LEARNING WITH MATCHING
IN
DATA-GENERATING EXPERIMENTS

Yannis G. Yatracos

Yau Mathematical Sciences Center

Tsinghua University, Beijing

e-mail: yatracos@tsinghua.edu.cn yannis.yatracos@gmail.com

June 20, 2021

Summary

In a Data-Generating Experiment, the observed sample, $\mathbf{x}$, has intractable or unavailable c.d.f., F_θ, and θ's statistical nature is unknown; θ is element of metric space $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$. *Matching estimates* of θ are introduced, *learning* from the "best" $\mathbf{x}$-matches with samples $\mathbf{X}^*$ from $F_{\theta^*}, \theta^* \in \mathbf{\Theta}$. Under mild conditions, these *nonparametric* estimates are uniformly consistent and the upper bounds on their rates of convergence in probability have the same rate and depend on the Kolmogorov entropies of an increasing sequence of sets covering $\mathbf{\Theta}$. When $\mathbf{\Theta} \subseteq R^m$ and the observations are *i.i.d.* the upper bounds can be, $\frac{\sqrt{\log n}}{\sqrt{n}}$ when m is known, and $\frac{\sqrt{m_n \cdot \log n}}{\sqrt{n}}$ when m is unknown; $m \geq 1, m_n \uparrow \infty$ at a desired rate. Upper bounds can also be obtained for dependent observations. These rates hold for observations in R^d, complementing recent results obtained for real, *i.i.d.* observations, under stronger assumptions and using weak probability distances; $d \geq 1$. In simulations, the Matching estimates are successful for the mixture of 2 normals and for Tukey's (a, b, g, h) and the (a, b, g, k) models. Computers' evolution will allow for more and faster comparisons, resulting in improved Matching estimates for universal use in Machine Learning.

Some key words: Data Generating Experiment; Intractable models; Kolmogorov entropy; Learning with Matching; Maximum Matching Support Probability Estimate; Minimum Matching Distance Estimates; Nonparametric Estimation

1 Introduction

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 $\mathcal{M}(\theta)$, *e.g.*, 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 $\mathbf{X}(\theta)$. The approach is nonparametric, extending the use of Minimum Distance estimation method for intractable or unavailable underlying c.d.fs, and introducing the Maximum Matching Support Probability estimates.

Matching and Fiducial Calibration ideas in Cochran and Rubin (1973) and in Rubin (1973, 1984, 2019) motivate finding the best match for the observed $\mathbf{x}(\theta)$, *learning* from generated $\mathbf{X}^*(\theta^*)$ for several θ^*, hence discovering the "best" parameter $\hat{\theta}^*$ among them 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 and faster comparisons, thus making it a useful tool in Machine Learning.

The Matching measure is a generic $\tilde{d}$-distance between empirical distributions $\hat{F}_{\mathbf{x}(\theta)}$ and $\hat{F}_{\mathbf{X}^*(\theta^*)}$, Two estimates are presented:
a) $\hat{\theta}_{MMDE}$ is the Minimum *Matching* Distance Estimate (MMDE),

$$\hat{\theta}_{MMDE} = \arg\{\min_{\theta^* \in \mathbf{\Theta}} \tilde{d}(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{x}})\}, \tag{1}$$

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

b) For $\epsilon > 0$, we measure for each $\theta^* \in \mathbf{\Theta}$ the proportion of the N_{rep} $\mathbf{X}^*(\theta^*)$ for which

$$\tilde{d}(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{x}}) \leq \epsilon, \tag{2}$$

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 7.2, 7.4, Remark 7.2 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 7.4.

In practice, the Matching estimates are obtained using a discretization, $\mathbf{\Theta}^*$, of $\mathbf{\Theta}$. Under *mild conditions* on the metric space $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$, on 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 θ; $\tilde{\theta}$ denotes either $\hat{\theta}_{MMDE}$ or $\hat{\theta}_{MMSPE}$. 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 coincide, as well as those on the $d_{\mathbf{\Theta}}$-rate of $\tilde{\theta}$ to θ and depend on the Kolmogorov entropy of metric space space $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$[1], or those of increasing sets $\mathbf{\Theta}_k$ covering $\mathbf{\Theta}$, *e.g.* when $\mathbf{\Theta}$ is 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 can be similarly obtained under mixing conditions and dependence when there is exponential bound on $P[d_K(\hat{F}_{\mathbf{X}}, F_\theta) > \epsilon]$ similar to the Dvoretzky-Kiefer-Wolfowitz-Massart bound in (43); $d \geq 1, \epsilon > 0$. The rates may change under 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 of unknown dimension, m, the uniform upper $d_{\mathbf{\Theta}}$-rate in Probability for $\hat{\theta}_{MMDE}$ and $\hat{\theta}_{MMSPE}$, has often order at most $\frac{\sqrt{m_n \log n}}{\sqrt{n}}$ with $m_n \uparrow \infty$ at any desired rate; see Example 7.1. Note that the MLE and other model-based estimation methods cannot be used with DGE and comparison with these Matching Estimates is meaningless. Both Matching Estimation methods apply for any $T(\mathbf{X})$ estimate of θ, replacing in (1) and (2) $\hat{F}_{\mathbf{x}}$ by $T(\mathbf{x})$ and $\hat{F}_{\mathbf{X}^*(\theta)}$ by $T(\mathbf{X}^*(\theta^*))$; $\tilde{d}$ is generic distance.

In Examples 6.1-6.3, matching distances and support probabilities are plotted over $\mathbf{\Theta}(\subseteq R^m, m = 1, 2)$ for several parametric models and have extremes near 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

[1]The Kolmogorov entropy of metric space $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$ is $\log_2 N(a)$, with $N(a)$ the minimum number of balls of radius a needed to cover $\mathbf{\Theta}$.

bias. Choosing a large K may be preferred than choosing various starting points when looking for a global maximum, as in MLE. In Examples 6.4-6.6, 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).

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. Kajihara *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.* (2019a, b) and Briol *et al.* (2020) use the empirical distribution, $\hat{\mu}_n(x) = n^{-1}\sum_{i=1}^n \delta_{X_i}(x)$, to provide estimates for $\theta; \delta_{x^*}(x)$ is the Dirac function with mass 1 at $x = x^*(\in R^d), \mathbf{X} = (X_1, \ldots, X_n)$. $\hat{\mu}_n$ is sufficient only for real observations and is neither the empirical c.d.f., $\hat{F}_{\mathbf{X}}$, nor the empirical measure, μ_n, that are indexed by Borel sets, $\mathcal{B}_d$, in $R^d, d \geq 1$. Main drawback of $\hat{\mu}_n$ is the inadequate information it provides for F_θ and the induced probability P_θ and so for θ, since it is evaluated at singletons, vanishes except for the sample (where it takes value $\frac{1}{n}$) and, *most important*, unlike $\hat{F}_{\mathbf{X}}$ and μ_n, $\hat{\mu}_n$ does not use the information in the sample about P_θ on $\mathcal{B}_d$ *which determines* F_θ *and* P_θ*;* $d \geq 1$. This information is valuable when matching $\mathbf{x}$ and $\mathbf{x}^*$, since probabilities P and Q in $(R^d, \mathcal{B}_d)$ are equal (*i.e.* P and Q "match") if and only if $P(A) = Q(A)$ for every $A \in \mathcal{B}_b$.

Bernton *et al.* (2019a, b) provide Minimum Wasserstein distance estimates for intractable models, with their rates of convergence and asymptotic distributions for real observations only (2019a, section 3.3, last paragraph, 2019b, section 2, line 4). Briol *et al.* (2020) use $\hat{\mu}_n$, which is not a probability, after embedding it with kernel, k, in a space of probability measures/c.d.fs, $\mathcal{P}_k$, and use a divergence measure, the Maximum Mean Discrepancy (MMD). The subjective choice of k is a serious concern since it shapes *arbitrarily* the meager information in $\hat{\mu}_n$ about F_θ that is used in MMD. This is confirmed also by the authors in Section 4. Theoretical results hold for *i.i.d.* observations; see section 3.1.

