrt{4-h^2(P,Q)}\leq 2h(P,Q)$

Proof: $\int_{\mathcal{X}}|p(x)-q(x)|\mu(dx)=\int_{\mathcal{X}}|\sqrt{p(x)}-\sqrt{q(x)}|\cdot|\sqrt{p(x)}+\sqrt{q(x)}|\mu(dx)\geq h^2(P,Q)$,

$$\int_{\mathcal{X}}|p(x)-q(x)|\mu(dx)=\int_{\mathcal{X}} p(x)|1-\frac{q(x)}{p(x)}|\mu(dx)=\int_{\mathcal{X}} p(x)|1-\frac{\sqrt{q(x)}}{\sqrt{p(x)}}|\cdot|1+\frac{\sqrt{q(x)}}{\sqrt{p(x)}}|\mu(dx)$$

$$\leq[\int_{\mathcal{X}} p(x)(1-\frac{\sqrt{q(x)}}{\sqrt{p(x)}})^2\mu(dx)]^{1/2}\cdot[\int_{\mathcal{X}} p(x)(1+\frac{\sqrt{q(x)}}{\sqrt{p(x)}})^2\mu(dx)]^{1/2}=h(P,Q)[\int_{\mathcal{X}}|\sqrt{p(x)}+\sqrt{q(x)}|^2\mu(dx)]^{1/2}$$

$$=h(P,Q)\cdot(2+2\rho(p,q))^{1/2}=h(P,Q)\cdot(4-h^2(P,Q))^{1/2}\leq 2h(P,Q)$$

Lemma 3.1 *Let* $0 \le u_i \le 1, i = 1, \dots, n$. *Show that*

$$1 - \Pi_{i=1}^{n}(1 - u_i) \le \sum_{i=1}^{n} u_i. \tag{14}$$

Proof: By induction: for $i = 1$, indeed it holds $1 - (1 - u_1) \le u_1$. We do it also for $i = 2$ to get a feeling for the general case:

$$1 - (1 - u_1)(1 - u_2) \le u_1 + u_2 \iff 1 - [1 - u_2 - u_1 + u_1 u_2] \le u_1 + u_2 \iff -u_1 u_2 \le 0.$$

Assume that for $n = k$ (14) holds,

$$1 - \Pi_{i=1}^{k}(1 - u_i) \le \sum_{i=1}^{k} u_i.$$

To show it also holds for $n = k + 1$,

$$1 - \Pi_{i=1}^{k+1}(1-u_i) = 1 - \Pi_{i=1}^{k}(1-u_i)(1-u_{k+1}) = 1 - \Pi_{i=1}^{k}(1-u_i) + u_{k+1}\Pi_{i=1}^{k}(1-u_i) \le \sum_{i=1}^{k} u_i + u_{k+1}$$

Proposition 3.1 *a) If* $X_i,\ i = 1, \dots, n$ *are independent r.v. with probabilities either* $P_i, i = 1, \dots, n$ *or* $Q_i, i = 1, \dots, n$ *with densities* $p_i, q_i, i = 1, \dots, n$, *then* $(X_1, \dots, X_n)$ *will have as joint probability the probabiliy defined by* $P_1 x \dots x P_n$ *(notation, well defined by products) or* $Q_1 x \dots x Q_n$ *and densities either* $p_1 \cdot p_2 \dots p_n$ *or* $q_1 \cdot q_2 \dots q_n$. *Then,*

$$h^2(P_1 x P_2 x \dots x P_n, Q_1 x Q_2 x \dots x Q_n) = 2[1 - \Pi_{i=1}^{n} \rho(p_i, q_i)] \le \sum_{i=1}^{n} h^2(P_i, Q_i). \tag{15}$$

b) When $P_1 = \dots = P_n = P,\ Q_1 = \dots = Q_n = Q$ *then for the corresponding* n*-product probabilities* $P^{(n)}$ *and* $Q^{(n)}$ *it holds*

$$h^2(P^{(n)}, Q^{(n)}) = 2[1 - \rho^n(p, q)] \le n \cdot h^2(P, Q). \tag{16}$$

Observe: From the equality in the middle of (16), the distance $h^2(P^{(n)}, Q^{(n)})$ increases to 2 with n, i.e. the probabilities $(P^{(n)}, Q^{(n)})$ separate and are easier to distinguish in estimation and testing!

Proof:

$$h^2(P_1 x P_2 x \dots P_n, Q_1 x Q_2 x \dots Q_n) = 2[1 - \Pi_{i=1}^{n} \rho(p_i, q_i)] = 2[1 - \Pi_{i=1}^{n}(1 - \frac{1}{2} h^2(p_i, q_i))]$$

Suffices to show that

$$1 - \Pi_{i=1}^{n}(1 - \frac{1}{2}h^2(p_i, q_i))] \leq \frac{1}{2}\sum_{i=1}^{n} h^2(P_i, Q_i).$$

Recall that

$$0 \leq u_i = \frac{1}{2}h^2(p_i, q_i) \leq 1, i = 1, \ldots, n$$

thus the result follows from (14).

• When $P_1 = \ldots = P_n = P$, $Q_1 = \ldots = Q_n = Q$ then for the corresponding n-product probabilities $P^{(n)}$ and $Q^{(n)}$ it holds

$$h^2(P^{(n)}, Q^{(n)}) = 2[1 - \rho^n(p, q)] \leq n \cdot h^2(P, Q).$$

• $||P - Q||_1^2 \leq 2 \cdot d_{KL}(p, q)$

Proof: A set that determines the L_1-distance between P and Q is: $A = \{x : p(x) > q(x)\}$. A will be used to prove the inequality by splitting the integral in d_{KL} in two parts, over A and its complement A^c.

Note that $I_A(x) = 1$, if $x \in A$, and 0 otherwise, and that $q(x)I_A(x)/\int_A q(x)dx$ is a density over the whole space where p, q are defined.

The convex function $f(y) = y \log y, y > 0$, is used and Jensen's inequality after creating f for $y = \frac{p(x)}{q(x)} > 0$. For convexity note: $f'(y) = \log y + 1$, $f''(y) = 1/y > 0$ when $y > 0$.

$$\int_A p(x) \log \frac{p(x)}{q(x)}dx = \int \frac{I_A(x)q(x)}{\int_A q(x)dx} \cdot \{\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)}\}dx \cdot \int_A q(x)dx$$

$$\geq f(\int \frac{I_A(x)q(x)}{\int_A q(x)dx}\frac{p(x)}{q(x)}dx) \cdot \int_A q(x)dx = \int \frac{I_A(x)p(x)}{\int_A q(x)dx}dx \cdot \log[\int \frac{I_A(x)p(x)}{\int_A q(x)dx}dx] \cdot \int_A q(x)dx$$

$$= P(A) \log \frac{P(A)}{Q(A)}.$$

Similarly, since the nature of A was not used to obtain the inequality, it also holds

$$\int_{A^x} p(x) \log \frac{p(x)}{q(x)}dx \geq P(A^c) \log \frac{P(A^c)}{Q(A^c)} = [1 - P(A)] \log \frac{1 - P(A)}{1 - Q(A)}.$$

Therefore,

$$d_{KL}(p, q) \geq P(A) \log \frac{P(A)}{Q(A)} + [1 - P(A)] \log \frac{1 - P(A)}{1 - Q(A)}. \quad (17)$$

Let $\tilde{p} = P(A), \tilde{q} = Q(A)$. Recall $A = \{x : p(x) > q(x)\}$ which implies $P(A) > Q(A)$. Then, (17) can be written denoting the lower bound $H(\tilde{p}, \tilde{q})$ as

$$d_{KL}(P, Q) \geq H(\tilde{p}, \tilde{q}) = \tilde{p} \log \frac{\tilde{p}}{\tilde{q}} + (1 - \tilde{p}) \log \frac{1 - \tilde{p}}{1 - \tilde{q}}.$$

Let $\tilde{p} = \tilde{q} + r$. Then,

$$r = P(A) - Q(A) = \frac{1}{2}||P - Q||_1,$$

$$H(\tilde{p}, \tilde{q}) = H(\tilde{q} + r, \tilde{q}) = (\tilde{q} + r) \log(1 + \frac{r}{\tilde{q}}) + (1 - \tilde{q} - r) \log(1 - \frac{r}{1 - \tilde{q}}).$$

We now bound $H(\tilde{q} + r, \tilde{q})$ using Taylor expansion. For H' the first derivative of H with respect to r we get:

$$H'(\tilde{q} + r, \tilde{q}) = \log(1 + \frac{r}{\tilde{q}}) + (\tilde{q} + r) \frac{\tilde{q}}{\tilde{q} + r} \cdot \frac{1}{\tilde{q}} - \log(1 - \frac{r}{1 - \tilde{q}}) + (1 - \tilde{q} - r) \frac{1 - \tilde{q}}{1 - \tilde{q} - r} \cdot (\frac{-1}{1 - \tilde{q}})$$

$$= \log(1 + \frac{r}{\tilde{q}}) - \log(1 - \frac{r}{1 - \tilde{q}}),$$

$$H''(\tilde{q} + r, \tilde{q}) = \frac{\tilde{q}}{\tilde{q} + r} \cdot \frac{1}{\tilde{q}} - \frac{1 - \tilde{q}}{1 - \tilde{q} - r} \cdot (\frac{-1}{1 - \tilde{q}}) = \frac{1}{\tilde{q} + r} + \frac{1}{1 - \tilde{q} - r} = \frac{1}{(\tilde{q} + r)(1 - \tilde{q} - r)} \geq 4,$$

$\forall \, r : 0 < r < 1 - \tilde{q}$.

Observe that $H(\tilde{q}, \tilde{q}) = H'(\tilde{q}, \tilde{q}) = 0$ then from a Taylor expansion with a remainder term

$$d_{KL}(P, Q) \geq H(\tilde{q} + r, \tilde{q}) \geq 4 \frac{r^2}{2} = 2r^2 = \frac{1}{2}||P - Q||_1^2.$$

Exercise: Show that $||P - Q||_1 \leq 2\sqrt{1 - \exp\{-d_{KL}(P, Q)\}}$. (Hint: $\log \frac{q(x)}{p(x)} = \log(\frac{q(x)}{p(x)} \wedge 1) + \log(\frac{q(x)}{p(x)} \vee 1)$.)

Proof of Exercise: Sometimes we write $\log$ but we mean $\ln$.

$$-d_{KL}(P, Q) = -\int p(x) \log \frac{p(x)}{q(x)} dx = \int p(x) [\log(\frac{q(x)}{p(x)} \wedge 1) + \log(\frac{q(x)}{p(x)} \vee 1)]$$

$$\leq \log[\int q(x) \wedge p(x) dx] + \log[\int q(x) \vee p(x) dx]$$

$$\rightarrow \exp\{-d_{KL}(P, Q)\} \leq \int q(x) \wedge p(x) dx \int q(x) \vee p(x) dx = [1 - \frac{1}{2}||P - Q||_1] \cdot [1 + \frac{1}{2}||P - Q||_1]$$

$$\rightarrow \exp\{-d_{KL}(P, Q)\} \leq 1 - \frac{1}{4}||P - Q||_1^2 \rightarrow ||P - Q||_1^2 \leq 4[1 - \exp\{-d_{KL}(P, Q)\}].$$

4 Characterizing the dimension of a space

Examples from compact subsets in R^d.

$N_1(a) = \#$ of intervals of length a needed to cover$(0,1) \sim \frac{1}{a}$

$N_2(a) = \#$ of rectangles, side length a needed to cover$(0,1)^2 \sim \frac{1}{a^2}$

$N_d(a) = \#$ of rectangles, side length a needed to cover$(0,1)^d \sim \frac{1}{a^d}$

Observe: $\log N_d(a)/\log(\frac{1}{a}) \sim d$, the dimension of the space where $[0,1]^d$ lives.

Definition 4.1 *Let $(\mathcal{F}, \rho)$ be a metric space. For $a > 0$, let*

$$N(a) = \textit{minimum } \# \textit{ of } \rho\textit{-balls of radius } a \textit{ needed to cover } \mathcal{F}.$$

Then, $\log_2 N(a)$ *is Kolmogorov entropy of the space* $(\mathcal{F}, \rho)$.

$N(a)$ is useful in determining the dimension of a space, in particular of a space of functions metrized with a distance.

Examples

Notation: If $x = (x_1, \ldots, x_d) \in R^d,\ a \in R$ and $s = (s_1, \ldots, s_d)$ is a d-tuple of non-negative integers,

$$x^s = (x_1^{s_1}, \ldots, x_d^{s_d}),\ xs = x_1 s_1 + \ldots + x_d s_d,\ ax = (ax_1, \ldots, ax_d),\ [s] = s_1 + \ldots + s_d;$$

for $y \in R^d$,

$$|x - y| = \max\{|x_i - y_i|, i = 1, \ldots, d\}.$$

For a real valued function g defined in R^d let $g^{(s)}(x_0)$ denote the $[s]$-th order mixed partial derivative of g at x_0, i.e.

$$g^{(s)}(x_0) = \frac{\partial^{[s]} g(x_0)}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}.$$

a) q-smooth functions defined on a compact in R^d

Let $\mathcal{X} = [0,1]^d$, the uniformly bounded functions in sup-norm $\mathcal{F} = \{f : [0,1]^d \longrightarrow R^+\}$, such that each f has p-derivatives and the p-th derivative satisfies a Lipschitz condition with parameters $0 < \alpha < 1, L > 0$,

$$|f^{(p)}(x) - f^{(p)}(y)| \leq L \cdot |x-y|^{\alpha}, q = p + \alpha.$$

Note: In the literature you may see exponents $\alpha_i, i = 1, \ldots, d$, for each of the components of $|x-y|$. Then, $\min\{\alpha_i; i = 1, \ldots, d\}, \max\{\alpha_i; i = 1, \ldots, d\}$ play different roles in estimation. In the sequel we use the "isotropic" case, with all α_i's equal to α.

Kolmogorov and Tikhomirov (1959) have shown that $\mathcal{F}$ metrized with the sup-norm,

$$||f - g||_{\infty} = sup_x |f(x) - g(x)|$$

is totally bounded and that for every $a > 0$ for the smallest number $N_{\infty}(a)$ of $|| \cdot ||_{\infty}$-balls of radius a needed to cover $\mathcal{F}$ it holds

$$C_1 \cdot 2^{(\frac{1}{a})^{d/q}} \leq N_{\infty}(a) \leq C_2 \cdot 2^{(\frac{1}{a})^{d/q}}, \ 0 < C_1 < C_2. \tag{18}$$

Clements (1966) showed that when $\mathcal{F}$ is metrized by the L_1-distance then inequalities similar to (18) with the same bounds in terms of a modulo the constants.

b) Functions with uniformly bounded modulus of continuity

Let $\mathcal{X} = [0,1], \ \mathcal{F} = \{f : [0,1] \to R^+ : \omega_f(\epsilon) = \sup |f(x+h) - f(x)|; x \in (0,1), |h| < \epsilon\} \leq \omega(\epsilon)\}$. By Lorentz (1966), $\mathcal{F}$ metrized with $|| \cdot ||_{\infty}$ is totally bounded,

$$N_{\infty}(a) \leq \frac{K}{\delta(\gamma \cdot a)},$$

for K, γ fixed constant s and $\delta = \delta(a)$ any root of the equation $\omega(\delta) = a$.

5 Statistical Experiments-The estimation problem

Definition 5.1 *A Statistical Experiment, $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, consists of sample space $\mathcal{X}$ with σ-field $\mathcal{A}$, the parameter space $\mathbf{\Theta}$ with distance d_{Θ} and probability measures $\mathcal{P} = \{P_{\theta^*}; \theta^* \in \mathbf{\Theta}\}$; see* e.g. *Le Cam (1986), Le Cam and Yang(2000).*

The problem: $\mathbf{X} \in \mathcal{X}$ is observed from P_θ, θ unknown and the aim is to estimate θ and study properties of the estimate, *e.g.* its rate of convergence to θ with respect to d_Θ.

Note: $\mathbf{X}$ could be a sample $X_1, \ldots, X_n$ with each X_i from P_θ.$\mathbf{X}$ could be seen as vector in R^n and then in the Statistical Experiment $\mathcal{X}$ is indeed $\mathcal{X}^n = R^n$, and the corresponding $\mathcal{P}$ includes product probabilities, or joint densities each indexed by $\theta \in \Theta$.

Instead of $\mathcal{P}$ one can use the corresponding *c.d.fs* $\mathcal{F}_{\mathbf{\Theta}} = \{F_{\theta^*}, \theta^* \in \mathbf{\Theta}\}$ with generic distance $\tilde{d}$ used also for functionals $T(F_{\theta^*}), \theta^* \in \mathbf{\Theta}$, and assume identifiability $i.e.$ $F_{\theta_1} = F_{\theta_2}$ implies $\theta_1 = \theta_2$. We will use $\mathcal{F}_{\mathbf{\Theta}}$ to denote c.d.fs or the corresponding densities.

6 Upper Rates of Convergence in Probability

Our goal is to define the upper rate of convergence of an estimate to a parameter in Probability. If $X_1, \ldots, X_n$ is *i.i.d.* sample with unknown mean μ and finite variance $\sigma^2 = 1$, we want $k_n \downarrow 0$ such that

$$\lim_{n\to\infty} P[|\bar{X}_n - \mu| > k_n] = 0. \tag{19}$$

We hear $\bar{X}_n$ (or the MLE) converge at rate $n^{-1/2}$. Can we use $k_n = n^{-1/2}$ in (19)?

From CLT, $n^{1/2}(\bar{X}_n - \mu)$ has asymptotic distribution the Normal. Is $k_n = \frac{C}{\sqrt{n}}$, $C > 0$? Observe that

$$P[|\bar{X}_n - \mu| > \frac{C}{\sqrt{n}}] \approx P[Z > C] \neq 0, \tag{20}$$

and for this probability to converge to 0 we need $C = C_n \uparrow \infty, C_n = o(n^{-1/2})$.

We complete (19) in view of (20): for every $\epsilon > 0$ there are $C_\epsilon, n(\epsilon)$:

$$P[|\bar{X}_n - \mu| > C_\epsilon k_n] < \epsilon \tag{21}$$

for every $n \geq n(\epsilon)$. For $n < n(\epsilon)$ there will be another constant that depends on ϵ that will make (21) hold for $1 \leq n < n(\epsilon)$. Thus, there is C_ϵ for which (21) holds for $n \geq 1$. We would prefer that the rate is the same for all μ, *i.e.* uniform, so we will add in front of the probability in (21) the

$\sup_\mu$,

$$\sup_\mu P[|\bar{X}_n - \mu| > C_\epsilon k_n] < \epsilon, n \geq 1. \tag{22}$$

Note also that (22) can be written

$$\lim_{C\to\infty} \sup_\mu P[|\bar{X}_n - \mu| > C k_n] = 0, n \geq 1. \tag{23}$$

Definition 6.1 Let $X_1, \ldots, X_n$ be a sample of d-dimensional vectors from unknown probability P_θ, element of a known family of probabilities $\mathcal{P}, \theta$ element of a metric space (Θ, ρ). A sequence $\hat{\theta}_n$ of estimates of θ is uniformly consistent estimate for θ in probability, with upper rate of convergence δ_n with respect to ρ if for every $\epsilon > 0$ there is $C(\epsilon)(> 1\ w.l.o.g.)$ such that

$$\sup_{\theta\in\Theta} P_\theta^{(n)}[\rho(\hat{\theta}_n, \theta) > C(\epsilon)\delta_n] \leq \epsilon,\ \forall\ n \geq 1. \tag{24}$$

(24) is briefly denoted "$\hat{\theta}_n$ has upper ρ-error rate, δ_n, in probability." It is expected δ_n converges to zero. $P_\theta^{(n)}$ in (24) denotes the joint probability of the sample.

7 Wolfowitz's Minimum Distance Estimates

Wolfowitz introduced Minimum Distance Estimation/Estimates (MDE) in a series of papers in the 50's (e.g. 1957) using as tools the empirical cumulative distribution of the sample, Kolmogorov distance d_K and Dvoretzky-Kiefer-Wolfowitz inequality (1956) for *iid* r.vs that was extended also for *i.i.d.* random vectors in R^d.

Kolmogorov distance, d_K, **between c.d.fs:** For c.d.fs F, G in R^d,

$$d_k(F, G) = \sup_{x\in R^d} |F(x) - G(x)|\,;$$

it is also called Kolmogorov-Smirnov distance.

Definition 7.1 *For any n-size sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ of random vectors in $R^d, n\hat{F}_n(y)$ denotes the number of Y_i's with all their components smaller or equal to the corresponding components of y. $\hat{F}_n$ is the empirical c.d.f. of $\mathbf{Y}$, denoted also $\hat{F}_{\mathbf{Y}}$.*

Theorem 7.1 *(Dvoretzky, Kiefer and Wolfowitz, 1956, and the tight upper bound by Massart, 1990) Let $\hat{F}_n$ denote the empirical c.d.f of the size n sample $\mathbf{Y}$ of i.i.d. random variables obtained from cumulative distribution F. Then, for any $\epsilon > 0$,*

$$P[d_K(\hat{F}_n, F) > \epsilon] \leq U_{DKWM} = 2e^{-2n\epsilon^2} \tag{25}$$

Inequality (80) implies that $\hat{F}_n$ converges in probability to F with respect to Kolmogorov distance. For example check that with $\epsilon_n = \frac{\sqrt{\ln n}}{\sqrt{n}}$ an upper rate of convergencein probability is obtained.

Generalizations of (80) in R^d have been obtained, at least, by Kiefer and Wolfowitz (1958), Kiefer (1961) and Devroye (1977); $d > 1$. The differences in upper bound U in (80) are in the multiplicative constant, in the exponent of the exponential and on the sample size for which the exponential bound holds which may also depend on ϵ. The constants used are not determined except for Devroye (1977).

i) In Kiefer and Wolfowitz (1958), the upper bound in (80) $U_{KW} = C_1(d)e^{-C_2(d)n\epsilon^2}$.

ii) In Kiefer (1961), the upper bound in (80) $U_K = C_3(b, d)e^{-(2-b)n\epsilon^2}$, for every $b \in (0, 2)$.

iii) In Devroye (1977), with the upper bound in (80) $U_{De} = 2e^2(2n)^d e^{-2n\epsilon^2}$ valid for $n\epsilon^2 \geq d^2$.

There are also exponential bounds under weak dependence and for non-exponential bounds for linear time series. I can provide the reference if you need it.

Definition 7.2 *For sample $\mathbf{X}$ having unknown c.d.f $F_\theta \in \mathcal{F}_{\mathbf{\Theta}}$, the Minimum Distance Estimate, $\tilde{\theta}_{MDE}$, of θ is defined such that:*

$$d_K(F_{\tilde{\theta}_{MDE}}, \hat{F}_n) \leq \inf_{\theta^* \in \Theta} d_K(F_{\theta^*}, \hat{F}_n) + \gamma_n, \tag{26}$$

with the user's choice of $\gamma_n \downarrow 0$ as $n \uparrow \infty$, when $\gamma_n = 0$ cannot be used.

The infimum in (42) may not be achievable and by including $\gamma_n > 0$, $\tilde{\theta}_{MDE}$ is element of

$$\tilde{\Theta}_n = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_{m_n}, \ldots\} \tag{27}$$

satisfying (42). Thus, $d_K(\hat{F}_{\tilde{\theta}_{MDE}}, \hat{F}_n)$ is kept small for $\tilde{\theta}_{MDE} \in \tilde{\Theta}_n$.

Key inequality for proving consistency and the uniform convergence rate $\frac{k_n}{\sqrt{n}}$ of $F_{\tilde{\theta}_{MDE}}$ to F_θ is:

$$d_K(F_{\tilde{\theta}_{MDE}}, F_\theta) \leq d_K(F_{\tilde{\theta}_{MDE}}, \hat{F}_n) + d_K(\hat{F}_n, F_\theta) \leq 2 \cdot d_K(\hat{F}_n, F_\theta) + \gamma_n, \tag{28}$$

the Dvoretzky, Kiefer, Wolfowitz (DKW) (1956) inequality for $d_K(\hat{F}_n, F_\theta)$ and controlled $\gamma_n \leq \frac{k_n}{\sqrt{n}},\ k_n = o(\sqrt{n})$ increasing as slowly as we wish with n to infinity.

Convergence in Probability of $\tilde{\theta}_{MDE}$ to θ from convergence of $d_K(F_{\tilde{\theta}_{MDE}}, F_\theta)$ to 0 in probability will hold when

$$d_\Theta(\theta_1, \theta_2) \leq h[d_K(F_{\theta_1}, F_{\theta_2})]$$

for every θ_1, θ_2 elements of Θ and h continuous at 0.

The MDE method can be used for any functional $T(F_\theta)$ for which consistent estimate T_n exists with respect to distance $\tilde{d}$, by replacing in (42) $d_K, \hat{F}_n, F_{\theta^*}$, respectively, by $\tilde{d}, T_n, T(F_{\theta^*})$, to obtain estimate $T(F_{\tilde{\theta}_{MDE}})$ (*e.g.* Yatracos, 2019, Lemma 3.1).

8 L_1-Estimate of a probability or density via MDE with upper rates of convergence in Probability

Set-up: The observations $Y_1, \ldots, Y_n$ are *i.i.d.* random vectors from a distribution with unknown parameter $\theta \in \Theta$.

Parametric estimation problems: Θ is finite dimensional, subset of R^k for some $k \in N$, *e.g.* for a sample from a multivariate normal distribution with unknown vector of means, $\mathbf{m}$ and unknown covariance matrix Σ and the space Θ of parameters $\theta = (\mathbf{m}, \Sigma)$.

Nonparametric estimation problem: Θ is not subset of R^k for any $k,\ e.g.$ when θ is either an unknown density $f \in \Theta$ or an unknown probability $P \in \Theta$ with Θ infinite dimensional space.

Observe: when θ is a density with polynomial form of degree k then θ has at most $k+1$ unknown parameters so it is a parametric problem. If $\Theta = \mathcal{F}$ is the set of densities in $[0,1]^k$ with

p-continuous derivatives is infinite dimensional and the problem is nonparametric.

Estimation via discretization of the parameter space Θ

When we have n *i.i.d.* observations, $Y_1, \ldots, Y_n$, we cannot estimate the unknown parameter $\theta \in \Theta$ without error. Thus, we cover metric space (Θ, d) with $N(a_n)$ d-balls of radius a_n and their centers, Θ_n, is a discretization of Θ. Then, we can choose one element of the discretization, Θ_n, as the estimate of θ. This will motivate the family of pseudodistances approximating the L_1-distance.

Nonparametric estimation of densities in Θ using its discretization Θ_n and tests of hypotheses among the elements of Θ_n, with calculations of rates of convergence in Probability and in risk were provided by Le Cam (1967, 1970, 1973) and Birgé (1983) for d Hellinger and L_p-distances, $p \geq 1$. We will present a Minimum Distance Estimate MDE) of the unknown parameter with calculation of L_1-upper convergence rates in probability to the true underlying θ, either probability or density, uniformly in Θ. All these results assume the family of the underlying probabilities $\mathcal{P}$ to be determined and known.

Under mild assumptions, similar results will be presented for the case $\mathcal{P}$ is either unknown or the probabilities indexed by $\theta \in \Theta$ are intractable, with calculation of rates of convergence to θ using MDE for the Kolmogorov distance, d_K.

Why not stay with Wolfowitz's MDE and d_K when $\mathcal{P}$ is known? For observations in R^d, the L_1-distance between probabilities P, Q is always greater than or equal to Kolmogorov distance, d_K. Therefore small L_1 distance between two probabilities P, Q "means more" than small $d_K(P, Q)$. Recall that if $\mathcal{B}$ is the underlying Borel σ-field, $P = Q$ if $P(A) = Q(A)$ for every $A \in \mathcal{B}$.

Assumption: $\Theta = \mathcal{P} = \{P_s : s \in \mathcal{S}\}$, a set of probability measures that is L_1-totally bounded, i.e. the cardinality $N(a_n)$ of L_1-balls of radius a_n needed to cover $\mathcal{P}$ is finite for each $a_n > 0$. The n independent observations, $Y_1, \ldots, Y_n$, follow an unknown probability $P \in \mathcal{P}$.

MDE for L_1-distance: Assume the probabilities in $\mathcal{P}$ are defined on the space $\mathcal{Y}$ with σ-field

$\mathcal{A}$. The tool used is the empirical measure,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(Y_i) = \frac{\#Y_i \in A}{n},\ A \in \mathcal{A}.$$

$I_A(Y_i) = 1$ if $Y_i \in A$, and is 0 otherwise, $i = 1, \ldots, n$.

Observe: If sets of the form $(u_1, \ldots, u_n) \in \mathcal{A}$ and the probabilities in $\mathcal{P}$ have continuous densities, then for every $P_s \in \mathcal{P}$

$$||\mu_n - P_s||_1 = 2\sup\{|\mu_n(A) - P_s(A)|; A \in \mathcal{A}\} = 2,$$

and cannot obtain MDE, $P_{\hat{\theta}_{MDE}}$.

Thus, a family of **pseudo-distances**, d_n, should be determined, taking **supremum over a subclass** $\mathcal{A}_n$ of $\mathcal{A}$ such that

$$d_n(P_s, P_t) \le ||P_s - P_t||_1 \le d_n(P_s, P_t) + \delta_n, \tag{29}$$

for every s, t in $\mathcal{S}$, with $\delta_n \downarrow 0$ as n increases to infinity.

The pseudo-distance d_n in (29) should be able to discriminate/separate measures equally well as with the L_1-distance at least for each a_n-discretization, $\Theta_n = \mathcal{P}_n$, of $\mathcal{P}$, and then hopefully for $\mathcal{P}$; a_n should play a role in the determination of δ_n in (29).

Since $\mathcal{P}$ is L_1-totally bounded, denote the cardinality of the most economical Θ_n by $N(a_n)$ and if there are more than one candidates for Θ_n simply pick one,

$$\Theta_n = \{P_1, \ldots, P_{N(a_n)}\}. \tag{30}$$

The sets determining the L_1-distance of Probabilities P_i and P_j have been shown in (11) to be

$$A_{ij} = \{x : p_i(x) > p_j(x)\} = \{p_i > p_j\}, i \neq j, \tag{31}$$

where p_i, p_j are densities with respect to dominating measure μ which exists since $\mathcal{P}$ is L_1-totally bounded (*Hint:* There is an L_1-countable dense subset of $\mathcal{P}$). Therefore, densities exist for all elements of Θ_n in (30) and since

$$||P_i - P_j||_1 = 2[P_i(p_j > p_i) - P_j(p_j > p_i)]$$

it is enough to use for the separation between each each P_i, P_j the set $A_{ij} = \{p_i > p_j\}$ in the pseudodistance, therefore the pseudo-distance $d_n(P_s, P_t)$ for any P_s, P_t in $\mathcal{P}$ is

$$d_n(P_s, P_t) = 2\sup\{|P_s(A) - P_t(A)|, A \in \mathcal{A}_n\} \tag{32}$$

with

$$\mathcal{A}_n = \{A_{ij}; 1 \le i < j \le N(a_n)\} = \{\{p_i > p_j\}; 1 \le i < j \le N(a_n)\}. \tag{33}$$

A key Lemma is now provided.

Lemma 8.1 *Let $\mathcal{P} = \{P_s : s \in \mathcal{S}\}$ be L_1-totally bounded family of probability measures on space $\mathcal{Y}$ with σ-field $\mathcal{A}$ such that the smallest number of L_1-balls of radius a_n covering $\mathcal{P}$ has cardinality $N(a_n)$. Then, for the class of sets $\mathcal{A}_n(\subset \mathcal{A})$ in (33) with cardinality $card(\mathcal{A}_n) \le N^2(a_n)$ it holds for every $s,\ t$ in $\mathcal{S}$,*

$$||P_s - P_t||_1 \le 4a_n + 2\sup\{|P_s(A) - P_t(A)|; A \in \mathcal{A}_n\} = 4a_n + 2d_n(P_s, P_t), \tag{34}$$

which has the form (29).

Proof: Let P_s, P_t be elements of $\mathcal{P}$. For $a_n > 0$ let $\mathcal{P}_n$ be the centers of L_1 balls covering $\mathcal{P}$. Let P_i and P_j be, respectively, the centers of the balls where P_s and P_t live, $1 \le i \le j \le N(a_n)$. From the triangular inequality it follows that

$$||P_s - P_t||_1 \le ||P_s - P_i||_1 + ||P_i - P_j||_1 + ||P_j - P_t||_1 \le 2a_n + 2|P_i(A_{ij}) - P_j(A_{ij})|$$

$$\le 2a_n + 2|P_i(A_{ij}) - P_s(A_{ij}| + 2|P_s(A_{ij}) - P_t(A_{ij})| + 2|P_t(A_{ij}) - P_j(A_{ij})|$$

$$\le 4a_n + 2\sup\{|P_s(A) - P_t(A)|; A \in \mathcal{A}_n\} = 4a_n + d_n(P_s, P_t).$$

MDE for L_1-totally bounded $\mathcal{P}$: The MDE $P_{\hat{\theta}_{MDE}}$ of P_θ is such that

$$d_n(\mu_n, P_{\hat{\theta}_{MDE}}) = \inf\{d_n(\mu_n, P_s); s \in \Theta)\}. \tag{35}$$

In (35) it is assumed the infimum is achieved. If not γ_n will be added as in (42). The infimum could be taken instead over $s \in \Theta_n$, the discretization of Θ.

Proposition 8.1 *Let $Y_1, \ldots, Y_n$ be i.i.d. random vectors with probability $P_\theta \in \mathcal{P}$, L_1 totally bounded with Kolmogorov entropy $N(a), a > 0$. Then, there is a uniformly consistent MDE, $P_{\hat{\theta}_{MDE}}$ of P_θ with rate of convergence a_n :*

$$a_n \sim [\frac{\ln N(a_n)}{n}]^{1/2}, \tag{36}$$

when $a_n \downarrow 0$ in (36); $a_n \sim b_n$ denotes $C_1 b_n \leq a_n \leq C_2 b_n, 0 < C_1 \leq C_2$.

Proof: $P_{\hat{\theta}_{MDE}}$ is defined in (35). We have then from (34),

$$||P_{\hat{\theta}_{MDE}} - P_\theta||_1 \leq 4a_n + d_n(P_{\hat{\theta}_{MDE}}, P_\theta) \leq 4a_n + d_n(P_{\hat{\theta}_{MDE}}, \mu_n) + d_n(\mu_n, P_\theta) \leq 4a_n + 2d_n(\mu_n, P_\theta). \tag{37}$$

From Hoeffding's inequality, since $Card(\mathcal{A}_n) \leq N^2(a_n)$, $P(\cup_{i=1}^m B_i) \leq \sum_{i=1}^m P(B_i)$ and for each A in $\mathcal{A}_n$ the corresponding Probability bound for $|\mu_n(A) - P_\theta(A)|$ in $d_n(\mu_n, P_\theta)$ is uniform, it follows that

$$P[d_n(\mu_n, P_\theta] > k_n) \leq 2 \cdot N^2(a_n) \cdot e^{-2nk_n^2} \tag{38}$$

and the result follows taking $k_n = c[\frac{\ln N(a_n)}{n}]^{1/2}$, with $c > 0$ such that the upper bound in (38) converges to zero as n increases to infinity and k_n is used to bound the last term in (37).