In Bernton *et al.* (2019a, last paragraph of section 3,) it is added: "In high dimensions, the rate of convergence of the Wasserstein distance between empirical measures is known to be slow (Talagrand, 1994)." and " Detailed analysis of WABC's dependence on dimension is an interesting avenue of future research." The statements hold for their counterpart in Bernton *et al.* (2019b), with focus on estimation, but also Briol *et al.* (2020) with high dimensional observations, because of the use of $\hat{\mu}_n$. In addition, unlike the proposed Matching Estimation methods, the dimension of $\mathbf{\Theta}$ needs to be known.

It is unavoidable, that several assumptions are required for $\hat{\mu}_n$ to convey information for θ, among which that $\hat{\mu}_n$, even though it is neither density, nor *c.d.f.*, nor probability, converges to P_θ in either probability or almost surely with respect to the *weak* Wasserstein distance, *e.g.*, see Assumptions 1 and 2 (Bernton *et al.*, 2019a), Assumption 2.1 (Bernton *et al.*, 2019b). Also, that c.d.fs $\{F_{\theta^*}, \theta^* \in \mathbf{\Theta}\}$ are subset of $\mathcal{P}_k$ and Assumption 1 on MMD which implies the bounds on Theorem 1 and Lemma 1 (Briol *et al.*, 2019). These assumptions naturally hold with *i.i.d.* and dependent observations for the empirical *c.d.f.*, $\hat{F}_{\mathbf{X}}$, and the empirical measure, $\mu_{\mathbf{X}}$, due to Glivenko-Cantelli Theorem and Large Deviations' inequalities.

In sections 2-5, the Matching estimation methods are briefly introduced, applications appear in section 6 and theoretical results with proofs are in sections 7 and 8.

2 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 $\mathbf{\Theta}$ with distance $d_{\mathbf{\Theta}}$, and probability measures $\mathcal{P} = \{P_{\theta^*}; \theta^* \in \mathbf{\Theta}\}$; 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 2.1 *A Data-Generating Experiment (DGE) consists of* $(\mathcal{X}, \mathcal{M}_{\mathcal{X}}, \boldsymbol{\Theta}, \mathcal{M}_{\boldsymbol{\Theta}})$, *with sample and parameter spaces, respectively,* $\mathcal{X}$ *and* $\boldsymbol{\Theta}$, *Samplers* $\mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{M}_{\mathcal{X}}$, *respectively, for random* Θ *and for* $\mathbf{X}$ *given* $\Theta = \theta^*$. *Underlying structure includes* σ*-fields* $\mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\boldsymbol{\Theta}}$, *prior* π *for* Θ, *c.d.f.* F_θ *for generated* $\mathbf{X}$ *given* $\Theta = \theta$, *non-available or intractable* c.d.fs $\mathcal{F}_{\boldsymbol{\Theta}} = \{F_{\theta^*}, \theta^* \in \boldsymbol{\Theta}\}$ *with distance* $\tilde{d}$, θ*-identifiability and distance* $d_{\boldsymbol{\Theta}}$ *on* $\boldsymbol{\Theta}$.

- $\mathbf{X} = \mathbf{X}(\theta) \in \mathcal{X}$ is observed and the aim is to estimate θ.
- The user can select $\theta^* \in \boldsymbol{\Theta}$ 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, $\tilde{d}$ is replaced for c.d.fs by the Kolmogorov distance, d_K.

Definition 2.2 *For any two distribution functions* F, G *in* $R^d, d \geq 1$, *their Kolmogorov distance*

$$d_K(F, G) = \sup\{|F(y) - G(y)|; y \in R^d\}. \tag{3}$$

3 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 d_K and with the empirical c.d.f., $\hat{F}_{\mathbf{X}}$, of sample $\mathbf{X}$ representing data, D, that is "matched" with a c.d.f. from a pool of c.d.fs.

Definition 3.1 *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 $F_\theta \in \mathcal{F}_{\boldsymbol{\Theta}}$, $\mathbf{X} = \mathbf{X}(\theta)$, $\hat{\theta}_{MDE}$ satisfies

$$d_K(F_{\hat{\theta}_{MDE}}, \hat{F}_{\mathbf{X}(\theta)}) \leq \inf_{\theta^* \in \boldsymbol{\Theta}} d_K(F_{\theta^*}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n, \tag{4}$$

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

The infimum in (4) may not be achievable and by including $\gamma_n > 0$, $\tilde{\theta}_{MDE}$ is element of

$$\tilde{\mathbf{\Theta}}_n = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_{m_n}, \ldots\} \tag{5}$$

satisfying (4). 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:

$$d_K(F_{\hat{\theta}_{MDE}}, F_\theta) \leq d_K(F_{\hat{\theta}_{MDE}}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta) \leq 2 \cdot d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta) + \gamma_n, \tag{6}$$

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 (4) $d_K, \hat{F}_{\mathbf{X}}, F_{\theta^*}$, respectively, by $\tilde{d}, T_n, T(F_{\theta^*})$, to obtain estimate $T(F_{\hat{\theta}_{MDE}})$ with the form of the functional; see, *e.g.*, Yatracos, 2019, Lemma 3.1.

4 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 for matching, respectively, by $\hat{F}_{\mathbf{X}(\theta)}, \hat{F}_{\mathbf{X}^*(\theta^*)}$.

Definition 4.1 *The Minimum Matching Distance Estimate (MMDE), $\hat{\theta}_{MMDE}$, satisfies*

$$d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{\mathbf{MMDE}})}, \hat{F}_{\mathbf{X}(\theta)}) \leq \inf_{\theta^* \in \mathbf{\Theta}} d_K(\hat{F}_{\mathbf{X}^*(\theta^*)}, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n, \tag{7}$$

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 (7) 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 $n \uparrow \infty$. $\log_2 N(a), a > 0$, is Kolmogorov's entropy of $(\mathbf{\Theta}, d_{\mathbf{\Theta}})$. In the sequel, $\log_2 N(a)$ and $\ln N(a)$ are used interchangeably.

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 (6) is

$$d_K(F_{\hat{\theta}_{MMDE}}, F_\theta) \le d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta). \quad (8)$$

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 8.1 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 $\theta; k_n$ as above. Under mild assumptions, an upper bound in Probability is obtained for $d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta)$. Details are in Proposition 7.1 and Corollary 7.1.

Remark 4.1 *The advantage of having Sampler, $\mathcal{M}_{\mathcal{X}}$, allows using N_{rep}(fixed) samples $\mathbf{X}^*(\theta^*)$ for each $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$. $\hat{\theta}_{MMDE}$ minimizes all the distances and gives much weight to one sample. The Mean Matching d_K-distances, one for each $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$, are also compared and their minimum provides $\hat{\theta}_{MMMDE}$, the Minimum Mean Matching Distance estimate(s).*

Remark 4.2 *MMDE applies for any estimate, $T_n(\mathbf{X})$, of $T(\theta)$ with generic distance $\tilde{d}$, replacing in (7) $\hat{F}_{\mathbf{X}(\theta)}$ by $T_n(\mathbf{X}(\theta))$ and $\hat{F}_{\mathbf{X}^*(\theta)}$ by $T_n(\mathbf{X}^*(\theta^*))$.*

5 The Maximum Matching Support Probability Method

Definition 5.1 *For $\theta^* \in \mathbf{\Theta}$, 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:*

$$A_\epsilon(\theta^*) = \{\mathbf{X}^*_j(\theta^*) : d_K(\hat{F}_{\mathbf{X}^*_j(\theta^*)}, \hat{F}_{\mathbf{x}(\theta)}) \le \epsilon, j = 1, \ldots, N_{rep}\}. \quad (9)$$