Exercise: Show that the upper convergence rate when $\mathcal{P}$ has densities the q-smooth functions in $[0, 1]^d$ is $n^{-\frac{q}{2q+d}}$.

9 Learning about parameters with Matching

The evolution of Statistics to Data Science with the positive influence of Computer Science and Big Data, motivates the search for new tools when the sample of size n, $\mathbf{X}(\in R^{nxd})$, is generated from $\mathcal{M}(\theta)$, a quantile function or a sampler or a "black-box", $\mathcal{M}$, with input $\theta \in \mathbf{\Theta}$; $\mathbf{X}$ is indexed by $\theta, \mathbf{X}(\theta)$. In this Data-Generating Experiment (DGE), the goal is statistical inference for θ with unknown statistical nature in the intractable or unavailable cumulative distribution function (c.d.f.), F_θ, of each observation in $\mathbf{X}(\theta)$.

Matching and Fiducial Calibration ideas in Cochran and Rubin (1973) and in Rubin (1973, 1984, 2019) motivate, instead of calibrating θ's estimates, to find the best match for the observed $\mathbf{x}(\theta)$ *learning* from generated $\mathbf{X}^*(\theta^*)$, hence discovering the "best" parameter θ^* matching θ. Matching Estimation is model-free. The luxury of having $\mathcal{M}$ allows using N_{rep} repeated $\mathbf{X}^*(\theta^*)$ for each $\theta^* \in \mathbf{\Theta}$. Since models for the Data are never accurate, *Matching Comparisons* as *Learning Tool* for θ can have universal use. Matching estimation will improve with the evolution of computing capabilities allowing for more prompt comparisons, thus making it a useful tool in Machine Learning.

Matching measure is generic $\tilde{d}$-distance between empirical distributions $\hat{F}_{\mathbf{x}(\theta)}$ and $\hat{F}_{\mathbf{X}^*(\theta^*)}$ and $\hat{\theta}_{MMDE}$ is the Minimum *Matching* Distance Estimate (MMDE), *w.l.o.g*

$$\hat{\theta}_{MMDE} = \arg\{\min_{\theta^* \in \mathbf{\Theta}} \tilde{d}(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{x}})\}, \tag{39}$$

extending the classical Minimum Distance Estimation method (*e.g.*, Wolfowitz, 1957) used when $\{F_{\theta^*}; \theta^* \in \mathbf{\Theta}\}$ are tractable.

For $\epsilon > 0$, *the Matching Support Proportion* among the N_{rep} $\mathbf{X}^*(\theta^*)$ for which

$$\tilde{d}(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{x}}) \leq \epsilon, \tag{40}$$

is calculated *w.l.o.g.* for each $\theta^* \in \mathbf{\Theta}$ and the Maximum *Matching* Support Probability Estimate, $\hat{\theta}_{MMSPE}$, is obtained.

Motivation for MMSPE is that for several models, as θ^* approaches θ the higher its Matching Support Probability is, increasing to 1 (Propositions 15.2, 15.4, Remark 15.2 and Yatracos, 2020, Proposition 5.2). MMSPE is a relative of *noisy* Approximate Bayesian Computation (ABC) MLE (Dean *et. al.*, 2014, Yildirim *et al.* 2015) and is more distant from Maximum Probability Estimator (Weiss and Wolfowitz, 1967, 1974); see Remark 15.4.

The estimates are obtained using a discretization $\mathbf{\Theta}^*$ of $\mathbf{\Theta}$. Under *mild conditions* on the metric space $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ and the underlying family of *c.d.fs* $\{F_{\theta^*}, \theta^* \in \mathbf{\Theta}\}$ which is either unavailable or intractable and with $\tilde{d}$ the Kolmogorov distance d_K, it is shown that the Matching Estimate, $\tilde{\theta}$, is uniformly consistent for $\theta; \tilde{\theta} = \hat{\theta}_{MMDE}, \hat{\theta}_{MMSEP}$. The convergence rate for $\tilde{\theta}$ to θ is obtained via that of the unavailable $F_{\tilde{\theta}}$ to F_θ. The upper bounds on the d_K-rate of convergence of $F_{\tilde{\theta}}$ to F_θ and on the $d_{\mathbf{\Theta}}$-rate of $\tilde{\theta}$ to θ depend on the Kolmogorov entropy either of $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$, or of increasing sets $\mathbf{\Theta}_k$ covering $\mathbf{\Theta}$, *e.g.* when $\mathbf{\Theta} = R^m$, with m either known or unknown;$k \uparrow \infty, m \geq 1$. The rates are presented for *i.i.d.* F_θ vectors in R^d and remain valid under mixing conditions and dependence when there is exponential bound on $P[d_K(\hat{F}_n, F_\theta) > \epsilon]$ similar to the Dvoretzky-Kiefer-Wolfowitz-Massart bound; $d \geq 1, \epsilon > 0$. The rates often change in other situations of dependence, as for example in Time Series where different probability bounds hold (see, e.g., Chen and Wu, 2018).

When $\mathbf{\Theta}$ is a Euclidean space, the uniform upper $d_{\mathbf{\Theta}}$-rate in Probability has often order at most $\frac{\sqrt{\log n}}{\sqrt{n}}$; see Example 15.1. The usual $n^{-.5}$ parametric rate, *e.g.* of the MLE $\hat{\theta}_n$, or of other estimates from model-based estimation methods, is attained when models are tractable. Both Matching Estimation methods apply for any $T_n(\mathbf{X})$ estimate of $T(\theta)$, replacing in (39) and (40) $\hat{F}_{\mathbf{x}}$ by $T_n(\mathbf{x})$ and $\hat{F}_{\mathbf{X}^*(\theta)}$ by $T_n(\mathbf{X}^*(\theta^*))$; $\tilde{d}$ is generic distance.

In Examples 14.1-14.3, matching distances and support probabilities are plotted over $\mathbf{\Theta}(\subset R^m, m = 1, 2)$ for several parametric models and have extremes pointing to the true parameters. Thus, preliminary applications of the methods with a discretization over $\mathbf{\Theta}$ will indicate a compact, K, where θ lives, and then a finer discretization for K is used to reduce estimation bias. Choosing a large K may be preferred than choosing various starting points when looking for a global maximum, as in MLE. In Examples 14.4-14.6, averages of $M = 50$ Matching Estimates are used successfully with the mixture of two normal densities and with the intractable Tukey's (a, b, g, h) and the (a, b, g, k)-models (respectively in Tukey, 1977, and Haynes et al., 1997).

In DGE, there is no indication about θ-identifiability or what n is needed to discriminate parameters' values within the acceptable bias' level. Thus, the Empirical Discrimination Index (EDI) is introduced, to provide insight on the quality of θ's estimates and/or compare DGEs. In Example 16.1, Tukey's g-and-h parameter discrimination improves that of g-and-k model which is further studied for local g-discrimination in Figures 7 and 8.

EDI's use is justified from the literature. Rayner and MacGillivray (2002) indicated the difficulty in samples to discriminate distributional shapes and parameters' values for small and moderate $n,$ *e.g.* for the g-and-k and the generalized g-and-h models: "... computational Maximum Likelihood procedures are very good for very large sample sizes, but they should not necessarily be assumed to be safe for even moderately large sample sizes" (p. 58); also, "... with moderately large positive (*i.e.* to the right) skewness, the MLE method fitting to the g-and-k distribution *cannot efficiently discriminate* between *moderate positive values* and *small negative values* of the kurtosis parameter." (p.64). For Tukey's asymmetric λ-distributions and Moments estimation it is observed: "An additional difficulty with the use of this distribution when fitting through moments, is that of *nonuniqueness*, where more than one member of the family may be realized when matching the first four moments ... " (Ramberg *et al.* 1979, Rayner and MacGillivray, 2002, p. 58). Thus, Matching estimates in DGE should be examined at least locally with EDI.

Dean *et al.* (2014) prove consistency and asymptotic normality of ABC based maximum likelihood estimates. Yildirim *et al.* (2015) use sequential Monte Carlo to provide consistent and asymptotically normal estimates for parameters in hidden Markov Models with intractable likelihoods. Takafumi *et. al.* (2018) estimate parameters for simulator-based statistical models with intractable likelihood using recursive application of kernel ABC and show consistency. Bernton *et al.* (2019) provide Minimum Wasserstein distance estimates for intractable models, with their rates of convergence and asymptotic distributions for real observations only (section 2, line 4) using strong model assumptions some of which hold for the empirical c.d.f. and Kolmogorov distance, d_K. The "empirical distribution", $\hat{\mu}_n$, in the Wasserstein distance denotes simply the data, neither the empirical c.d.f., $\hat{F}_n$, nor the empirical measure, μ_n.

10 From Statistical Experiments to Data-Generating Experiments (DGE)

A Statistical Experiment, $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, consists of sample space $\mathcal{X}$ with σ-field $\mathcal{A}$, the parameter space Θ with distance d_Θ and probability measures $\mathcal{P} = \{P_{\theta^*}; \theta^* \in \Theta\}$; see *e.g.* Le Cam (1986), Le Cam and Yang(2000). $\mathbf{X} \in \mathcal{X}$ is observed from P_θ and the aim is to estimate θ and study properties of the estimate.

Instead of $\mathcal{P}$ one can use the corresponding *c.d.fs* $\mathcal{F}_{\mathbf{\Theta}} = \{F_{\theta^*}, \theta^* \in \mathbf{\Theta}\}$ with generic distance $\tilde{d}$ used also for functionals $T(F_{\theta^*}), \theta^* \in \mathbf{\Theta}$, and assume identifiability $i.e.$ $F_{\theta_1} = F_{\theta_2}$ implies $\theta_1 = \theta_2$.

Definition 10.1 *A Data-Generating Experiment (DGE) consists of* $(\mathcal{X}, \mathcal{M}_{\mathcal{X}}, \mathbf{\Theta}, \mathcal{M}_{\mathbf{\Theta}})$, *with sample and parameter spaces, respectively,* $\mathcal{X}$ *and* $\mathbf{\Theta}$, *Samplers* $\mathcal{M}_{\mathbf{\Theta}}, \mathcal{M}_{\mathcal{X}}$, *respectively, for random* Θ *and for* $\mathbf{X}$ *given* $\Theta = \theta^*$. *Underlying structure includes* σ*-fields* $\mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathbf{\Theta}}$, *prior* π *for* Θ, *c.d.f.* F_θ *for generated* $\mathbf{X}$ *given* $\Theta = \theta$, *non-available or intractable* c.d.fs $\mathcal{F}_{\mathbf{\Theta}} = \{F_{\theta^*}, \theta^* \in \mathbf{\Theta}\}$ *with distance* $\tilde{d}$, θ*-identifiability, distance* $d_{\mathbf{\Theta}}$ *for* $\mathbf{\Theta}$.

- $\mathbf{X} = \mathbf{X}(\theta) \in \mathcal{X}$ is observed and the aim is to estimate θ.
- The user can select $\theta^* \in \mathbf{\Theta}$ to draw one or more $\mathbf{X}^*(\theta^*)$ via $\mathcal{M}_{\mathcal{X}}(\theta^*)$.

DGE examples include those where data is obtained via either a Quantile function, or a Sampler, or a "Black-Box".

In the sequel, for c.d.fs $\tilde{d} = d_K$, Kolmogorov distance.

Definition 10.2 *For any two distribution functions* F, G *in* $R^d, d \geq 1$, *their Kolmogorov distance*

$$d_K(F, G) = \sup\{|F(y) - G(y)|; y \in R^d\}. \tag{41}$$

11 The Minimum Distance Method for Statistical Experiments

Wolfowitz introduced Minimum Distance Estimates (MDEs) in a series of papers in the 50's (e.g. 1957) using Kolmogorov distance d_K and empirical c.d.f. $\hat{F}_{\mathbf{X}}$ of sample $\mathbf{X}$ representing data D that is "matched" with a model from a pool of models.

Definition 11.1 *For any n-size sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ of random vectors in $R^d, n\hat{F}_{\mathbf{Y}}(y)$ denotes the number of Y_i's with all their components smaller or equal to the corresponding components of y. $\hat{F}_{\mathbf{Y}}$ is the empirical c.d.f. of $\mathbf{Y}$.*

For a Statistical Experiment with $\mathbf{X}$ having c.d.f $F_\theta \in \mathcal{F}_{\mathbf{\Theta}}$, $\mathbf{X} = \mathbf{X}(\theta)$, $\hat{\theta}_{MDE}$ satisfies

$$d_K(F_{\hat{\theta}_{MDE}}, \hat{F}_{\mathbf{X}(\theta)}) \leq \inf_{\theta^* \in \mathbf{\Theta}} d_K(F_{\theta^*}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n, \tag{42}$$

with the user's choice of $\gamma_n \downarrow 0$ as $n \uparrow \infty$, when $\gamma_n = 0$ cannot be used.

The infimum in (42) may not be achievable and by including $\gamma_n > 0$, $\tilde{\theta}_{MDE}$ is element of

$$\tilde{\mathbf{\Theta}}_n = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_{m_n}, \ldots\} \tag{43}$$

satisfying (42). Thus, $d_K(\hat{F}_{\hat{\theta}_{MDE}}, \hat{F}_{\mathbf{X}(\theta)})$ is kept small for $\hat{\theta}_{MDE} \in \tilde{\mathbf{\Theta}}_n$.

Tools for proving consistency and the uniform convergence rate $\frac{k_n}{\sqrt{n}}$ of $F_{\hat{\theta}_{MDE}}$ to F_θ are:

$$d_K(F_{\hat{\theta}_{MDE}}, F_\theta) \leq d_K(F_{\hat{\theta}_{MDE}}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta) \leq 2 \cdot d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta) + \gamma_n, \tag{44}$$

the Dvoretzky, Kiefer, Wolfowitz (DKW) (1956) inequality for $d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta)$ and controlled $\gamma_n \leq \frac{k_n}{\sqrt{n}},\ k_n = o(\sqrt{n})$ increasing as slowly as we wish with n to infinity.

The MDE method can be used for any functional $T(F_\theta)$ for which consistent estimate T_n exists with respect to distance $\tilde{d}$, by replacing in (42) $d_K, \hat{F}_{\mathbf{X}}, F_{\theta^*}$, respectively, by $\tilde{d}, T_n, T(F_{\theta^*})$, to obtain estimate $T(F_{\hat{\theta}_{MDE}})$ (*e.g.* Yatracos, 2019, Lemma 3.1).

12 The Minimum Matching Distance Method

In observational studies, Rubin (1973) matched data D with data D^* from a big data reservoir to reduce bias, using a mean matching method and nearest available pair-matching methods. In a DGE, $D = \mathbf{X} = \mathbf{X}(\theta)$ is available generated by unknown θ to be estimated, and $D^* = \mathbf{X}^*(\theta^*)$ become available via $\mathcal{M}_{\mathcal{X}}, \theta^* \in \mathbf{\Theta}$. D and D^* are replaced, respectively, by $\hat{F}_{\mathbf{X}(\theta)}, \hat{F}_{\mathbf{X}^*(\theta^*)}$.

Definition 12.1 *The Minimum Matching Distance Estimate (MMDE), $\hat{\theta}_{MMDE}$, satisfies*

$$d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{\mathbf{MMDE}})}, \hat{F}_{\mathbf{X}(\theta)}) \leq \inf_{\theta^* \in \mathbf{\Theta}} d_K(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n, \tag{45}$$

with $\gamma_n = 0$ or $\gamma_n \downarrow 0$ as $n \uparrow \infty$.

$\hat{\theta}_{MMDE}$ is not necessarily unique. γ_n appears in the upper rate of convergence of $F_{\hat{\theta}_{MMDE}}$ to F_θ and has rate smaller than the other additive components.

$(\mathcal{D})$ *Discretizations of* $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$: $\mathbf{\Theta}$'s *finite* $d_{\mathbf{\Theta}}$-discretization, $\mathbf{\Theta}^*_{\mathbf{n}}$, is used in (45) instead of $\mathbf{\Theta}$, $\mathbf{\Theta}^*_{\mathbf{n}} \uparrow \mathbf{\Theta}, Card(\mathbf{\Theta}^*_{\mathbf{n}}) = N_n$. $\theta^*_{ap,n}(s)$ is the element of $\mathbf{\Theta}^*_{\mathbf{n}}$ closest to s. When $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ is totally bounded, $\mathbf{\Theta}^*_{\mathbf{n}}$ consists of the $N_n = N(a_n)$ centers of the smallest number of $d_{\mathbf{\Theta}}$-balls of radius a_n covering $\mathbf{\Theta}$; $a_n > 0, a_n \downarrow 0$ as $n \uparrow \infty$.

The convergence rate for $\hat{\theta}_{MMDE}$ to θ is obtained via that of $F_{\hat{\theta}_{MMDE}}$ to F_θ. The parallel, matching inequality to (44) is

$$d_K(F_{\hat{\theta}_{MMDE}}, F_\theta) \leq d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta). \tag{46}$$

In a nutshell, $d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta)$ decreases to 0 in Probability, bounded above by $\frac{k_n}{\sqrt{n}}, k_n = o(\sqrt{n})$, with $k_n \uparrow \infty$ with n as slowly as we wish. $d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})})$ is bounded above in Probability by $\frac{\sqrt{\ln N_n}}{\sqrt{n}}$ by Lemma 17.1 with $\hat{\theta}_{MMDE}$ one of N_n *selected* $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$, $\frac{\ln N_n}{n} \downarrow 0, N_n \uparrow \infty$ as $n \uparrow \infty$. The "matching term", $d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}, \hat{F}_{\mathbf{X}(\theta)})$, is bounded above in Probability by a multiple of $\gamma_n + \frac{k_n}{\sqrt{n}} + d_K(F_\theta, F_{\theta^*_{ap,n}(\theta)})$ and depends on θ; k_n as above. Under mild assumptions, an upper bound in Probability is obtained for $d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta)$. Details are in Proposition 15.1 and Corollary 15.1.

Remark 12.1 *The advantage of having Sampler $\mathcal{M}_{\mathcal{X}}$ allows using N_{rep}(fixed) samples $\mathbf{X}^*(\theta^*)$ for each $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$. $\hat{\theta}_{MMDE}$ minimizing all the distances gives much weight to one sample. The Mean Matching d_K-distances, one for each θ^*, are also compared using their minimum to obtain $\hat{\theta}_{MMMDE}$, Minimum Mean Matching Distance estimate(s).*

Remark 12.2 *MMDE applies for any estimate, $T_n(\mathbf{X})$, of $T(\theta)$ with generic distance $\tilde{d}$, replacing in (45) $\hat{F}_{\mathbf{X}(\theta)}$ by $T_n(\mathbf{X}(\theta))$ and $\hat{F}_{\mathbf{X}^*(\theta)}$ by $T_n(\mathbf{X}^*(\theta^*))$.*

13 The Maximum Matching Support Probability Method

N_{rep} $\mathbf{X}^*(\theta^*)$ are used for $\theta^* \in \boldsymbol{\Theta}$.

Definition 13.1 *For $\theta^* \in \boldsymbol{\Theta},$ N_{rep} samples $\mathbf{X}_1^*(\theta^*), \ldots, \mathbf{X}_{N_{rep}}^*(\theta^*)$ are drawn via $\mathcal{M}_{\mathcal{X}}(\theta^*)$ and for $\epsilon > 0$ those supporting ϵ-matching with $\mathbf{X}(\theta) = \mathbf{x}$ are:*

$$A_\epsilon(\theta^*) = \{\mathbf{X}_j^*(\theta^*) : d_K(\hat{F}_{\mathbf{X}_j^*(\theta^*)}, \hat{F}_{\mathbf{x}(\theta)}) \leq \epsilon, j = 1, \ldots, N_{rep}\}. \tag{47}$$

The ϵ-Matching Support Proportion for θ^* is:

$$p_{\epsilon,match}(\theta^*) = \frac{Card[A_\epsilon(\theta^*)]}{N_{rep}} > 0. \tag{48}$$

The Maximum ϵ-Matching Support Probability Estimate (MMSPE) is

$$\hat{\theta}_{MMSPE} = \arg\{\max_{\theta^* \in \boldsymbol{\Theta}} p_{\epsilon,match}(\theta^*)\}. \tag{49}$$

Observe that:

a) for large N_{rep} and n,

$$p_{\epsilon,match}(\theta^*) \text{ estimates } P_{\theta^*}[\mathbf{X}^*(\theta^*) : d_K(\hat{F}_{\mathbf{X}^*(\theta^*)}, F_\theta) \leq \epsilon], \tag{50}$$

b) for all $s \in \boldsymbol{\Theta}$ and for all n by construction,

$$p_{\epsilon,match}(\hat{\theta}_{MMSPE}) \geq p_{\epsilon,match}(\theta^*_{ap,n}(s)). \tag{51}$$

Small ϵ in (47) with $p_{\epsilon,match}(\hat{\theta}_{MMSPE})$ at least .7 is the goal in practice.

In MMDE, with N_{rep} $\mathbf{X}^*(\theta^*)$ drawn for each $\theta^* \in \boldsymbol{\Theta}^*_{\mathbf{n}}$ and several candidates to choose from as $\hat{\theta}_{MMDE}$, (48) is used with ϵ equal to the upper bound in (45) and generated data supports $\arg\{max_{\theta^* \in \tilde{\boldsymbol{\Theta}}} p_{\epsilon,match}(\theta^*)\}$ as MMDE. The upper bound on the convergence rate in Proposition 15.1 holds for $\hat{\theta}_{MMSPE}$ which is also MMDE.

The convergence rate for $\hat{\theta}_{MMSPE}$ to θ is obtained via that of $F_{\hat{\theta}_{MMSPE}}$ to F_θ. Inequalities to determine the rate for $F_{\hat{\theta}_{MMSPE}}$, with $p_{\epsilon,match}(\hat{\theta}_{MMSPE})$ involved, are:

$$d_K(\hat{F}_{\hat{\theta}_{MMSPE}}, F_\theta) \leq d_K(F_{\hat{\theta}_{MMSPEE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}) + d_K(F_{\hat{\theta}_{MMSPEE}}, F_\theta)$$

$$\leq d_K(F_{\hat{\theta}_{MMSPEE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}) + d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta). \tag{52}$$

The first and the last term in upper bound (52) have uniform upper bounds in Probability with order, respectively, $\frac{\sqrt{\ln N_n}}{\sqrt{n}}$ and $\frac{k_n}{\sqrt{n}}, k_n = o(\sqrt{n})$, as explained in the paragraph after (46); choose $k_n \sim \sqrt{\ln N_n}$. The middle "matching term" is bounded by ϵ in (47).

Lemma 13.1 *For the Maximum ϵ-Matching Support Probability estimate, $\hat{\theta}_{MMSPE}$, in (49), $\mathbf{\Theta} = \mathbf{\Theta_n^*}$ with cardinality N_n,*

$$d_K(\hat{F}_{\hat{\theta}_{MMSPE}}, F_\theta) \leq C \cdot [\epsilon + \frac{\sqrt{ln\ N_n}}{\sqrt{n}}] \leq C \cdot \max\{\epsilon, \frac{\sqrt{ln\ N_n}}{\sqrt{n}}\}, \quad C > 0. \tag{53}$$

From (53) the question arises, whether uniformly in θ the order of ϵ can be at most $\frac{\sqrt{lnN_n}}{\sqrt{n}}$, with $p_{\epsilon,match}(\hat{\theta}_{MMSPE}) \uparrow 1$ as $n \uparrow \infty$. From (51), it seems clear the latter holds when there is $\theta^* \in \mathbf{\Theta_n^*}$ such that $d_K(F_{\theta^*}, F_\theta) < \epsilon$. In simulations with *i.i.d. r.vs.*, small $\epsilon > 0$, n, N_n, N_{rep} moderately large, $p_{\epsilon,match}(\hat{\theta}_{MMSPE})$ is at least .70 for Normal, Cauchy, Weibull, Uniform, Poisson models with one parameter unknown and $\hat{\theta}_{MMSPE}$ is near θ, competing well with MMDE. The results are confirmed in Propositions 15.2, 15.4 for the probabilities and in Propositions 15.3, 15.5 for the upper bounds on the convergence rates.

Remark 13.1 *When any of $\hat{\theta}_{MMDE}, \hat{\theta}_{MMMDE}, \hat{\theta}_{MMSPE}$ takes more than one values, the average is reported as the corresponding estimate.*

14 Matching Estimation Examples

The Examples have two goals. In parametric models, readers to compare the values of Matching Estimates and mainly observe how plots of matching Kolmogorov distances and matching support probabilities over $\mathbf{\Theta}$ point to the parameters and can provide indications for a compact K in R^d where θ lives via preliminary Matching Estimation. The second goal is for readers to observe the performance of Matching Estimates with intractable models: Tukey's g-and-h model (Tukey, 1977), the g-k model (Haynes *et al.*, 1997) and the mixtures of two normal distributions. M

repeated estimates are obtained with each method and their average is used with its estimated standard deviation, providing density plots for the estimates of each parameter.

In Figures 1-3, observe for several parametric models the "path" towards the unknown parameter(s) with the mean matching distances of N_{rep} $\mathbf{X}^*(\theta^*)$ getting smaller and the matching support probabilities larger along the θ^*-values, confirmed by the results in Section 15; see Propositions 15.2, 15.4 and Remark 15.2. Preliminary Matching Estimation with distant θ^* over R^d will provide paths to determine the large compact K. Alternatively, increasing compacts covering R^d can be used and K is determined concurrently with the Matching estimates.

In Examples 14.1-14.3, $\theta \in R$ for the exponential, normal and Poisson models and $\theta \in R^2$, either with equal coordinates for the Weibull, Cauchy and normal models or with different coordinates for the normal model. For MMSPE, the choice of ϵ is crucial. To determine ϵ one may use Empirical Quantiles of Kolmogorov distance between $\hat{F}_{\mathbf{X}}$ and $\hat{F}_{\mathbf{X}^*}$ (Yatracos, 2020, Section 3.1, Table 1). In the Examples, $\epsilon = .13$ is used which is the 90th Empirical quantile for the Kolmogorov distance of $\hat{F}_{\mathbf{X}(0)}$ and $\hat{F}_{\mathbf{X}^*(0)}$ from a normal distribution with mean zero and variance 1. Alternatively, ϵ can be chosen by trial with a satisfactory matching support probability and avoiding very many MMSEP candidates, starting with ϵ-value $C \cdot \frac{\sqrt{\ln n}}{\sqrt{n}}$; $.5 \leq C \leq 1.5$ is preferred for small d. When more than one elements of discretization $\mathbf{\Theta}^*$ satisfy a method's criterion, the reported estimate is their average.

Example 14.1 *The observed* $\mathbf{X}$ *consists of* $n = 100$ *i.i.d. r.vs from the exponential and Poisson models, each with parameter 5 , and from normal model with mean 5 and assumed known standard deviation* $\sigma = 1$. *It is assumed the unknown* θ *(i.e. 5) is in the compact* $[3, 8]$, *divided in 49 equal sub-intervals with their end-points elements of discretization* $\mathbf{\Theta}^*, N = 50$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta}^*$ *and the value* $\epsilon = .13$ *is used for MMSPE. Estimates appear in Table 1*[1] *and, most important, plots pointing to the parameters are in Figure 1.*

[1]Standard deviations of estimates for intractable models appear after Example 14.3.

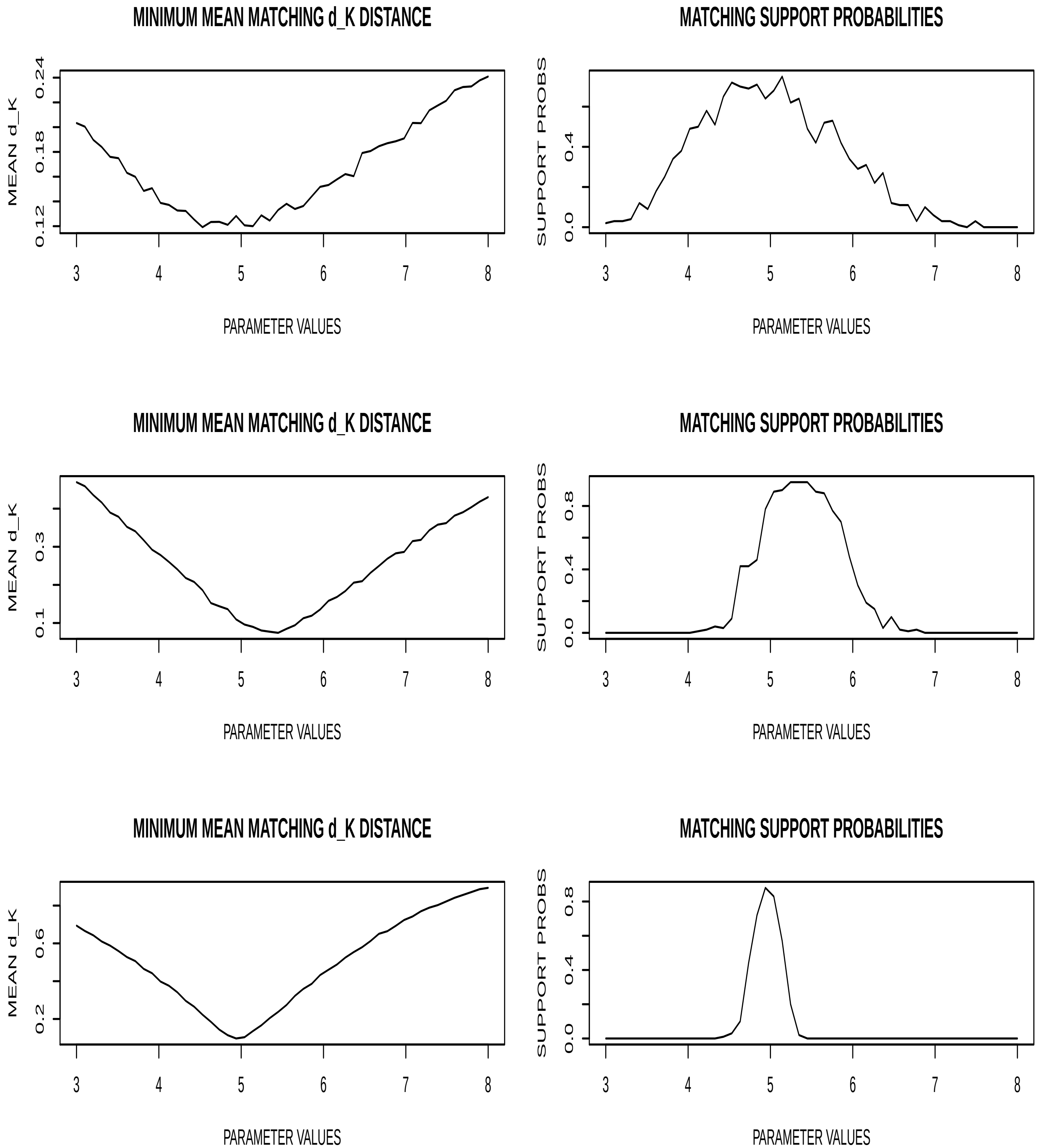

Figure 1: Row-wise, Exponential, Poisson with parameters 5, Normal mean 5, known $\sigma = 1$. Plots along $\mathbf{\Theta}$ with optima pointing to the parameters.

MATCHING ESTIMATES				
Model	MMDE	MMMDE	MMSPE	$p_{\epsilon,match}$
Exponential	5.11	4.53	5.14	0.75
Poisson	5.48	5.45	5.35	0.95
Normal	4.84	4.94	4.94	0.88

Table 1: Matching Estimation for one parameter with value 5

Example 14.2 *The observed* **X** *consists of* $n = 100$ *i.i.d. r.vs from the Weibull, Cauchy and the normal models, with both parameters equal to 5. For Matching estimation it is assumed known that these parameters are equal and only the discretization of* $[3, 8]$ *is used. The rest is as in Example 14.1. Results appear in Table 2 and plots pointing to the parameters are in Figure 2.*

MATCHING ESTIMATES				
Model	MMDE	MMMDE	MMSPE	$p_{\epsilon,match}$
Weibull	5.14	5.14	5.14	0.85
Cauchy	4.79	4.94	4.84	0.92
Normal	5.16	4.94	4.84	0.75

Table 2: Matching Estimation for two equal parameters with value 5

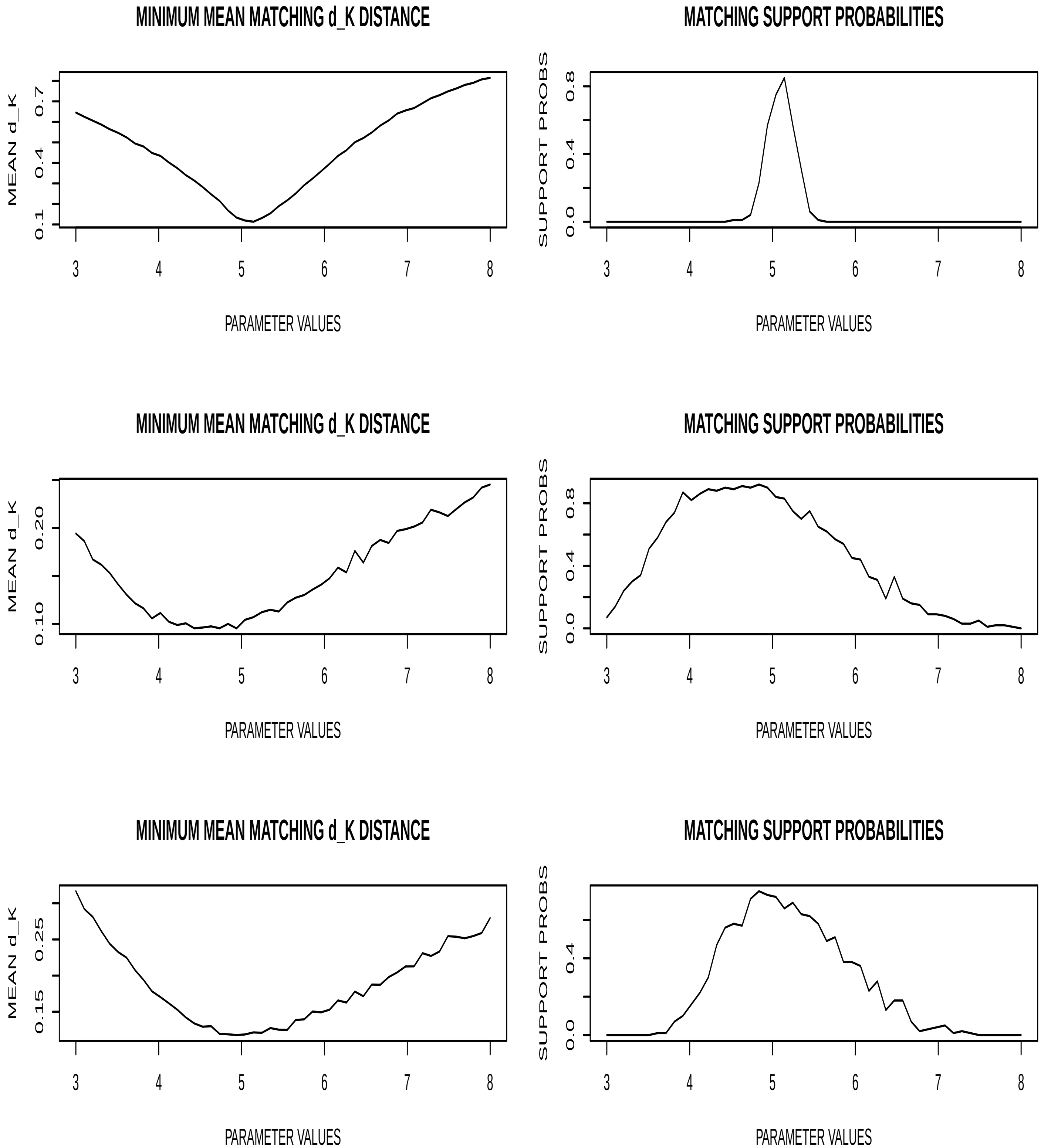

Figure 2: Row-wise, Weibull, Cauchy, Normal, Both Parameters 5. Plots along Θ with optima pointing to the parameters.