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

$$p_{\epsilon,match}(\theta^*) = \frac{Card[A_\epsilon(\theta^*)]}{N_{rep}} > 0. \tag{10}$$

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

$$\hat{\theta}_{MMSPE} = \arg\{\max_{\theta^* \in \mathbf{\Theta}} p_{\epsilon,match}(\theta^*)\}. \tag{11}$$

Observe that:

a) for large N_{rep} and n,

$$p_{\epsilon,match}(\theta^*) \text{ estimates } P_{\theta^*}[\mathbf{X}^*(\theta^*) : d_K(\hat{F}_{\mathbf{X}^*(\theta^*)}, F_\theta) \leq \epsilon], \tag{12}$$

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

$$p_{\epsilon,match}(\hat{\theta}_{MMSPE}) \geq p_{\epsilon,match}(\theta^*_{ap,n}(s)). \tag{13}$$

Remark 5.1 *$\hat{\theta}_{MMSPE}$ depends crucially on ϵ and the cardinality of discretization, $\mathbf{\Theta}^*_{\mathbf{n}}$, that replaces $\mathbf{\Theta}$ in (11). When $A_\epsilon(\theta^*)$ is empty, ϵ is increased. When the histogram of the matching support probabilities, $p_{\epsilon,match}(\theta^*), \theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$, is nearly flat on a large neighborhood, ϵ is decreased. A finer discretization is needed when the smooth histogram forms an open palm. When $\mathbf{\Theta} \subseteq R^m, m \geq 2$, the size of discretization depends on the difficulty in estimating each θ's coordinate. This holds also for $\hat{\theta}_{MMDE}$.*

Observe that when N_{rep} $\mathbf{X}^*(\theta^*)$ are drawn for each $\theta^* \in \mathbf{\Theta}^*_{\mathbf{n}}$ and the upper bound in (7) is used as ϵ in (10), $\hat{\theta}_{MMSPE}$ is MMDE as element of $\arg\{max_{\theta^* \in \tilde{\mathbf{\Theta}}} p_{\epsilon,match}(\theta^*)\}$, with upper bound on the convergence rate as in Proposition 7.1.

For other values of ϵ, 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:

$$\begin{aligned} d_K(F_{\hat{\theta}_{MMSPE}}, F_\theta) &\leq d_K(F_{\hat{\theta}_{MMSPE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}) + d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}, F_\theta) \\ &\leq d_K(F_{\hat{\theta}_{MMSPE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}) + d_K(\hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMSPE})}, \hat{F}_{\mathbf{X}(\theta)}) + d_K(\hat{F}_{\mathbf{X}(\theta)}, F_\theta). \end{aligned} \tag{14}$$

The first and the last term in upper bound (14) 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 (8); choose $k_n \sim \sqrt{\ln N_n}$. The middle "matching term" is bounded by ϵ in (9).

Lemma 5.1 *For the Maximum ϵ-Matching Support Probability estimate, $\hat{\theta}_{MMSPE}$, in (11), $\mathbf{\Theta} = \mathbf{\Theta_n^*}$ with cardinality N_n,*

$$d_K(F_{\hat{\theta}_{MMSPE}}, F_\theta) \leq C \cdot [\epsilon + \frac{\sqrt{ln\ N_n}}{\sqrt{n}}] \leq C \cdot \max\{\epsilon, \frac{\sqrt{ln\ N_n}}{\sqrt{n}}\}, \quad C > 0. \qquad (15)$$

From (15) 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 (13), 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 7.2, 7.4 for the probabilities and in Propositions 7.3, 7.5 for the upper bounds on the convergence rates.

Remark 5.2 *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.*

6 Applications

For tractable, parametric models, observe in Examples 6.1-6.3, Figures 1-3, the "path" towards the unknown parameter(s), as the mean matching distances of N_{rep} $\mathbf{X}^*(\theta^*), \theta^* \in \mathbf{\Theta_n^*}$, are getting smaller and the matching support probabilities are getting larger, confirmed by the results in Section 7; see Propositions 7.2, 7.4 and Remark 7.2. Preliminary Matching Estimation with distant θ^* over R^m will provide a path to determine the large compact, K, where θ lives. Alternatively, increasing compacts covering R^d can be used and K is determined concurrently with the Matching estimates.

In all the MMSPE applications, 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 between $n^{-.5}$

and $3n^{-.5}$. When more than one elements of discretization $\mathbf{\Theta_n^*}$ satisfy a method's criterion, the reported estimate is their average. Standard deviations of estimates for intractable models appear in Examples 6.4-6.6.

Example 6.1 *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* $\mathbf{\Theta_n^*}$ *with cardinality* $N = 50$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta_n^*}$ *and the value* $\epsilon = .13$ *is used for MMSPE. Estimates appear in Table 1 and, most important, plots with paths pointing to the parameters are in Figure 1.*

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 6.2 *The observed* $\mathbf{X}$ *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 6.1. Results appear in Table 2 and plots pointing to the parameters are in Figure 2.*

Example 6.3 *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 forming the discretization* $\mathbf{\Theta_n^*}$ *with cardinality* $N = 2,500$. $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta_n^*}$ *and* $\epsilon = .13$ *is used. Estimates appear in Table 3 and the plot pointing to the parameters is in Figure 3.*

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

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 6.4-6.6 present Matching estimates for Tukey's g-and-h model (Tukey, 1977), the g-k model (Haynes *et al.*, 1997) and the mixtures of two normal distributions. The estimation is repeated $M = 50$ times and MMDE, MMMDE and MMSEP denote the averages accompanied by their estimated standard deviation in $(\cdot)$, all in Tables 4-6. Density plots for the $M = 50$ estimates of each parameter are in Figures 4-6.

Example 6.4 *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), 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*

$$X_i = a + b\frac{e^{gZ_i} - 1}{g}e^{.5hZ_i^2}, \; i = 1, \ldots, n. \tag{16}$$

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* $\mathbf{\Theta}^*_{\mathbf{n}}$ *with cardinality* $N = 11^4$*.* $N_{rep} = 100$ *samples of size* n *are obtained using each element of* $\mathbf{\Theta}^*_{\mathbf{n}}$ *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 6.5 *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*

$$X_i = a + b[1 + c \cdot \frac{1 - e^{-gZ_i}}{1 + e^{-gZ_i}}](1 + Z_i^2)^k Z_i, \; i = 1, \ldots, n; \tag{17}$$

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 (17) 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 6.4 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.*

Example 6.6 *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* $\mathbf{\Theta_n^*}$ *with cardinality* $N = 11^5$*.* $N_{rep} = 100$ *samples of size* n *are obtained using each element of*

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.

$\mathbf{\Theta_n^*}$ *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.

Example 6.7 *Rates of convergence of Matching etimates are obtained for* $\mathbf{\Theta} \subseteq R^m$*, with* m *either known or unknown, under the assumptions and with the results in Section 7. Example 7.1 is presented for* $\hat{\theta}_{MMDE}$ *but holds also for* $\hat{\theta}_{MMSPE}$*, since the upper bounds on the rates of convergence coincide; see (19).*

7 Rates of Convergence for Matching Estimates

7.1 Assumptions and Results

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

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

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

$$d_K(F_{\theta_1}, F_{\theta_2}) \sim w(d_{\mathbf{\Theta}}(\theta_1, \theta_2)), \qquad \forall\ \theta_1, \theta_2 \in \mathbf{\Theta}, \tag{18}$$

or for small neighborhoods of F_{θ_1}.

(A1) holds for most parametric models in R^d. ($\mathcal{A}2$) holds for sets $\mathbf{\Theta} = [-\frac{L}{2}, \frac{L}{2}]^m$, $L > 0$, 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.