Example 14.3 *The observed* $\mathbf{X}$ *consists of* $n = 100$ *i.i.d. r.vs from the Normal model with mean* $\mu = 5$ *and standard deviation* $\sigma = 2$. *It is assumed for* $\theta = (\mu, \sigma)$ *that* $\mathbf{\Theta} = [3, 8]x[.5, 4.5]$, *discretized by dividing each interval in 49 equal sub-intervals with their end-points elements of discretization* $\mathbf{\Theta}^*$, $N = 2,500$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta}^*$ *and* $\epsilon = .13$ *is used. Estimates appear in Table 3 and the plot pointing to the parameters in Figure 3.*

MATCHING ESTIMATES FOR THE NORMAL MODEL			
Parameters	MMDE	MMMDE	MMSPE, $p_{\epsilon,match} = .9$
μ	5	5.04	4.94
σ	2.1	2.05	2.13

Table 3: Matching Estimation for parameter $\theta = (5, 2)$

Examples 14.4-14.6 present Matching estimates for intractable models. The estimation is repeated $M = 50$ times and MMDE, MMMDE and MMSEP are the averages accompanied by their standard deviation in $(\cdot)$, all in Tables 4-6.

Example 14.4 *The observed* $\mathbf{X}$ *consists of* $n = 200$ *i.i.d. r.vs,* $X_1, \ldots, X_n$, *from Tukey's* g*-and-*h *model (see, e.g., Tukey, 1977, or Yan and Genton, 2019) which accommodates data with non-Gaussian distribution, with* g *real-valued controlling skewness, non-negative* h *controlling tail heaviness and with location and scale parameters* $a \in R, b > 0$. *Standard normal* $Z_1, \ldots, Z_n$ *are used,* $a = 3, b = 4, g = 3.5, h = 2.5$ *and*

$$X_i = a + b\frac{e^{gZ_i} - 1}{g}e^{.5hZ_i^2},\ i = 1, \ldots, n. \tag{54}$$

Parameter spaces $\Theta_g, \Theta_h, \Theta_a, \Theta_b$ *are each the interval* $[2, 5]$, *divided in 10 equal sub-intervals with the 11 end-points used to obtain for* $\mathbf{\Theta} = \Theta_a x \Theta_b x \Theta_g x \Theta_h$ *discretization* $\mathbf{\Theta}^*$ *with cardinality* $N = 11^4$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta}^*$ *for Matching Estimation with* $\epsilon = .13$. *The process is repeated* $M = 50$ *times and the average Matching estimates and their estimated standard deviations are in Table 4. The distributions of the* $M = 50$ *obtained estimates for each of* g, h, a, b *are in Figure 4.*

MEAN MATCHING ESTIMATES FOR TUKEY'S g-and-h MODEL			
Parameters	MMDE & SD	MMMDE & SD	MMSPE & SD
$a = 3$	2.98 (.03)	3.04 (.04)	3.03 (.04)
$b = 4$	3.91 (.08)	4.06 (.12)	3.77 (.09)
$g = 3.5$	3.42 (.08)	3.52 (.09)	3.52 (0.07)
$h = 2.5$	2.72 (.05)	2.57 (.07)	2.93 (0.05)

Table 4: Matching Estimates with independent observations, n=200.

Example 14.5 *The observed* $\mathbf{X}$ *consists of* $n = 50$ *dependent r.vs,* $X_1, \ldots, X_n$, *from* g*-and-*k *model (Haynes* et al.*, 1997), with* g *real-valued controlling skewness,* $k > -.5$ *controlling kurtosis and with location and scale parameters* $a \in R, b > 0$. *The* g*-and-*k *distributions accommodate distributions with more negative kurtosis than the normal distribution and some bimodal distributions (Rayner and MacGillivray, 2002, p. 58). Standard normal* $Z_1, \ldots, Z_n$ *are used and*

$$X_i = a + b[1 + c \cdot \frac{1 - e^{-gZ_i}}{1 + e^{-gZ_i}}](1 + Z_i^2)^k Z_i,\ i = 1, \ldots, n; \tag{55}$$

c *is a parameter used to make the sample correspond to a density; usually* $c = .8$. *The normal variables used have covariance .5 and are obtained using* R *as one vector of size* n *from a multivariate normal. The parameters in (55) are:* $a = 3, b = 4, g = 3.5, h = 2.5$; $c = .8$. *Parameter spaces* $\Theta_g, \Theta_k, \Theta_a, \Theta_b$, *the discretization of* $\mathbf{\Theta}$ *and* ϵ *are as in Example 14.4 and Matching Estimation follows. The process is repeated* $M = 50$ *times and the average Matching estimates and their estimated standard deviations are in Table 5. The distributions of the* $M = 50$ *obtained estimates for each of* g, k, a, b *are in Figure 5.*

MEAN MATCHING ESTIMATES FOR g-and-k MODEL			
Parameters	MMDE & SD	MMMDE & SD	MMSPE & SD
$a = 3$	2.96 (.07)	3.31 (.15)	3.09 (.1)
$b = 4$	3.66 (.07)	3.81 (.14)	3.98 (.09)
$g = 3.5$	3.35 (.05)	3.54 (.12)	3.36 (.1)
$k = 2.5$	2.98 (.06)	3.08 (.12)	2.78 (.08)

Table 5: Matching Estimates with dependent observations, n=50.

Example 14.6 *The observed* $\mathbf{X}$ *consists of* $n = 200$ *independent r.vs, from a Normal mixture with two components, means* $\mu_1 = 1, \mu_2 = 6$, *standard deviations* $\sigma_1 = 1, \sigma_2 = 1.5$ *and weights, respectively,* $p = p_1 = .3, p_2 = 1 - p = .7$. *Parameter spaces* $\Theta_p = [0, 1], \Theta_{\mu_1} = [.5, 3.5], \Theta_{\mu_2} = [3.5, 6.5], \Theta_{\sigma_1} = \Theta_{\sigma_2} = [.5, 2]$, *are divided each in 10 equal sub-intervals with the 11 end-points used to obtain for* $\mathbf{\Theta} = \Theta_p x \Theta_{\mu_1} x \Theta_{\sigma_1} x \Theta_{\mu_2} x \Theta_{\sigma_2}$ *discretization* $\mathbf{\Theta}^*$ *with cardinality* $N = 11^5$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta}^*$ *for Matching Estimation with* $\epsilon = .13$. *The process is repeated* $M = 50$ *times and the average Matching estimates and their estimated standard deviations are in Table 6. The distributions of the* $M = 50$ *obtained estimates for each of* $p, \mu_1, \sigma_1, \mu_2, \sigma_2$, *are in Figure 6, using for the means* $m1, m2$ *and for the standard deviations* $s1, s2$.

MEAN MATCHING ESTIMATES FOR $pN(\mu_1, \sigma_1) + (1 - p)N(\mu_2, \sigma_2)$			
Parameters	MMDE & SD	MMMDE & SD	MMSPE & SD
$p = .3$	.31 (.002)	.32 (.006)	.34 (.002)
$\mu_1 = 1$	1.06 (.03)	1.14 (.04)	1.26 (.016)
$\sigma_1 = 1$	1.11 (.03)	1.15 (.05)	1.33 (.006)
$\mu_2 = 6$	6 (.02)	6.06 (.03)	6.12 (.02)
$\sigma_2 = 1.5$	1.51 (0.02)	1.43 (.03)	1.41 (.02)

Table 6: Matching Estimates with independent observations, n=200.

15 Rates of Convergence for Matching Estimates

15.1 Assumptions and Results

Notation: a_n has order b_n, $a_n \sim b_n$: for large n, $C_1 b_n \leq a_n \leq C_2 b_n, 0 < C_1 \leq C_2$;

$a_n \approx b_n \Longleftrightarrow \lim_{n\to\infty} \frac{a_n}{b_n} = 1.$

Assumptions used in MMDE and MMSPE

$(\mathcal{A}1)$ Continuity of F_θ: $\forall\, \theta, \theta_n \in \mathbf{\Theta},\ \lim_{n\to\infty} d_{\mathbf{\Theta}}(\theta_n, \theta) = 0 \to \lim_{n\to\infty} d_K(F_{\theta_n}, F_\theta) = 0.$

$(\mathcal{A}2)$ Dimension of $\mathbf{\Theta}$: there are $a_n \to 0$ such that $\frac{\ln N(a_n)}{n} \to 0, N(a_n) \uparrow \infty$ as $n \uparrow \infty$.

$(\mathcal{A}3)$ From F_θ to θ : w is continuous, increasing function defined on R^+ with $w(0) = 0$ and

$$d_K(F_{\theta_1}, F_{\theta_2}) \sim w(d_{\mathbf{\Theta}}(\theta_1, \theta_2)), \qquad \forall\, \theta_1, \theta_2 \in \mathbf{\Theta}, \tag{56}$$

or for small neighborhoods of F_{θ_1}.

(A1) holds for most parametric models in R^d. $(\mathcal{A}2)$ holds for sets $\mathbf{\Theta} = [-\frac{L}{2}, \frac{L}{2}]^d \subset R^d,\ L > 0$, with $a_n \sim n^{-k}, k > 0$, but also for families of functions, *e.g.* densities in a compact in R^d that have p mixed partial derivatives and the p-th derivative satisfying a Lipschitz condition with parameter, *e.g.* $\alpha \in (0, 1)$. Observe that $(\mathcal{A}3)$ implies $(\mathcal{A}1)$. $(\mathcal{A}3)$ holds for several parametric families in R with bounded densities, at least locally using the mean value theorem. $(\mathcal{A}3)$ provides the upper bound on the error rate for θ from the error rate for F_θ.

Uniform consistency of $F_{\hat\theta_{MMDE}}, F_{\hat\theta_{MMSPE}}$ to F_θ and upper bounds on the d_K-rates of convergence in Probability are initially established when $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ is totally bounded or is the union of increasing totally bounded sets. Under $(\mathcal{A}1), (\mathcal{A}2)$, the upper bound in Probability, ϵ_n^*, for the matching estimate $F_{\tilde\theta}, \tilde\theta = \hat\theta_{MMDE}, \hat\theta_{MMSPE}$, of F_θ is

$$d_K(F_{\tilde\theta}, F_\theta) \leq \epsilon_n^* \sim \max\{\sup_{s\in\mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\};$$

see (59), (71), (77). When, in addition $(\mathcal{A}3)$ holds,

$$\epsilon_n^* \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n);$$

see (60), (72), (78). The upper bound on the $d_{\mathbf{\Theta}}$-rate for $\hat{\theta}_{MMDE}, \hat{\theta}_{MMSPE}$ to θ depends on the relation between $d_K(F_{\theta_1}, F_{\theta_2})$ and $d_{\mathbf{\Theta}}(\theta_1, \theta_2)$ determined by $(\mathcal{A}3)$. The results are obtained for $i.i.d.$ vectors in R^d and it is indicated how the results are extended under dependence, *e.g.* see Roussas and Yatracos (1997).

15.2 Upper bound on the rates of convergence for MMDE

We find instructive the reader to observe the passage from the data to the parameters via the empirical c.d.fs and the intractable or unavailable models.

Proposition 15.1 *In a DGE, let $\mathbf{X} = (X_1, \ldots, X_n)$ consist of i.i.d. r.vs with c.d.f. $F_\theta \in \mathcal{F}_{\mathbf{\Theta}}$. Assume that $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ is totally bounded with discretization $\mathbf{\Theta}_{\mathbf{n}}^*$ and associated notation $a_n, N(a_n), \theta^*_{ap,n}(\theta)$ in $(\mathcal{D})$, section 4. $\mathbf{X}^*(\theta^*)$ are drawn via $\mathcal{M}_{\mathcal{X}}(\theta^*)$ for $\theta^* \in \mathbf{\Theta}_{\mathbf{n}}^*$. Obtain $\hat{\theta}_{MMDE}$ in (45) with $\mathbf{\Theta} = \mathbf{\Theta}_{\mathbf{n}}^*$.*

a) For any $\epsilon_n > 0, a_n \downarrow 0,$

$$P[d_K(\hat{F}_{\hat{\theta}_{MMDE}}, F_\theta) > \epsilon_n] \leq 6 \cdot N(a_n) \cdot \exp\{-\frac{n}{18}(\epsilon_n - d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) - \gamma_n)^2\}. \tag{57}$$

When

$$\epsilon_n = \epsilon_n(\theta) = d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + 6\frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} + \gamma_n, \tag{58}$$

the upper bound in (57) is $\frac{6}{N(a_n)}$ and converges to zero as n increases to infinity.
b) Under assumptions $(\mathcal{A}1), (\mathcal{A}2)$, ϵ_n in (58) decreases to zero in probability:
$b_1)$ The uniform upper d_K-rate of convergence, ϵ_n^*, for $\hat{F}_{\hat{\theta}_{MMDE}}$ to F_θ is:

$$\epsilon_n^* \sim \max\{\sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \tag{59}$$

$b_2)$ Using the upper bound of (56) in $(\mathcal{A}3)$, the uniform upper rate of convergence for $d_K(\hat{F}_{\hat{\theta}_{MMDE}}, F_\theta)$ in Probability to zero is:

$$\epsilon_n^* \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{60}$$

b_3) Under $(\mathcal{A}3)$, from ϵ_n^* in (60) the uniform upper rate of convergence for $d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta)$ in Probability to zero is $w^{-1}(\epsilon_n^*)$.
c) Under $(\mathcal{A}2), (\mathcal{A}3)$, with $a_n = w^{-1}(n^{-1/2})$, an upper rate in b_2) is $u_n = \sqrt{\ln N(w^{-1}(n^{-1/2}))}/\sqrt{n}$ and in b_3) is $w^{-1}(u_n)$.

Similar results hold when $\mathbf{\Theta}$ is union of increasing sequence of totally bounded sets.

Corollary 15.1 *Under the assumptions of Proposition 15.1, with $\mathbf{\Theta} = \cup_{k=1}^{\infty}\mathbf{\Theta}_k,\ \mathbf{\Theta}_k \subseteq \mathbf{\Theta}_{k+1}$, $\mathbf{\Theta}_k$ $d_{\mathbf{\Theta}}$-totally bounded, $N_k(a)$ the smallest number of $d_{\mathbf{\Theta}}$-balls of radius a covering $\mathbf{\Theta}_k$, for every $\theta \in \mathbf{\Theta}_k$ the uniform upper d_K-rate of convergence, ϵ_n^*, for $\hat{F}_{\hat{\theta}_{MMDE}}$ to F_θ is:*

$$\epsilon_n^* \sim \frac{\sqrt{\ln N_k(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{61}$$

For each $\theta \in \mathbf{\Theta}$, eventually in n, upper rates of convergence for $d_K(\hat{F}_{\hat{\theta}_{MMDE}}, F_\theta)$ and $d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta)$ are as in Proposition 15.1, $b_3), c)$ with $k = k(n) \uparrow \infty$ as $n \uparrow \infty$.

Remark 15.1 *The MMDE rates of convergence in Proposition 15.1 and Corollary 15.1 hold with observations in $R^d, d > 1$, using Lemma 17.1 with probability bound (80) U_{KW} in Remark 17.1. Similar rates hold under dependence, with the upper bound in (80) and therefore (57)-(59) all including mixing coefficient ϕ (Roussas and Yatracos, 1997, page 339, equations (8),(30)-(33)). The rates change, e.g. in Linear Time Series, using an upper probability bound in Chen and Wu (2018, p. 3, equation (8)): for $z \geq \sqrt{n}\log(n)$*

$$P[\sup_{t\in R}|\sum_{i=1}^{n} I(X_i \leq t) - F(t)| > z] \leq C_1 \frac{n}{z^{q\beta}\log^{r_0}(z)},$$

β is dependence parameter, with larger β indicating weaker dependence, q, r_0 are parameters measuring tail heaviness, $q > 1$ and $r_0 > 1; I$ is indicator function, C_1 constant. The upper probability bound is sharp.

Example 15.1 Use the assumptions of Proposition 15.1, with $\mathbf{\Theta} = R^m, m \geq 1, d_{\mathbf{\Theta}}$ the sup-norm, $w(a) = a, a \geq 0$.

a) When $\theta \in (-L/2, L/2)^m, L \geq 1, m$ known, for $a_n > 0$

$$N_L(a_n) = (\frac{L}{a_n})^m. \tag{62}$$

From (60), the upper rate of convergence in probability for $d_K(F_{\hat{\theta}_{MMDE}}, F_\theta), \theta \in [-L/2, L/2]^m$,

$$\epsilon_n^* \sim \frac{m^{1/2}(\ln L - \ln a_n)^{1/2}}{n^{1/2}} \sim a_n \tag{63}$$

and with $a_n = \frac{1}{\sqrt{n}}$ the rate of convergence is

$$m^{1/2}\frac{(\ln L + .5\ln n)^{1/2}}{n^{1/2}} \sim \frac{\sqrt{\ln n}}{\sqrt{n}}.$$

Since $d_K(F_{\theta_1}, F_{\theta_2}) \sim d_{\mathbf{\Theta}}(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in \mathbf{\Theta}$,

$$d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta) \leq C \cdot \frac{\sqrt{\ln n}}{\sqrt{n}},\ C > 0.$$

b) When $\theta \in R^m = \cup_{n=1}^{\infty}(\frac{L_n}{2}, \frac{L_n}{2})^m, m$ known and $a_n > 0$, there is n^* such that $\theta \in (-\frac{L_{n^*}}{2}, \frac{L_{n^*}}{2})^m$. Then , for $n \geq n^*$, from (63), the upper rate of convergence in probability for $d_K(F_{\hat{\theta}_{MMDE}}, F_\theta)$ is

$$\epsilon_n^* \sim \frac{m^{1/2}(\ln L_n - \ln a_n)^{1/2}}{n^{1/2}} \sim a_n. \tag{64}$$

When $a_n = \frac{1}{\sqrt{n}}$ and $L_n \leq \sqrt{n}$, for each $\theta \in R^m$, eventually in n,

$$d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta) \sim d_K(F_{\hat{\theta}_{MMDE}}, F_\theta) \leq C \cdot \frac{\sqrt{\ln n}}{\sqrt{n}},\ C > 0.$$

In a Statistical Experiment, with $\theta \in R^m$ and F_θ known but possibly inaccurate, the order of convergence in probability of an estimate to θ is often $\frac{k_n}{\sqrt{n}}, k_n = o(\sqrt{n})$ with $k_n \uparrow \infty$ as desired with n.

c) When m is unknown in a) and b), it is replaced by m_n in (63) and (64) and the rate for the upper bound is $\frac{\sqrt{m_n \cdot \ln n}}{\sqrt{n}}$, with m_n increasing to infinity as slow as desired.

15.3 Upper bound on the rates of convergence for MMSPE

Confirmation that $p_{\epsilon,match}(\hat{\theta}_{MMSPE}) \uparrow 1$ as $n \uparrow \infty$, follows for real observations, under conditions holding for mentioned models and several other parametric families, namely that $d_K(F_s, F_\theta) =$

$\Delta(>0)$ is achieved at single $x_{s,\theta} \in R$, where the difference of densities $f_s(x) - f_\theta(x)$ changes sign. Tools in the proof are limiting distributions of Kolmogorov-Smirnov type statistics for one and two samples under the Alternative (Raghavachari, 1973). By Glivenko-Cantelli theorem, *w.l.o.g.* $\hat{F}_{\mathbf{x}(\theta)}$ is replaced by F_θ in the middle matching term of (52), suggested also by the inequality preceding (52), and the result for one sample is used.

Proposition 15.2 *In a DGE, let $\mathcal{F}_{\mathbf{\Theta}}$ be a family of continuous c.d.fs in R and for $s \neq \theta$,*

$$\Delta(s,\theta) = d_K(F_s, F_\theta), \tag{65}$$

$$K_1 = \{x : F_s(x) - F_\theta(x) = \Delta(s,\theta)\}, \qquad K_2 = \{x : F_s(x) - F_\theta(x) = -\Delta(s,\theta)\}. \tag{66}$$

(A4) One of K_1, K_2 in (66) is singleton and the other empty, w.l.o.g.

$$K_1 = \{x_{s,\theta}\}, \qquad\qquad K_2 = \emptyset. \tag{67}$$

Assume $(\mathcal{A}1)$ holds and fix $\theta \in \mathbf{\Theta}, \epsilon > 0$. Then, for large n there is $s^* \in \mathbf{\Theta}$, such that

$$\Delta(s^*,\theta) \leq \epsilon - \frac{k_n^*}{\sqrt{n}},\ k_n^* = o(\sqrt{n}),\ k_n^* \uparrow \infty \text{ with } n. \tag{68}$$

If $\mathbf{X}^*(s^*)$ is a vector of n i.i.d. F_{s^*} observations obtained via $\mathcal{M}_{\mathcal{X}}(s^*)$,

$$P_{s^*}[d_K(\hat{F}_{\mathbf{X}^*(s^*)}, F_\theta) \leq \epsilon] \geq \Phi(2 \cdot k_n^*)) \uparrow 1, \text{ as } n \uparrow \infty; \tag{69}$$

Φ is the c.d.f. of standard normal. The lower bound in (69) is independent of θ, therefore it holds uniformly in θ.

Upper bounds follow on the rate of convergence of estimates for real observations and $\mathbf{\Theta} \subseteq R$.

Proposition 15.3 *In a DGE with the assumptions $(\mathcal{A}1)$ and $(\mathcal{A}4)$ in Proposition 15.2, let the observed $\mathbf{X}(\theta) = (X_1, \ldots, X_n)$ consist of i.i.d. r.vs with unknown c.d.f. $F_\theta \in \mathcal{F}_{\mathbf{\Theta}}$, $\mathbf{\Theta} \subseteq R, d_{\mathbf{\Theta}} = |\cdot|$.*

a) Assume $(\mathbf{\Theta}, |\cdot|)$ is totally bounded, w.l.o.g. $(-\frac{L}{2}, \frac{L}{2})$, with discretization $\mathbf{\Theta}^*_{\mathbf{n}}$ and notation $a_n, N(a_n), \theta^*_{ap,n}(s)$ in $(\mathcal{D})$, section 4. For every $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$, N_{rep} $\mathbf{X}^*(\theta^*)$ are drawn via $\mathcal{M}_{\mathcal{X}}(\theta^*)$.

Obtain $\hat{\theta}_{MMSPE}$ in (49) with $\mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}}$ and in (47)

$$\epsilon = \epsilon_n = \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) + \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}. \tag{70}$$

$a_1)$ *The rate of the uniform upper bound in (53) is:*

$$\tilde{\epsilon}^*_n \sim \max\{\sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \tag{71}$$

$a_2)$ *Under $(\mathcal{A}3)$, with $a_n \downarrow 0$ as $n \uparrow \infty$, $\tilde{\epsilon}^*_n$ converges to zero,*

$$\tilde{\epsilon}^*_n \sim \frac{\sqrt{-\ln a_n}}{\sqrt{n}} \sim w(a_n). \tag{72}$$

For $s^* = \theta^*_{ap,n}(\theta)$, n large, (69) holds, and the uniform upper rate of of convergence for $d_K(F_{\hat{\theta}_{MMSPE}}, F_\theta)$ in Probability to 0 is $\tilde{\epsilon}^*_n$ in (72).
$a_3)$ *Under $(\mathcal{A}3)$, the uniform upper rate of convergence for $|\hat{\theta}_{MMSPE} - \theta|$ in Probability to 0 is $w^{-1}(\tilde{\epsilon}^*_n)$, with $\tilde{\epsilon}^*_n$ in (72).*

b) Assume $(\mathcal{A}3)$ holds and $\mathbf{\Theta} = R = \cup_{n=1}^{\infty}(-\frac{k(n)}{2}, \frac{k(n)}{2})$. Then, eventually in n, the upper rate of convergence in probability for $d_K(F_{\hat{\theta}_{MMSPEE}}, F_\theta)$,

$$\tilde{\epsilon}^*_n \sim \frac{\sqrt{\ln k(n) - \ln a_n}}{\sqrt{n}} \sim w(a_n), \tag{73}$$

and for $d_{\mathbf{\Theta}}(\hat{\theta}_{MMSPEE}, \theta)$ is $w^{-1}(\tilde{\epsilon}^*_n)$.

c) Assume $(\mathcal{A}3)$ holds and $a_n = w^{-1}(n^{-1/2})$. Then, an upper rate in $a_2)$ is $u_n = \sqrt{-\ln(w^{-1}(n^{-1/2}))}/\sqrt{n}$ and in $a_3)$ is $w^{-1}(u_n)$. In b) the upper rates are, respectively, $\tilde{u}_n = \max(\sqrt{\ln k(n)}, \sqrt{-\ln(w^{-1}(n^{-1/2}))})/\sqrt{n}$ and $w^{-1}(\tilde{u}_n)$.

Proposition 15.2 is extended for $i.i.d.$ observations in R^d.

Proposition 15.4 *For $\theta \in \mathbf{\Theta}$, $\mathbf{\Theta}^*_{\mathbf{n}}$ discretization of $\mathbf{\Theta}$, $\theta^*_{ap,n}(\theta)$ the element of $\mathbf{\Theta}^*_{\mathbf{n}}$ closest to θ and n i.i.d. random vectors in R^d with c.d.f. $F_{\theta^*_{ap,n}(\theta)}$, n large:*

$$P_{\theta^*_{ap,n}(\theta)}[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_\theta) \leq \epsilon_n] \geq 1 - C_1(d) \cdot \exp\{-C_2(d) \cdot n \cdot [\epsilon_n - \sup_{s \in \Theta} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]^2\}; \tag{74}$$

$C_1(d),\ C_2(d)$ *are positive constants.*

Lower bound (74) is uniform in θ *and increases to 1 as* n *increases to infinity when*

$$n \cdot [\epsilon_n - \sup_{s \in \Theta} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]^2 \uparrow \infty \textit{ with } n. \tag{75}$$

Remark 15.2 $(\mathcal{A}3)$ *with (68), (69), (74) and (75) confirm that when* s^* *approaches* θ $p_{\epsilon,match}(s^*)$ *increases, as seen in Figures 1 and 2. Preliminary simulations indicate a large compact where* θ *lives.*

Proposition 15.3 is extended for $i.i.d.$ observations in R^d. Similar results hold under mixing conditions, as for MMDE, and when $\mathbf{\Theta}$ is union of increasing sequence of totally bounded sets, as in Corollary 15.1.

Proposition 15.5 *In a DGE, let the observed* $\mathbf{X}(\theta) = (X_1, \ldots, X_n)$ *consist of i.i.d. random vectors in* R^d *with unknown c.d.f.* $F_\theta \in \mathcal{F}_{\mathbf{\Theta}}$. *Assume that* $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ *is totally bounded with discretization* $\mathbf{\Theta}^*_{\mathbf{n}}$ *and notation* $a_n, N(a_n), \theta^*_{ap,n}(s)$ *in* $(\mathcal{D})$, *section 4.* N_{rep} $\mathbf{X}^*(\theta^*)$ *are drawn via* $\mathcal{M}_{\mathcal{X}}(\theta^*)$ *for every* $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$.

Obtain $\hat{\theta}_{MMSPE}$ *in (49) with* $\mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}}$ *and in (47)*

$$\epsilon = \epsilon_n = \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) + \frac{\sqrt{\log N(a_n)}}{\sqrt{n}}. \tag{76}$$

$a)$ *The rate of the uniform upper bound in (53) is:*

$$\tilde{\epsilon}^*_n \sim \max\{\sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \tag{77}$$

$b)$ *Under* $(\mathcal{A}2), (\mathcal{A}3), \tilde{\epsilon}^*_n$ *converges to zero with Probability increasing to 1 uniformly in* $\theta \in \mathbf{\Theta}$,

$$\tilde{\epsilon}^*_n \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{78}$$

$c)$ *Under* $(\mathcal{A}2), (\mathcal{A}3)$, *the uniform upper rate of convergence for* $d_{\mathbf{\Theta}}(\hat{\theta}_{MMSPE}, \theta)$ *in Probability to zero is* $w^{-1}(\epsilon^*_n)$, *with* ϵ^*_n *in (78).*
d) Under $(\mathcal{A}2), (\mathcal{A}3)$, *with* $a_n = w^{-1}(n^{-1/2})$, *an upper rate in* $b)$ *is* $u_n = \sqrt{\ln N(w^{-1}(n^{-1/2}))}/\sqrt{n}$ *and in* $c)$ *is* $w^{-1}(u_n)$.

Remark 15.3 *$p_{\epsilon,match}(\theta^*)$ in (48) has been introduced in F-ABC (Yatracos, 2020), an alternative to ABC with N_{rep} $\mathbf{X}^*(\theta^*)$ drawn for each θ^* to reduce the variation effect of a single $\mathbf{X}^*(\theta^*)$ in the selection of θ^*. $p_{\epsilon,match}(\theta^*)$ is used in the approximate posterior of θ if θ^* is selected.*

Remark 15.4 *MMSPE is a relative of ABC MLE (Dean* et. al.*, 2014, Yildirim* et. al. *2015) where an ϵ-neighborhood like that in (47) is used, but in ABC MLE an approximate likelihood is maximized, constructed assuming a Hidden Markov Model. MMSPE is less related with Maximum Probability Estimator (MPE) Z_n (Weiss and Wolfowitz, 1967). The reason for calling Z_n MPE is that if θ can be estimated with increasing accuracy as n increases, then MPE maximizes the asymptotic value of the expected $0-1$ gain at each point in $\mathbf{\Theta}$ among a class of decision rules (Weiss, 1983, p. 268). With $f(\mathbf{x}|\theta)$ the conditional density of $\mathbf{X}$ given θ, MPE Z_n is d maximizing*

$$\int_{\{\theta: d_{\mathbf{\Theta}}(d,\theta)\leq\epsilon/\sqrt{n}\}} f(\mathbf{x}|\theta)d\theta, \tag{79}$$

(Weiss and Wolfowitz, 1974, p. 15), which is expected to be an average of $f(\mathbf{x}|\theta)$ in a θ-neighborhood of the MLE: (79) is not a probability, it is defined via a neighborhood in $\mathbf{\Theta}$ and does not have the frequentist interpretation (48) of $p_{\epsilon,match}(\theta^*)$ for a particular θ^*.

Remark 15.5 *Rates (60), (61), (72), (73) and (78) have the form of the upper convergence rate in estimation of a density and a regression type function via Kolmogorov entropy, $\log N(a_n)$, of the corresponding space of functions that is a_n-discretized and $w(a_n) = a_n$ (see, e.g., Yatracos, 1983, 1989, 2019).*

16 Empirical Discrimination of DGE

In Rayner and MacGillivray (2002) it is indicated that there is plethora among Tukey's asymmetric-λ and g-and-h models and the g-and-k model that have shapes affected concurrently by more than one parameters and valid ML estimation requires a very large sample but Moments' estimation cannot discriminate between parameters.

Related information on θ-discrimination is missing with DGEs, since the underlying model is unknown or intractable and the "discrimination" of parameters, $d_{\mathbf{\Theta}}(\theta, \theta^*)$, cannot be associated with models' shapes via plots or their distance, *e.g.*, $d_K(F_\theta, F_{\theta^*})$.

The alternative is to use the data: estimate empirically $d_K(F_\theta, F_{\theta^*})$ by drawing $\mathbf{X}(\theta)$ and $\mathbf{X}^*(\theta^*)$, calculate $d_K(\hat{F}_{\mathbf{X}(\theta)}, \hat{F}_{\mathbf{X}^*(\theta^*)})$ and compare it with $d_{\mathbf{\Theta}}(\theta, \theta^*)$. If $\tilde{D}$ is the θ-discrimination tolerance, it is desired that when $d_{\mathbf{\Theta}}(\theta, \theta^*)$ exceeds $\tilde{D}$ then $d_K(\hat{F}_{\mathbf{X}(\theta)}, \hat{F}_{\mathbf{X}^*(\theta^*)})$ is large enough, discriminating F_θ and F_{θ^*}. The distance between the empirical c.d.fs is random and its size is reflected in the P-value of a two-sided test of hypotheses under the null, i.e. models' equality. This leads to the DGE's Empirical Discrimination Index.

When $m = 1$, the P-value for the Kolmogorov-Smirnov two-sample test of F_θ against F_{θ^*} is calculated under the null repeatedly with M samples, $\mathbf{X}(\theta)$ and $\mathbf{X}^*(\theta^*)$, and the average of P-values is the Empirical Discrimination Index, EDI$(\theta, \theta^*; DGE, n, M)$. EDI-values denoting significance indicate discrimination of models F_θ, F_{θ^*}. For $m = 2$ and $m > 2$, the approaches in Peacock (1983) and Polonik (1999) can be used to obtain P-values.

EDI can be used to evaluate locally each coordinate of the estimate $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ by calculating EDI$(\tilde{\theta}, (\tilde{\theta}_1, \ldots, \tilde{\theta}_i + \tilde{D}_i + \delta_i, \tilde{\theta}_{i+1}, ..., \tilde{\theta}_m); DGE, n, M)$, where $\tilde{D}_i$ is the tolerance for $\theta_i, \delta_i > 0, i = 1, \ldots, m$.

EDI can be used to compare DGEs. Tukey's g-and-h model (DGE 1) and the g-and-k model (DGE 2) are now compared g-locally with EDI. The same normal sample is used to obtain the i-th samples from DGE 1 and 2 and DGE with the minimum P-value is identified, $i = 1, \ldots, M$.