In a nutshell, 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 $(\mathcal{A}1), (\mathcal{A}2)$ and with notation $a_n, N(a_n), \theta^*_{ap,n}(\theta)$ in $(\mathcal{D})$, section 4, 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

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

see (22), (34), (40). When, in addition ($\mathcal{A}3$) holds,

$$\epsilon^*_n \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n); \tag{19}$$

see (23), (35), (41). 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 Remark 7.1.

7.2 Upper bound on the rates of convergence for MMDE

The reader can observe in Proposition 7.1 *a)-c)* the passage of the rates, from the data to the parameters, via the empirical c.d.fs and the intractable or unavailable c.d.fs.

Proposition 7.1 *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 (7) with* $\mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}}$.
a) For any $\epsilon_n > 0, a_n \downarrow 0$,

$$P[d_K(F_{\hat{\theta}_{MMDE}}, F_\theta) > \epsilon_n] \leq 6 \cdot N(a_n) \cdot \exp\{-\frac{n}{18}(\epsilon_n - d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) - \gamma_n)^2\}. \tag{20}$$

When

$$\epsilon_n = \epsilon_n(\theta) = d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + 6\frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} + \gamma_n, \tag{21}$$

the upper bound in (20) 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 (21) decreases to zero in probability:*
b_1) *The uniform upper* d_K*-rate of convergence,* ϵ^*_n, *for* $F_{\hat{\theta}_{MMDE}}$ *to* F_θ *is:*

$$\epsilon^*_n \sim \max\{\sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \tag{22}$$

b_2) *Using the upper bound of (18) 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:*

$$\epsilon^*_n \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{23}$$

b_3) *Under* $(\mathcal{A}3)$, *from* ϵ^*_n *in (23) 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 7.1 *Under the assumptions of Proposition 7.1, 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* $F_{\hat{\theta}_{MMDE}}$ *to* F_θ *is:*

$$\epsilon_n^* \sim \frac{\sqrt{\ln N_k(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{24}$$

For each $\theta \in \mathbf{\Theta}$*, eventually in* n*, upper rates of convergence for* $d_K(F_{\hat{\theta}_{MMDE}}, F_\theta)$ *and* $d_{\mathbf{\Theta}}(\hat{\theta}_{MMDE}, \theta)$ *are as in Proposition 7.1,* $b_3), c)$ *with* $k = k(n) \uparrow \infty$ *as* $n \uparrow \infty$.

Remark 7.1 *The MMDE rates of convergence in Proposition 7.1 and Corollary 7.1 hold with observations in* $R^d, d > 1$*, using Lemma 8.1 with probability bound (43)* U_{KW} *in Remark 8.1. Similar rates hold under dependence, with the upper bound in (43) and therefore (20)-(22) 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 7.1 Upper rates of convergence of $\hat{\theta}_{MMDE}$ are obtained under the assumptions of Proposition 7.1, with $\mathbf{\Theta} \subseteq R^m, m \geq 1, d_{\mathbf{\Theta}}$ the sup-norm, $w(a) = a, a \geq 0$. The same rates hold also for $\hat{\theta}_{MMSPE}$.

a) When $\theta \in (-L/2, L/2)^m, L \geq 1, m$ known, for $a_n > 0$ used in the discretization of the parameter space,

$$N_L(a_n) = (\frac{L}{a_n})^m. \tag{25}$$

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

$$\epsilon_n^* \sim \frac{m^{1/2}(\ln L - \ln a_n)^{1/2}}{n^{1/2}} \sim a_n \tag{26}$$

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 (26), the upper rate of convergence in probability for $d_K(F_{\hat{\theta}_{MMDE}}, F_\theta)$ is

$$\epsilon_n^* \sim \frac{m^{1/2}(\ln L_n - \ln a_n)^{1/2}}{n^{1/2}} \sim a_n. \tag{27}$$

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 (26) and (27) 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.

7.3 Upper bound on the rates of convergence for MMSPE

The confirmation that $p_{\epsilon,match}(\hat{\theta}_{MMSPE}) \uparrow 1$ as $n \uparrow \infty$, follows for real observations, under conditions holding for models used in Example 6.1 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 (14), suggested also by the inequality preceding (14), and the result for one sample is used.

Proposition 7.2 *In a DGE, let $\mathcal{F}_{\mathbf{\Theta}}$ be a family of continuous c.d.fs in R and for $s \neq \theta$,*

$$\Delta(s,\theta) = d_K(F_s, F_\theta), \tag{28}$$

$$K_1 = \{x : F_s(x) - F_\theta(x) = \Delta(s,\theta)\}, \qquad K_2 = \{x : F_s(x) - F_\theta(x) = -\Delta(s,\theta)\}. \tag{29}$$

(A4) *One of K_1, K_2 in (29) is singleton and the other empty, w.l.o.g.*

$$K_1 = \{x_{s,\theta}\}, \qquad K_2 = \emptyset. \tag{30}$$

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

$$\Delta(s^*,\theta) \leq \epsilon - \frac{k_n^*}{\sqrt{n}}, \;\; k_n^* = o(\sqrt{n}), \;\; k_n^* \uparrow \infty \text{ with } n. \tag{31}$$

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

$$P_{s^*}[d_K(\hat{F}_{\mathbf{X}^*(s^*)}, F_\theta) \leq \epsilon] \geq \Phi(2 \cdot k_n^*)) \uparrow 1, \;\; \text{as } n \uparrow \infty; \tag{32}$$

Φ is the c.d.f. of standard normal. The lower bound in (32) 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 7.3 *In a DGE with the assumptions $(\mathcal{A}1)$ and $(\mathcal{A}4)$ in Proposition 7.2, 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 (11) with $\mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}}$ and in (9)

$$\epsilon = \epsilon_n = \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) + \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}. \tag{33}$$

a_1) The rate of the uniform upper bound in (15) is:

$$\tilde{\epsilon}_n^* \sim \max\{\sup_{s\in\mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \quad (34)$$

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

$$\tilde{\epsilon}_n^* \sim \frac{\sqrt{-\ln a_n}}{\sqrt{n}} \sim w(a_n). \quad (35)$$

For $s^* = \theta^*_{ap,n}(\theta), n$ large, (32) 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 (35).
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 (35).

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

$$\tilde{\epsilon}_n^* \sim \frac{\sqrt{\ln k(n) - \ln a_n}}{\sqrt{n}} \sim w(a_n), \quad (36)$$

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 7.2 is extended for $i.i.d.$ observations in R^d.

Proposition 7.4 *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:*

$$P_{\theta^*_{ap,n}(\theta)}[d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_\theta) \le \epsilon_n] \ge 1 - C_1(d)\cdot\exp\{-C_2(d)\cdot n\cdot[\epsilon_n - \sup_{s\in\Theta} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]^2\}; \quad (37)$$

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

Lower bound (37) is uniform in θ and increases to 1 as n increases to infinity when

$$n\cdot[\epsilon_n - \sup_{s\in\Theta} d_K(F_{\theta^*_{ap,n}(s)}, F_s)]^2 \uparrow \infty \text{ with } n. \quad (38)$$

Remark 7.2 ($\mathcal{A}3$) *with (31), (32), (37) and (38) 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 7.3 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 7.1.

Proposition 7.5 *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 (11) with* $\mathbf{\Theta} = \mathbf{\Theta}^*_{\mathbf{n}}$ *and in (9)*

$$\epsilon = \epsilon_n = \sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s) + \frac{\sqrt{\log N(a_n)}}{\sqrt{n}}. \tag{39}$$

a) The rate of the uniform upper bound in (15) is:

$$\tilde{\epsilon}^*_n \sim \max\{\sup_{s \in \mathbf{\Theta}} d_K(F_{\theta^*_{ap,n}(s)}, F_s), \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}}\}. \tag{40}$$

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

$$\tilde{\epsilon}^*_n \sim \frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} \sim w(a_n). \tag{41}$$

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 (41).*
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 7.3 $p_{\epsilon,match}(\theta^*)$ *in (10) 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 7.4 *MMSPE is a relative of ABC MLE (Dean* et. al., *2014, Yildirim* et. al. *2015) where an ϵ-neighborhood like that in (9) 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*

$$\int_{\{\theta: d_{\mathbf{\Theta}}(d,\theta)\leq\epsilon/\sqrt{n}\}} f(\mathbf{x}|\theta)d\theta, \tag{42}$$

(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: (42) is not a probability, it is defined via a neighborhood in $\mathbf{\Theta}$ and does not have the frequentist interpretation (10) of $p_{\epsilon,match}(\theta^*)$ for a particular θ^*.

Remark 7.5 *Rates (23), (24), (35), (36) and (41) 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).*

8 Appendix

Proposition 8.1 *(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 $\epsilon > 0$,*

$$P[d_K(\hat{F}_{\mathbf{Y}}, F) > \epsilon] \leq U_{DKWM} = 2e^{-2n\epsilon^2} \tag{43}$$

Lemma 8.1 *Let $\mathbf{X}$ be a sample of i.i.d. F_θ r.vs, with $\theta \in \mathbf{\Theta} = \mathbf{\Theta_n^*} = \{\theta_1^*, \ldots, \theta_{N_n}^*\}$. For any $\zeta > 0$ it holds for $\hat{\theta}_{MMDE}$ in (7),*

$$P[d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) > \zeta] \leq 2 \cdot N_n \cdot e^{-2n\zeta^2}. \tag{44}$$

When $\zeta = \frac{\sqrt{\ln N_n}}{\sqrt{n}}$, *the upper bound in (44) is* $\frac{2}{N_n}$ *and converges to zero as* N_n *increases to infinity with* n.

Proof of Lemma 8.1:

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

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

with the last inequality by Proposition 8.1. When $\zeta = \frac{\sqrt{\ln N_n}}{\sqrt{n}}$ the upper bound is $\frac{2}{N_n}$. □

Remark 8.1 *Extensions of Proposition 8.1 in* $R^d, d > 1$, *appeared at least by Kiefer and Wolfowitz (1958), Kiefer (1961) and Devroye (1977) with corresponding upper bounds* U *in (43):* $U_{KW} = C_1(d)e^{-C_2(d)n\epsilon^2}$, $U_K = C_3(b,d)e^{-(2-b)n\epsilon^2}$ *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 8.1 holds in* R^d *at least when using* U_{KW} *and different constants.*

Proof of Lemma 5.1: The first and the last term in upper bound (14) have uniform upper bounds in Probability with order, respectively, $\frac{\sqrt{\ln N_n}}{\sqrt{n}}$ (from Lemma 8.1) and $\frac{k_n}{\sqrt{n}}, k_n = o(\sqrt{n})$ from (43); choose $k_n \sim \sqrt{\ln N_n}$. □

Proof of Proposition 7.1: a) From (7), with $\mathbf{\Theta_n^*}$ instead of $\mathbf{\Theta}$, the "matching term"

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

$$\leq d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n. \quad (45)$$

From (8) and (45),

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

$$\leq d_K(F_{\hat{\theta}_{MMDE}}, \hat{F}_{\mathbf{X}^*(\hat{\theta}_{MMDE})}) + d_K(\hat{F}_{\mathbf{X}^*(\theta^*_{ap,n}(\theta))}, F_{\theta^*_{ap,n}(\theta)}) + d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) + 2d_K(F_\theta, \hat{F}_{\mathbf{X}(\theta)}) + \gamma_n. \quad (46)$$

Using (46), Lemma 8.1, the Dvoretzky-Kiefer-Wilfowitz-Massart inequality (43) and

$$\tilde{\epsilon} = \epsilon_n - d_K(F_{\theta^*_{ap,n}(\theta)}, F_\theta) - \gamma_n, \quad (47)$$

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

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

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

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

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

$$\leq 6 \cdot N(a_n) \cdot e^{-n\tilde{\epsilon}^2/18}. \tag{48}$$

From (21) and (47),

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

and upper bound (48) 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) (22) follows from (21) 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 (21)

$$\epsilon_n \leq C \cdot w(a_n) + 6\frac{\sqrt{\ln N(a_n)}}{\sqrt{n}} + \gamma_n, \ 1 \leq C, \tag{49}$$

and the uniform upper rate of convergence (23) follows ignoring γ_n.

b_3) Follows from (23) and the properties of w.

c) For b_2), u_n follows from (49) with $a_n = w^{-1}(n^{-1/2})$ and ($\mathcal{A}3$) implies the rate for b_3). □

Proof of Corollary 7.1: (24) follows from (23). 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 (24) holds, with $k = k(n), n \geq n^*$. Rates follow taking $a_n = w^{-1}(n^{-1/2})$ as in Proposition 7.1, b_3), c), replacing N by N_k. □

Proof of Proposition 7.2: 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 $F_s, \delta \in R$,

$$\lim_{m\to\infty} P_s[\sqrt{m}(d_K(\hat{F}_{\mathbf{X}^*(s)}, F_\theta) - \Delta(s,\theta) \leq \delta] = \Phi(\frac{\delta}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}). \tag{50}$$

When $\delta > 0$,

$$\Phi(\frac{\delta}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}) \geq \Phi(2 \cdot \delta). \tag{51}$$

From (50), for the given ϵ, θ and large m,

$$P_s[d_K(\hat{F}_{\mathbf{X}^*(s)}, F_\theta) \leq \epsilon] \approx \Phi(\frac{\sqrt{m}(\epsilon - \Delta(s,\theta))}{\sqrt{F_s(x_{s,\theta})(1 - F_s(x_{s,\theta})}}), \tag{52}$$

with "$\approx$" denoting asymptotic equality.

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

$$\Delta(s^*, \theta) \leq \epsilon - \frac{k_n^*}{\sqrt{n}}, \ k_n^* = o(\sqrt{n}), \ k_n^* \uparrow \infty \text{ with } n. \tag{53}$$

For $s = s^*, m = n$ in (52) and from (51),

$$P_{s^*}[d_K(\hat{F}_{\mathbf{X}^*(s^*)}, F_\theta) \leq \epsilon] \approx \Phi(\frac{\sqrt{n} \cdot (\epsilon - \Delta(s^*, \theta))}{\sqrt{F_{s^*}(x_{s^*,\theta})(1 - F_{s^*}(x_{s^*,\theta}))}} \geq \Phi(2 \cdot \sqrt{n} \cdot (\epsilon - \Delta(s^*, \theta)) \geq \Phi(2 \cdot k_n^*). \quad \square. \tag{54}$$

Proof of Proposition 7.3: a_1) $\tilde{\epsilon}_n^*$ follows from (15), with $\epsilon = \epsilon_n$ in (33), $N_n = N(a_n)$. 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 (35) follows from (25) 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 (33). Then, for large n, (53) (same with (31)) holds with $s^* = \theta^*_{ap,n}(\theta)$ and $k_n^* = .5 \cdot \sqrt{\ln N(a_n)}$. Hence, from (54) 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 (12), (13).
a_3) Follows from ($\mathcal{A}2$), ($\mathcal{A}3$), (35) 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 (25), 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 (35) and (36) 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 $|\hat{\theta}_{MMSPE} - \theta|$. $\square$

Proof of Proposition 7.4: Since

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

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

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

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

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

with the last inequality obtained using U_{KW} in the upper bound (43) as suggested in Remark 8.1. (37) and (38) follow. $\Box$

Proof of Proposition 7.5: *a*) $\tilde{\epsilon}^*_n$ follows from (15), with $\epsilon = \epsilon_n$ in (39), $N_n = N(a_n)$.
b) Follows from assumptions $(\mathcal{A}2), (\mathcal{A}3)$, (37), (38) The result for $\hat{\theta}_{MMSPE}$ follows from its construction and (12), (13).
c) Follows from $(\mathcal{A}2), (\mathcal{A}3)$, (41) and the properties of w.
d) For *b*), u_n follows from (41) with $a_n = w^{-1}(n^{-1/2})$ and $(\mathcal{A}3)$ implies the rate for *c*). $\Box$

References

[1] Bernton, E. , Jacob, P. E., Gerber, M. and Robert, C. P. (2019a) Approximate Bayesian computation with the Wasserstein distance.*JRSS B*, **81**, 235-269. arXiv:1905.03747v1

[2] Bernton, E. , Jacob, P. E., Gerbery, M. and Robert, C. P. (2019b) On parameter estimation with the Wasserstein distance. Information and Inference: A Journal of the IMA (2019) Page 1 of 23. *Information and Inference: A Journal of the IMA* **8**, 657-676.