Example 16.1 *Samples $\mathbf{X}_1(g_1, h), \mathbf{X}_1^*(g_2, h)$ of size n are generated from Tukey's g-and-h model (DGE 1) with $g_1 = 5, g_2 = 3, h = 2.5$ and with the same standard normal variables $\mathbf{X}_2(g_1, k), \mathbf{X}_2^*(g_2, k)$ are generated from the g-and-k model (DGE 2), with $k = h$. The corresponding P-values are obtained. The experiments are repeated $M = 1000$ times for $n = 50, 100, 200, 500, 1000, 1500, 2500, 5000$ and the EDIs for DGE 1 and DGE 2 are calculated for each n, with Tukey's g-and-h model having better θ-discrimination. This is confirmed by the number of times P-value(g-and-k) is smaller than or equal to the P-value(g-and-h), which decreases as n increases; similar observation for*

$M = 10000$ *including also* $n = 10000$ *with the results available but not presented. The results appear in Table 7.*

g-LOCAL DISCRIMINATION: TUKEY'S g-and-h AND g-and-k			
n	EDI (g-and-h)	EDI (g-and-k)	# PV(g-and-k)$\leq$ # PV(g-and-h)
50	8.9 e-01	9.52 e-01	369
100	7.95 e-01	8.98 e-01	291
200	6.11 e-01	7.59 e-01	248
500	2.69 e-01	3.95 e-01	221
1000	7.29 e-02	1.29 e-01	174
1500	1.99 e-02	4.21 e-02	149
2500	1.82 e-03	5.15 e-03	144
5000	4.26 e-06	2.61 e-05	77

Table 7: Model parameters: $g_1 = 5, g_2 = 3, h = k = 2.5$. EDI-values for g based on M=1000 repeats, PV=P-value.

The results in Example 16.1 for the g-and-k model suggest comparing also estimated density plots using the data. Plots appear in Figures 7 and 8, respectively, with $g = 5$ and $g = 3.5$ and also for $g = 5$ and $g = 4.5$, with the corresponding sample size, n, and P-value for discriminating the corresponding models; $k = 2.5$. The results are in agreement with the findings in Rayner and MacGillivray (2002) but the problem seems to be the family of distributions and not the estimation methods.

17 Appendix

Theorem 17.1 *(Dvoretzky, Kiefer and Wolfowitz, 1956, and Massart, 1990, providing the tight constant) Let* $\hat{F}_{\mathbf{Y}}$ *denote the empirical c.d.f of the size* n *sample* $\mathbf{Y}$ *of i.i.d. random variables* *obtained from cumulative distribution F. Then, for any $\epsilon > 0$,*

$$P[d_K(\hat{F}_{\mathbf{Y}}, F) > \epsilon] \leq U_{DKWM} = 2e^{-2n\epsilon^2} \tag{80}$$

Lemma 17.1 *Let $\mathbf{X}$ be a sample of i.i.d. F_θ r.vs, with $\theta \in \mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}} = \{\theta^*_1, \ldots, \theta^*_{N_n}\}$. For any $\zeta > 0$ it holds for $\hat{\theta}_{MMDE}$ in (45),*

$$P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) > \zeta] \leq 2 \cdot N_n \cdot e^{-2n\zeta^2}. \tag{81}$$

When $\zeta = \frac{\sqrt{\ln N_n}}{\sqrt{n}}$, the upper bound in (81) is $\frac{2}{N_n}$ and converges to zero as N_n increases to infinity with n.

Proof of Lemma 17.1:

$$P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) > \zeta] = \sum_{i=1}^{N_n} P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) > \zeta \ \& \ \hat{\theta}_{MMDE} = \theta^*_i]$$

$$\leq \sum_{i=1}^{N_n} P^{(n)}_{\theta^*_i}[d_K(F_{\theta^*_i}, \hat{F}_{\mathbf{X}^*(\theta^*_i)}) > \zeta] \leq 2 \cdot N_n \cdot e^{-2n\zeta^2},$$

with the last inequality by Theorem 17.1. When $\zeta = \frac{\sqrt{\ln N_n}}{\sqrt{n}}$ the upper bound is $\frac{2}{N_n}$. $\square$

Remark 17.1 *Extensions of Theorem 17.1 in $R^d, d > 1$, appeared at least by Kiefer and Wolfowitz (1958), Kiefer (1961) and Devroye (1977) with corresponding upper bounds U in (80): $U_{KW} = C_1(d)e^{-C_2(d)n\epsilon^2}$, $U_K = C_3(b,d)e^{-(2-b)n\epsilon^2}$ for every $b \in (0,2)$, and $U_{De} = 2e^2(2n)^d e^{-2n\epsilon^2}$ valid for $n\epsilon^2 \geq d^2$. Thus, Lemma 17.1 holds in R^d at least when using U_{KW} and different constants.*

Proof of Lemma 13.1: The first and the last term in upper bound (52) have uniform upper bounds in Probability with order, respectively, $\frac{\sqrt{\ln N_n}}{\sqrt{n}}$ (from Lemma 17.1) and $\frac{k_n}{\sqrt{n}}, k_n = o(\sqrt{n})$ from (80); choose $k_n \sim \sqrt{\ln N_n}$. $\square$

Proof of Proposition 15.1: a) From (45), with $\mathbf{\Theta}^*_{\mathbf{n}}$ instead of $\mathbf{\Theta}$, the "matching term"

$$d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}, \hat{F}_{\mathbf{X}(\theta)}) \leq \inf_{\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}} d_K(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n \leq d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n$$

$$\leq d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n. \quad (82)$$

From (46) and (82),

$$d_K(F_{\hat{\theta}_{MMDE}}, F_\theta)$$

$$\leq d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + 2d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n. \quad (83)$$

Using (83), Lemma 17.1, the Dvoretzky-Kiefer-Wilfowitz-Massart inequality (80) and

$$\tilde{\epsilon} = \epsilon_n - d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) - \gamma_n, \quad (84)$$

$$P[d_K(\hat{F}_{\hat{\theta}_{MMDE}}, F_\theta) > \epsilon_n]$$

$$\leq P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta_{ap,n}(\theta)}, F_\theta) + 2\cdot d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n > \epsilon_n]$$

$$= P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + 2\cdot d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) > \tilde{\epsilon}]$$

$$\leq P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) > \frac{\tilde{\epsilon}}{3}] + P[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n(\theta)}}) > \frac{\tilde{\epsilon}}{3}] + P[d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) > \frac{\tilde{\epsilon}}{6}]$$

$$\leq 2\cdot N(a_n)\cdot e^{-2n\tilde{\epsilon}^2/9} + 2\cdot e^{-2n\tilde{\epsilon}^2/9} + 2\cdot e^{-2n\tilde{\epsilon}^2/36} = 2\cdot[N(a_n)+1]e^{-2n\tilde{\epsilon}^2/9} + 2\cdot e^{-n\tilde{\epsilon}^2/18} \leq [2N(a_n)+4]e^{-n\tilde{\epsilon}^2/18}$$

$$\leq 6\cdot N(a_n)\cdot e^{-n\tilde{\epsilon}^2/18}. \quad (85)$$

From (58) and (84),

$$\tilde{\epsilon} = \epsilon_n - d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) - \gamma_n = 6\frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}$$

and upper bound (85) becomes.

$$6\cdot N(a_n)\cdot e^{-n\tilde{\epsilon}^2/18} = 6\cdot N(a_n)\cdot e^{-2\ln N(a_n)} = \frac{6}{N(a_n)}.$$

b_1) (59) follows from (58) since γ_n can be of smaller order than the other terms.

b_2) Since $d_{\mathbf{\Theta}}(\theta^*_{ap,n}(s), s) \leq a_n$ and w is increasing, from (58)

$$\epsilon_n \leq C\cdot w(a_n) + 6\frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} + \gamma_n,\ 1 \leq C, \quad (86)$$

and the uniform upper rate of convergence (60) follows ignoring γ_n.

b_3) Follows from (60) and the properties of w.

c) For b_2), u_n follows from (86) with $a_n = w^{-1}(n^{-1/2})$ and $(\mathcal{A}3)$ implies the rate for b_3). □

Proof of Corollary 15.1: (61) follows from (60). Let $k = k(n) \uparrow \infty$ as $n \uparrow \infty$. Then, for each $\theta \in \mathbf{\Theta}$ there is $k^* = k(n^*) : \theta \in \mathbf{\Theta}_{k(n)}$ for $n \geq n^*$. Then for θ (61) holds, with $k = k(n), n \geq n^*$. Rates follow taking $a_n = w^{-1}(n^{-1/2})$ as in Proposition 15.1, $b_3), c)$, replacing N by N_k. □

Proof of Proposition 15.2: Under $(\mathcal{A}4)$ and a result in Raghavachari (1973, Theorem 2, p. 68, or Serfling, 1980, p. 112), for the given θ, any other $s \in \mathbf{\Theta}$ and $\mathbf{X}^*(s)$ $i.i.d$ sample of size m from $F_s, \delta \in R$,

$$\lim_{m\to\infty} P_s[\sqrt{m}(d_K(\hat{F}_{\mathbf{X}^*(s)}, F_\theta) - \Delta(s,\theta) \leq \delta] = \Phi(\frac{\delta}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}). \tag{87}$$

When $\delta > 0$,

$$\Phi(\frac{\delta}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}) \geq \Phi(2 \cdot \delta). \tag{88}$$

From (87), for the given ϵ, θ and large m,

$$P_s[d_K(\hat{F}_{\mathbf{X}^*(s)}, F_\theta) \leq \epsilon] \approx \Phi(\frac{\sqrt{m}(\epsilon - \Delta(s,\theta))}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}), \tag{89}$$

with "$\approx$" denoting asymptotic equality.

From $(\mathcal{A}1)$, for large n there is $s^* \in \mathbf{\Theta}$:

$$\Delta(s^*,\theta) \leq \epsilon - \frac{k_n^*}{\sqrt{n}},\ k_n^* = o(\sqrt{n}),\ k_n^* \uparrow \infty \text{ with } n. \tag{90}$$

For $s = s^*, m = n$ in (89) and from (88),

$$P_{s^*}[d_K(\hat{F}_{\mathbf{X}^*(s^*)}, F_\theta) \leq \epsilon] \approx \Phi(\frac{\sqrt{n} \cdot (\epsilon - \Delta(s^*,\theta))}{\sqrt{F_{s^*}(x_{s^*,\theta})(1 - F_{s^*}(x_{s^*,\theta}))}} \geq \Phi(2 \cdot \sqrt{n} \cdot (\epsilon - \Delta(s^*,\theta)) \geq \Phi(2 \cdot k_n^*). \tag{91}$$
□.

Proof of Proposition 15.3: a_1) $\tilde{\epsilon}_n^*$ follows from (53), with $\epsilon = \epsilon_n$ in (70), $N_n = N(a_n)$.
a_2) Since $a_n \downarrow 0$ as $n \uparrow \infty$, from $(\mathcal{A}1)$ and $(\mathcal{A}3)$, $\tilde{\epsilon}_n^*$ decreases to zero as n increases and (72) follows from (62) with $d = 1$. For $\theta^*_{ap,n}(\theta)$,

$$\Delta(\theta^*_{ap,n}(\theta),\theta) \leq \sup_{s\in\mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) \leq \epsilon_n - \frac{.5 \cdot \sqrt{\ln N(a_n)}}{\sqrt{n}},$$

with the last inequality due to (70). Then, for large n, (90) (same with (68)) holds with $s^* = \theta^*_{ap,n}(\theta)$ and $k^*_n = .5 \cdot \sqrt{\ln N(a_n)}$. Hence, from (91) for large n,

$$P_{\theta^*_{ap,n}(\theta)}[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_\theta) \leq \epsilon_n] \geq \Phi(2 \cdot \sqrt{n} \cdot (\epsilon_n - \Delta(\theta^*_{ap,n}(\theta), \theta)) \geq \Phi(2 \cdot k^*_n) \uparrow 1 \text{ with } n \uparrow \infty.$$

Convergence in Probability for $\hat{\theta}_{MMSPE}$ follows from its construction and (50), (51).

a_3) Follows from $(\mathcal{A}2), (\mathcal{A}3)$, (72) and the properties of w.

b) When $\mathbf{\Theta} = R = \cup_{n=1}^{\infty}(-\frac{k(n)}{2}, \frac{k(n)}{2})$, there is n^* such that $\theta \in (-\frac{k(n^*)}{2}, \frac{k(n^*)}{2})$ and for $n \geq n^*$, from (62), the upper rate of convergence in probability for $d_K(F_{\hat{\theta}_{MMSPEE}}, F_\theta)$

$$\epsilon^*_n \sim \frac{(\ln k(n) - \ln a_n)^{1/2}}{n^{1/2}} \sim w(a_n).$$

c) Replace $a_n = w^{-1}(n^{-1/2})$ in (72) and (73) to obtain the upper rates u_n and $\tilde{u}_n$ for $d_K(F_{\hat{\theta}_{MMSPE}}, F_\theta)$. Their images for w^{-1} are upper rates for $|\hat{\theta}_{MMSPE} - \theta|$. □

Proof of Proposition 15.4: Since

$$d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_\theta) \leq d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta)$$

$$\leq d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s),$$

$$P[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_\theta) > \epsilon_n] \leq P[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) > \epsilon]$$

$$= P[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) > \epsilon_n - \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]$$

$$\leq C_1(d) \cdot \exp\{-C_2(d) \cdot n \cdot [\epsilon_n - \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]^2\},$$

with the last inequality obtained using U_{KW} in the upper bound (80) as suggested in Remark 17.1. (74) and (75) follow. □

Proof of Proposition 15.5: $a)$ $\tilde{\epsilon}^*_n$ follows from (53), with $\epsilon = \epsilon_n$ in (76), $N_n = N(a_n)$.

$b)$ Follows from assumptions $(\mathcal{A}2), (\mathcal{A}3)$, (74), (75) The result for $\hat{\theta}_{MMSPE}$ follows from its construction and (50), (51).

c) Follows from $(\mathcal{A}2), (\mathcal{A}3)$, (78) and the properties of w.

d) For $b)$, u_n follows from (78) with $a_n = w^{-1}(n^{-1/2})$ and $(\mathcal{A}3)$ implies the rate for $c)$. □

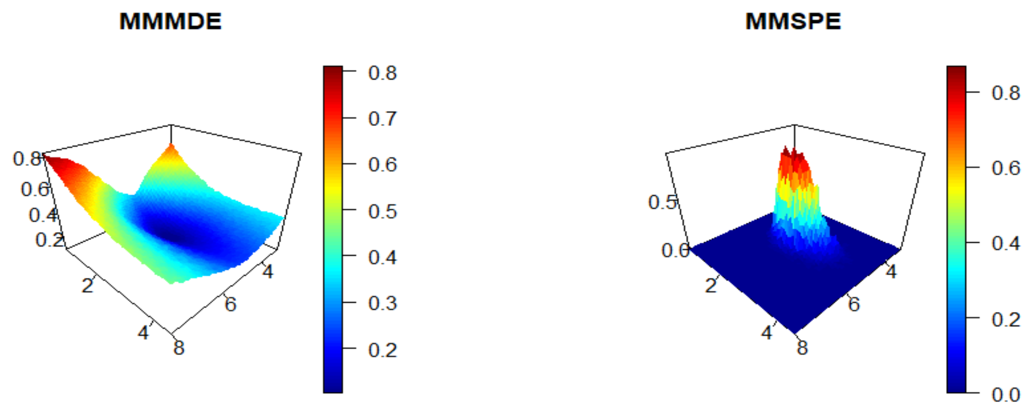

Figure 3: Parameter space $\mathbf{\Theta} = [3, 8]x[0.5, 4.5]$, Model Parameter $\theta = (\mu = 5, \sigma = 2)$. Plot along $\mathbf{\Theta}$ with optimum pointing to the parameters.

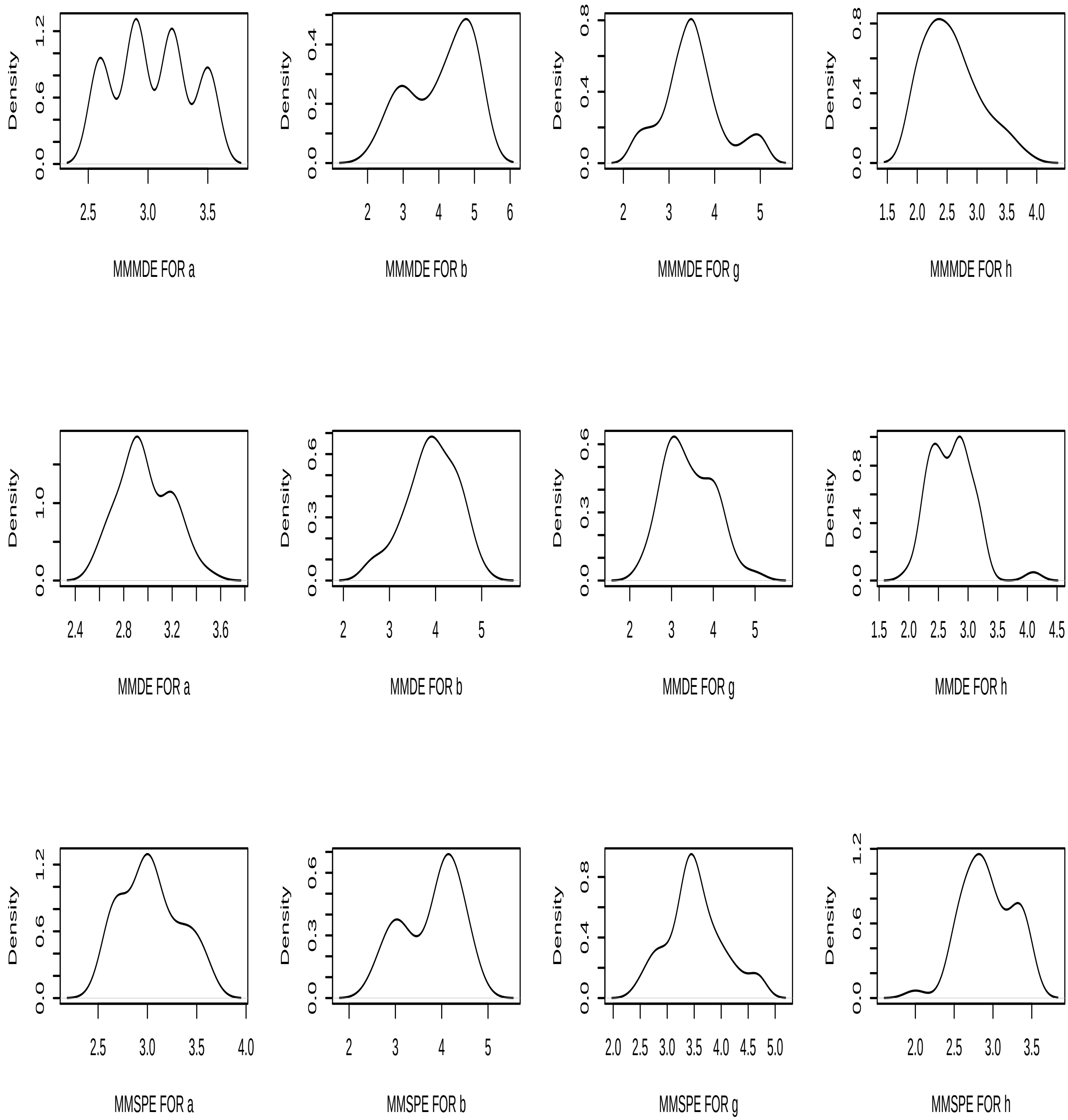

Figure 4: Density plots for the 50 estimates of Tukey's g-and-h model with independent samples, $n = 200$. The parameters are $a = 3, b = 4, g = 3.5, h = 2.5$.