[3] Briol, F.-X., Barp, A., Duncan, A. B. and Girolami, M. (2019) Statistical Inference for Generative Models with Maximum Mean Discrepancy. arXiv:1906.05944v1 [stat.ME] 13 Jun 2019

[4] Chen, L. and Wu, W. B. (2018) Concentration inequalities for empirical processes of linear time series. *J. Machine Learning Research* **18**, 1-46.

[5] Cochran, W. G. and Rubin, D. B. (1973) Controlling Bias in Observational Studies: A Review, *Sankhya,Ser. A*, **35**, 417-446.

[6] Dean, T. A., Singh, S. S., Jasra, A, and Peters, G. W. (2014) Parameter estimation for hidden Markov models with intractable Likelihoods. *Scandinavian J. of Statistics*, **41**, 970-987.

[7] Devroye, L. P. (1977) A Uniform Bound for the Deviation of Empirical Distribution Functions. *J. Multiv. Anal.* **7**, 594-597.

[8] Dvoretzky, A., Kiefer, J. and and Wolfowitz, J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27**, 642-669.

[9] Haynes M.A., MacGillivray H.L., and Mengersen K.L. (1997) Robustness of ranking and selection rules using generalized g-and-k distributions. *J. Statist. Plan. and Infer.* **65**, 45-66.

[10] Jasra, A., Singh, S. S., Martin, J. S. and McCoy, E. (2012). Filtering via Approximate Bayesian Computation. *Stat. Comput.* **22**, 1223-1237.

[11] Kajihara, T. , Kanagawa, M., Yamazaki, K. and Fukumizu, K. (2018) Kernel Recursive ABC: Point Estimation with Intractable Likelihood. arXiv:1802.08404v2

[12] Kiefer, J. (1961) On Large Deviations of the Empiric D. F. of Vector Chance Variables and a Law of the Iterated logarithm. *Pacific J. of Mathematics* **11**, 649-660

[13] Kiefer, J. and Wolfowitz, J. (1958) On the deviations of the empiric distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87**, 173-186

[14] Le Cam, L. M. and Yang, G. L. (2000) *Asymptotics in Statistics: Some Basic Concepts.* Springer-Verlag, New York.

[15] Le Cam, L. M. (1986) *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[16] Massart, P. (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Prob.* **18**, 1269-1283

[17] Peacock, J. A. (1983) Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices Royal Astronomy Society* **202**, 615–627.

[18] Polonik, W. (1999) Concentration and goodness-of-fit in higher dimensions:(asymptotically) distribution-free methods. *Ann. Statist.*, **27**, 1210–1229.

[19] Raghavachari, M. (1973) Limiting distributions of Kolmogorov-Smirnov type statistics under the Alternative. *Ann. Stat.* **1**, 67-73.

[20] Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J. and Mykytka, E. F. (1979) A probability distribution and its uses in fitting data. *Technometrics* **21**, 201–214.

[21] Rayner, G. D. and MacGillivray, H. L. (2002) Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing* **12**, 57-75.

[22] Roussas, G. G.and Yatracos, Y. G. (1997) Minimum distance estimates with rates under ϕ-mixing. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics,* p. 337-345. Editors: D. Pollard, E. Torgersen, G. L. Yang. Springer, New York.

[23] Rubin, D. B. (2019) Conditional Calibration and the Sage Statistician. *Survey Methodology* **45**, 187-198.

[24] Rubin, D. B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Ann. Statist.* **12**, pp. 213-244.

[25] Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159-183. Correction (1974) **30**, 728.

[26] Tukey, J. W. (1977) Modern techniques in data analysis. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.

[27] Weiss, L. (1983) Small-Sample Properties of Maximum Probability Estimators. *Stoch. Proc. and Appl.* **14**, 267-277.

[28] Weiss, L. and Wolfowitz, J. (1974) *Maximum Probability Estimators and Related Topics.* Lecture Notes in Mathematics, Vol. 424, Springer-Verlag.

[29] Weiss, L. and Wolfowitz, J. (1967) Maximum probability estimators. *Ann. Inst. Stat. Math.* **19**, Article 193.

[30] Wolfowitz, J. (1957) The Minimum Distance Method. *Ann. Math. Statist.* **28** , 75-88.

[31] Yatracos, Y. G. (2020) Fiducial Matching for the Approximate Posterior: F-ABC. Submitted for publication.

[32] Yatracos, Y. G. (2019) Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture. *Statistics* **53**, 1251-1268.

[33] Yatracos, Y. G. (1989) A regression type problem. *Ann. Statist.* **17**, 1597-1607.

[34] Yatracos, Y. G. (1985) Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13**, 768-774.

[35] Yildirim, S., Singh, S. S., Dean, T. and Jasra, A. (2015) Parameter Estimation in Hidden Markov Models With Intractable Likelihoods Using Sequential Monte Carlo *J. Comp. Graph. Stat.* **24**, 846-865.

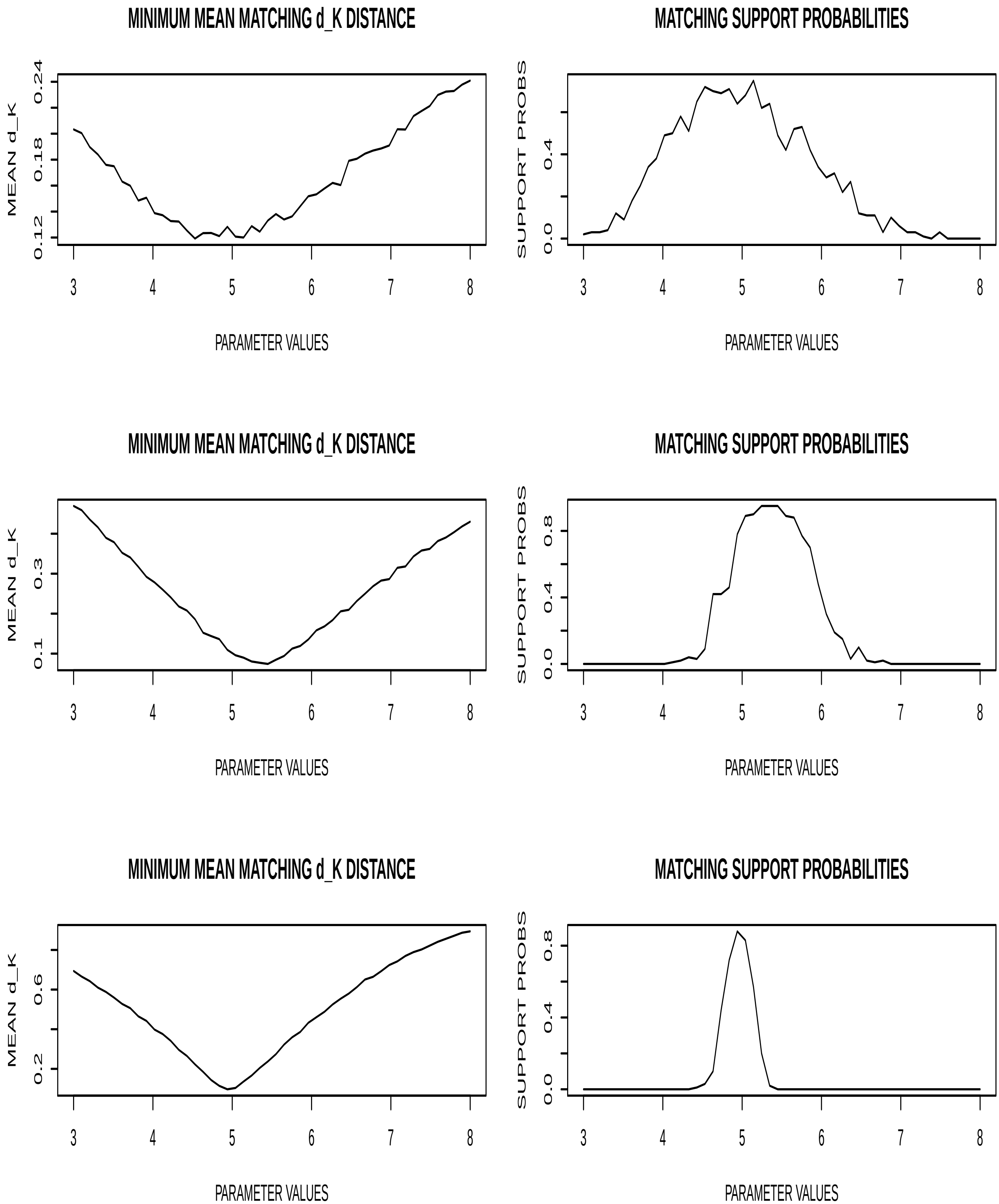

Figure 1: Row-wise, Exponential, Poisson with parameters 5, Normal mean 5, known $\sigma = 1$. Plots along $\mathbf{\Theta}$ with extremes pointing to the parameters.

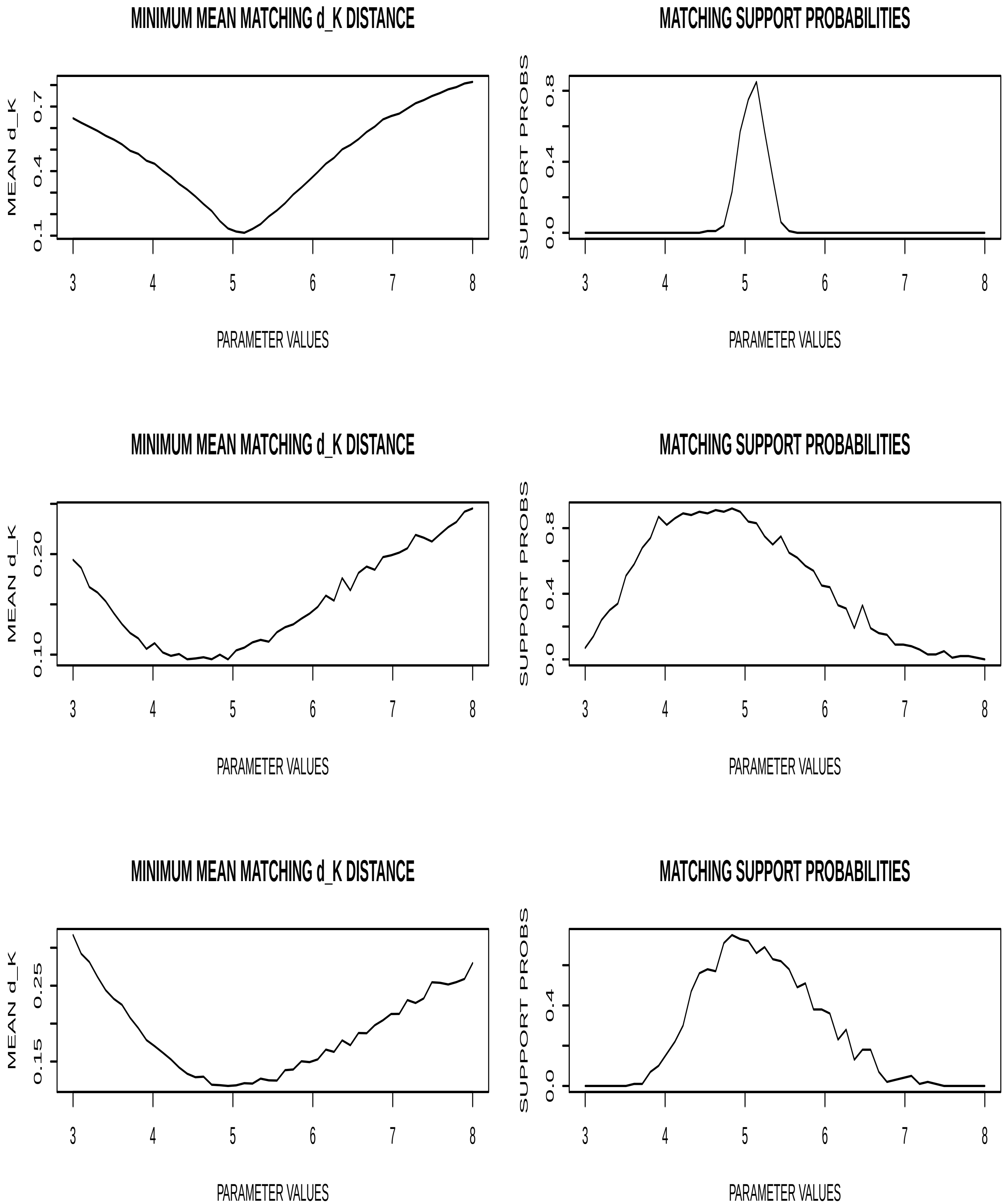

Figure 2: Row-wise, Weibull, Cauchy, Normal Both Parameters 5. Plots along $\boldsymbol{\Theta}$ with extremes pointing to the parameters.

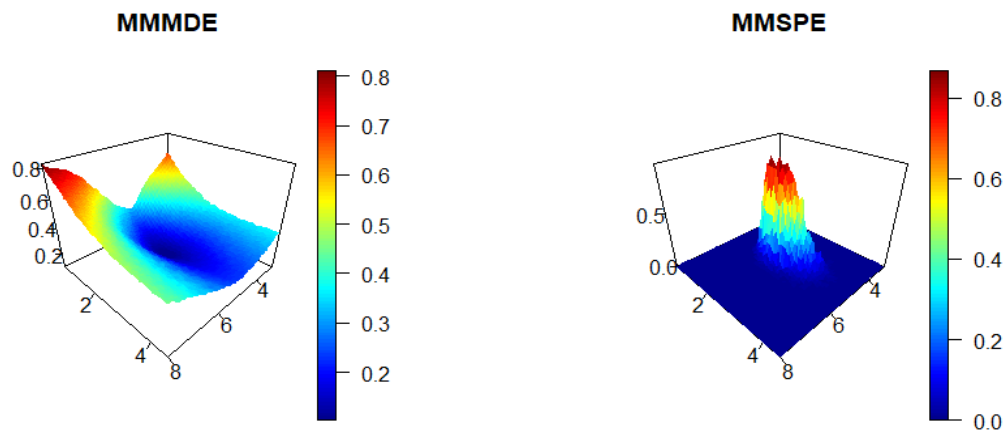

Figure 3: Parameter space $\mathbf{\Theta} = [3, 8]x[0.5, 4.5]$, Model Parameter $\theta = (\mu = 5, \sigma = 2)$. Plot along $\mathbf{\Theta}$ with extremes pointing to the parameters.