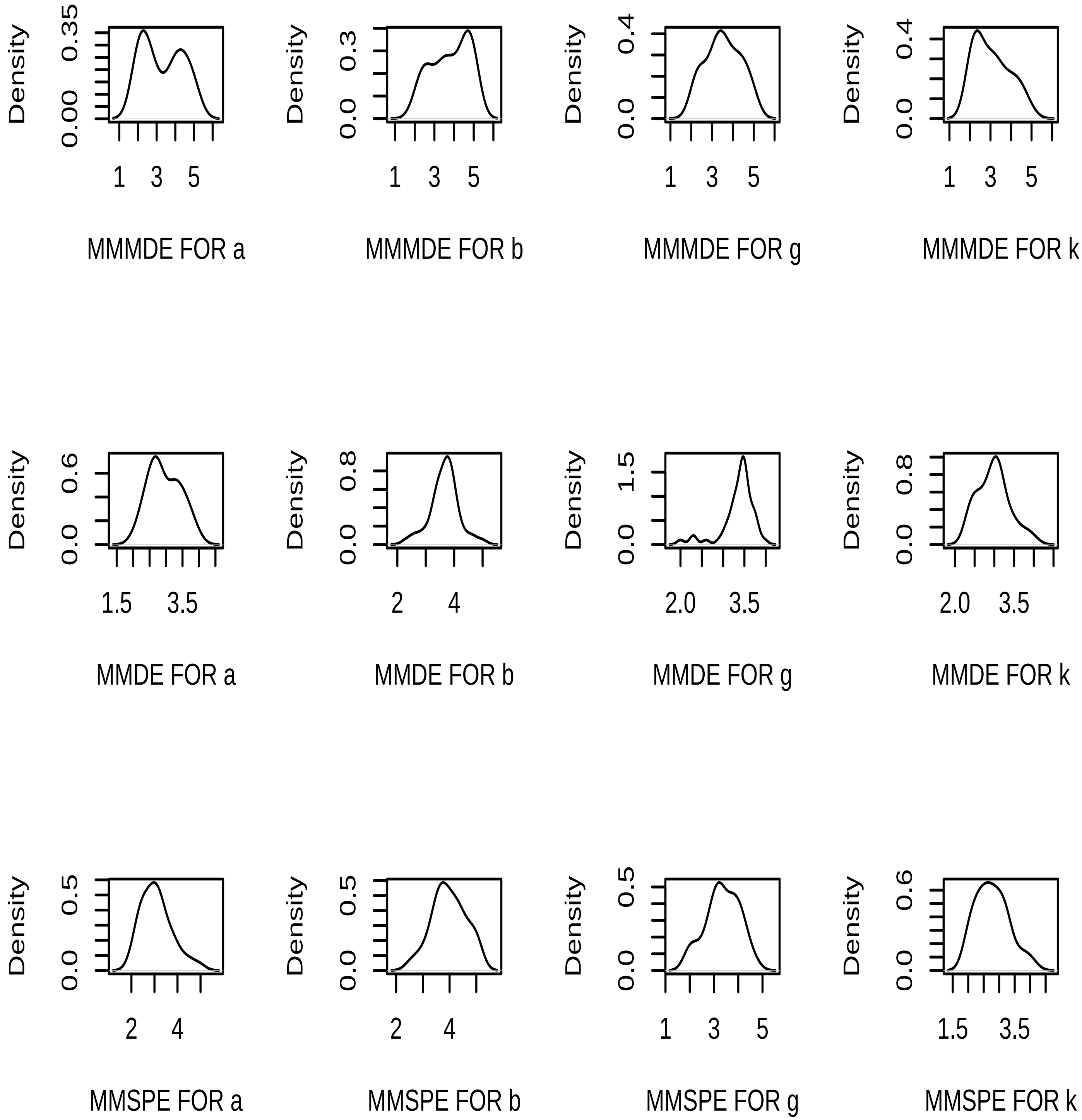

Figure 5: Density plots for 50 estimates of g-and-k model with dependent samples, $n = 50$. The parameters are $a = 3, b = 4, g = 3.5, k = 2.5$

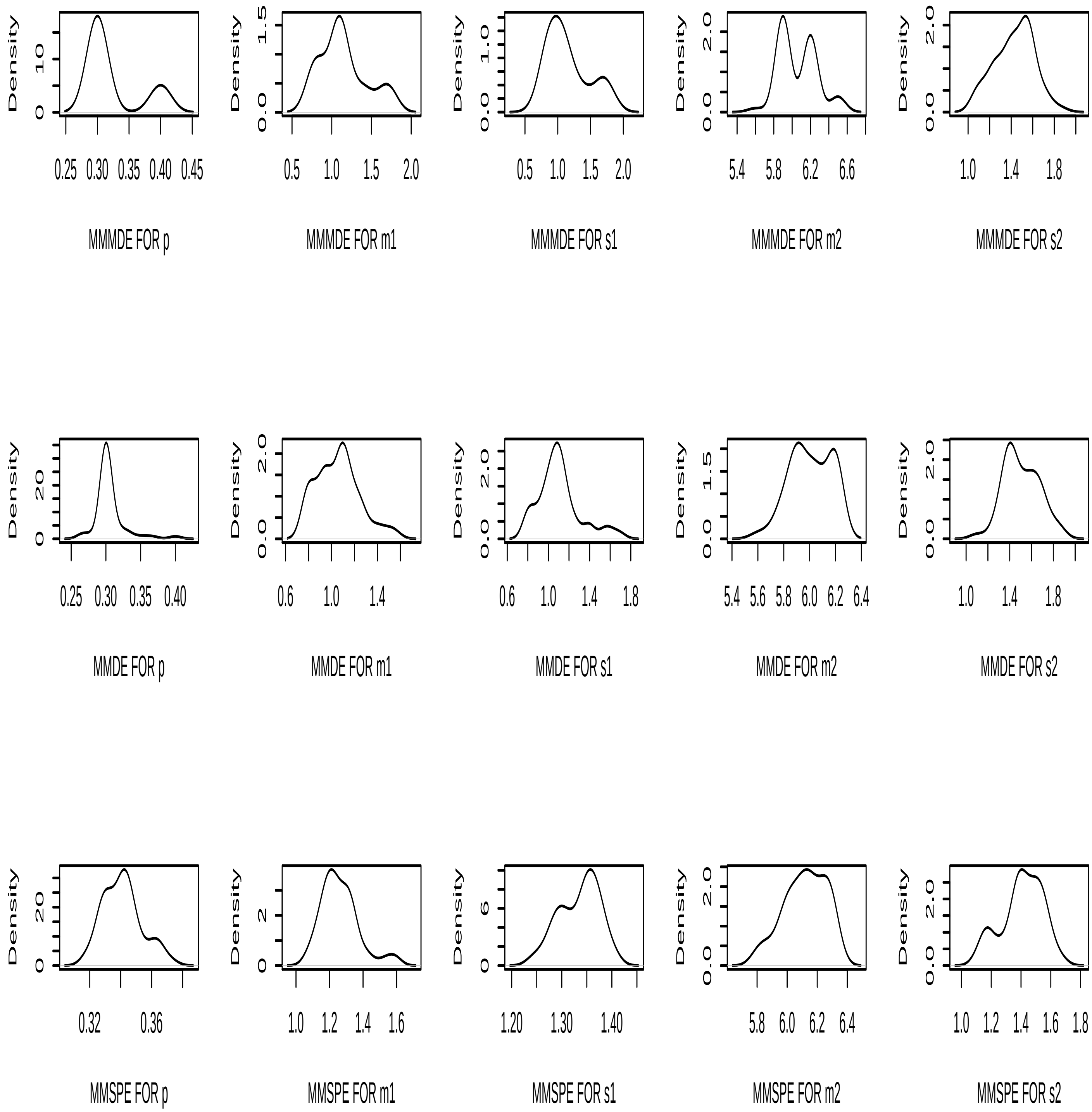

Figure 6: Density plots for the 50 estimates of the normal mixture with independent samples, $n = 200$; the parameters are p=.3, μ_1=m1=1, σ_1=s1=1, μ_2=m2=6, σ_2=s2=1.5.

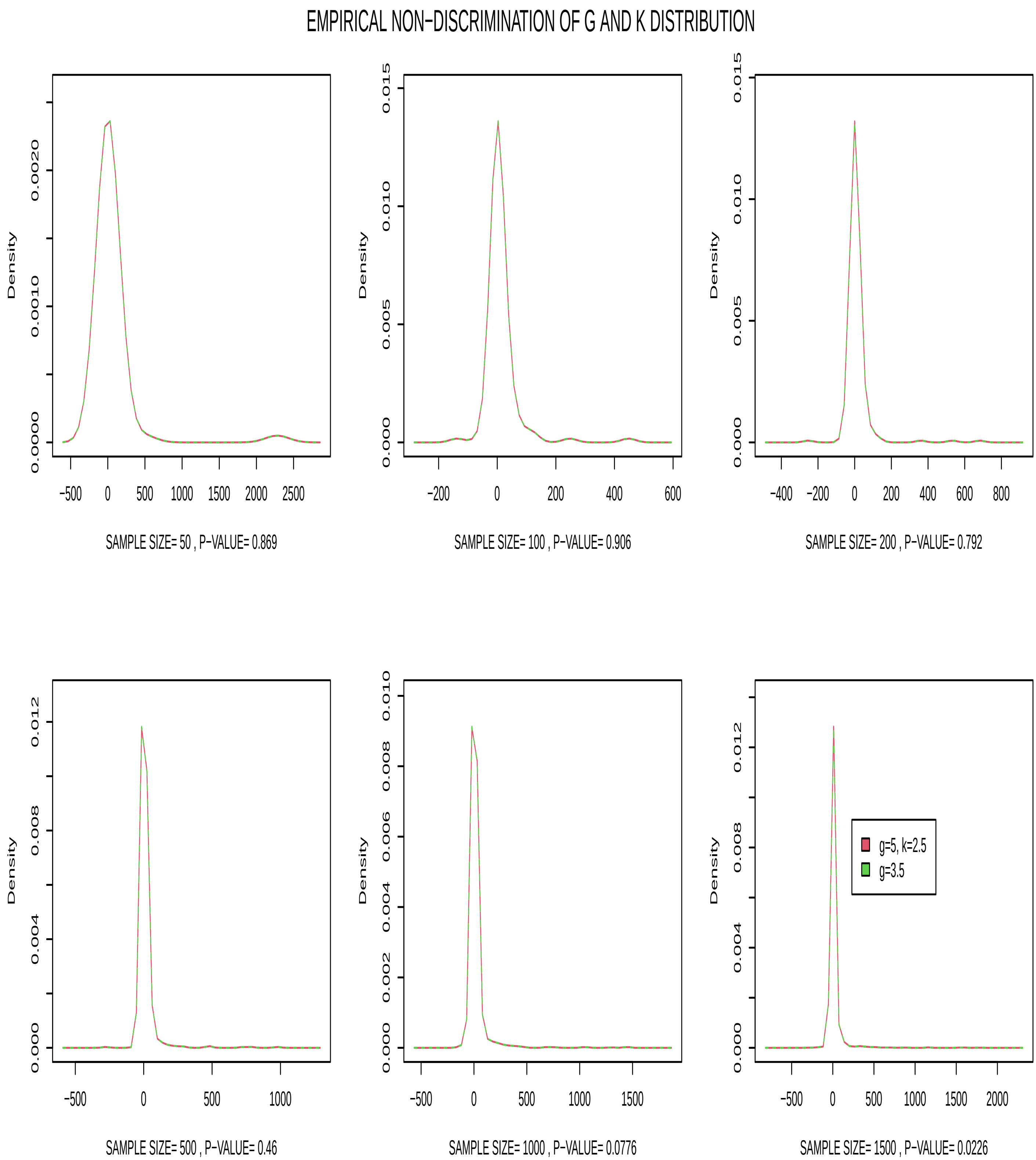

Figure 7: Visual comparison of estimated density plots for g-and-k data and Kolmogorov-Smirnov P-values.

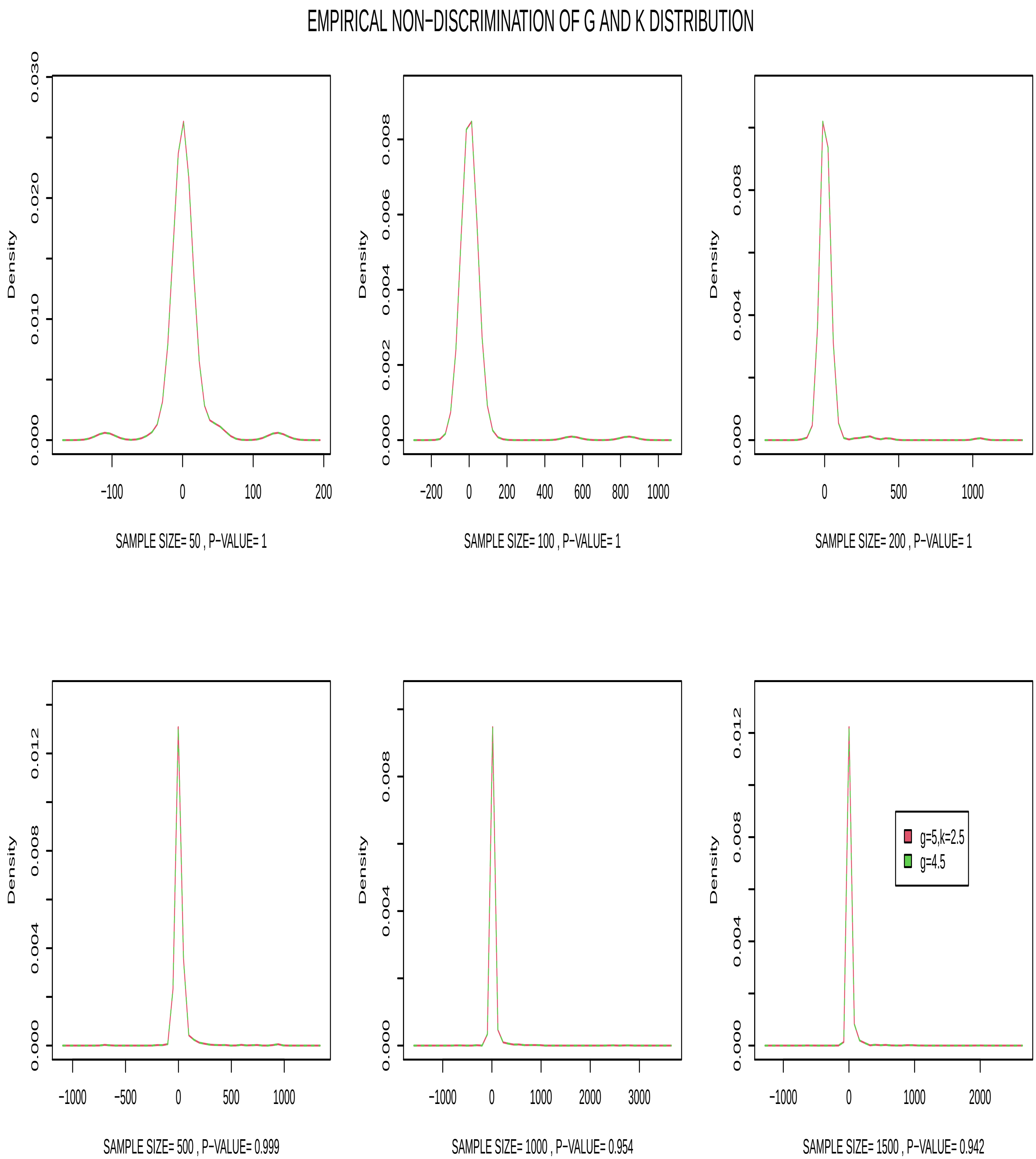

Figure 8: Visual comparison of estimated density plots for g-and-k data and Kolmogorov-Smirnov P-values.