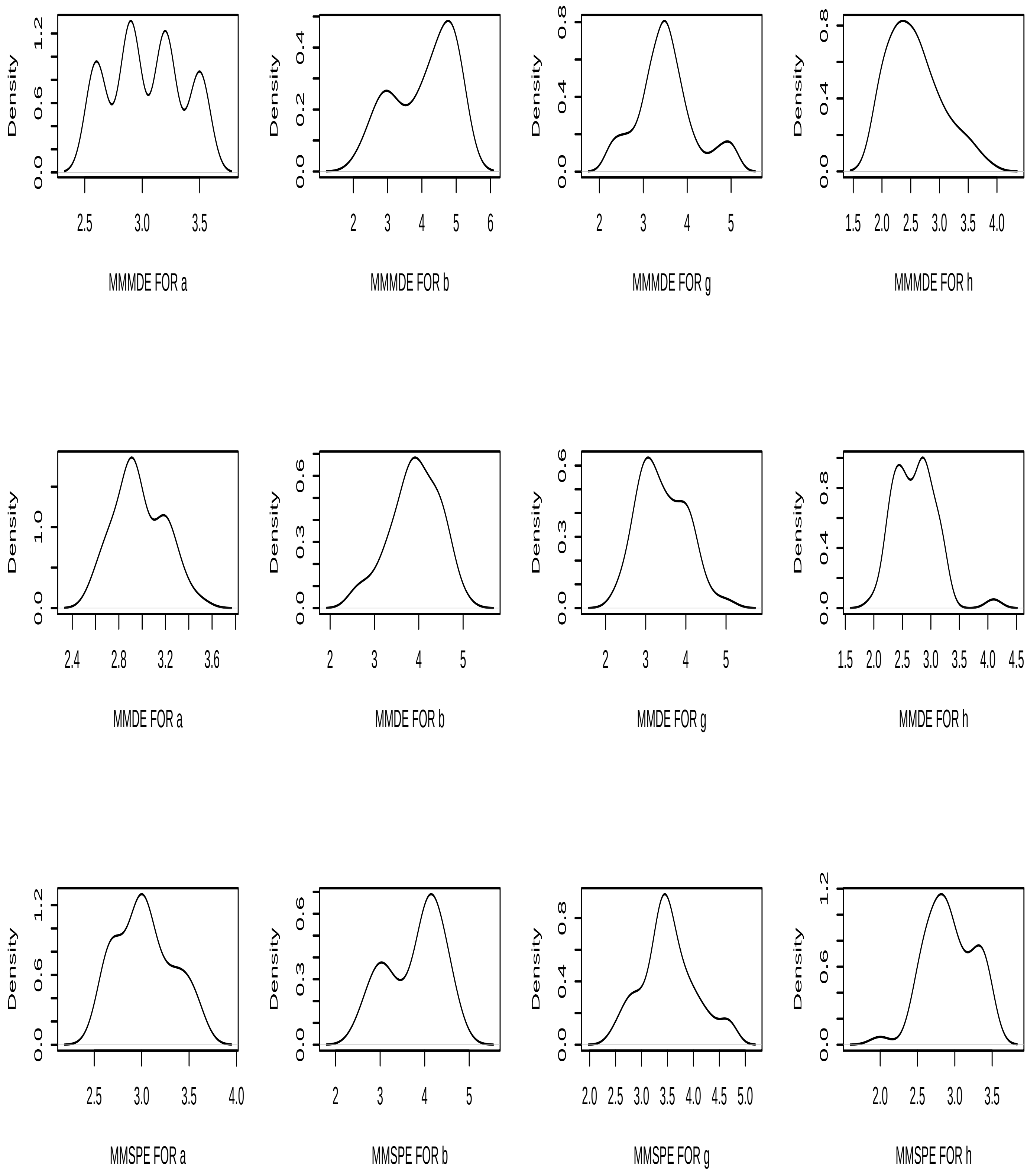

Figure 4: Density plots for the 50 estimates of Tukey's g-and-h model with independent samples, $n = 200$. The parameters are $a = 3, b = 4, g = 3.5, h = 2.5$.

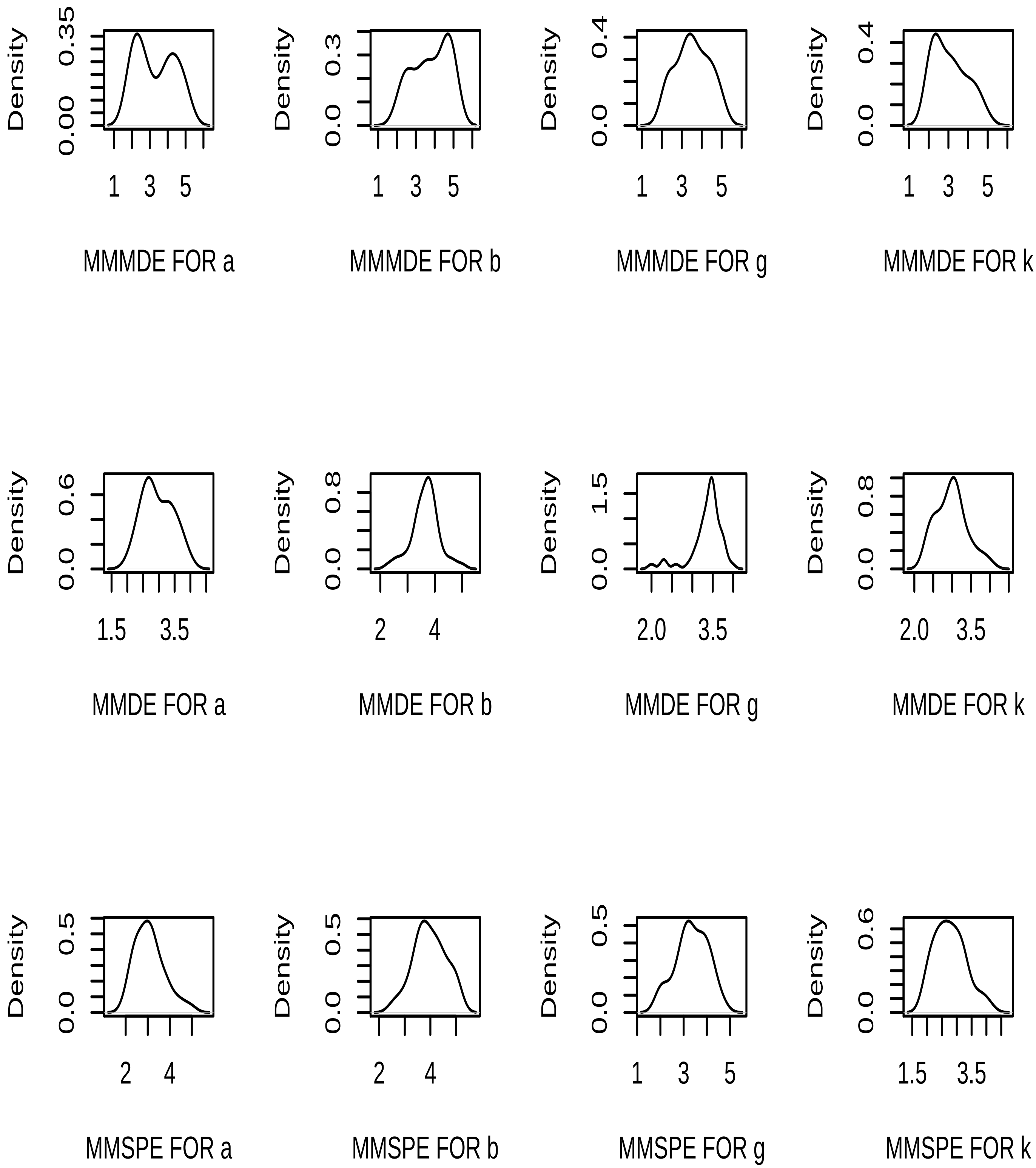

Figure 5: Density plots for 50 estimates of g-and-k model with dependent samples, $n = 50$. The parameters are $a = 3, b = 4, g = 3.5, k = 2.5$

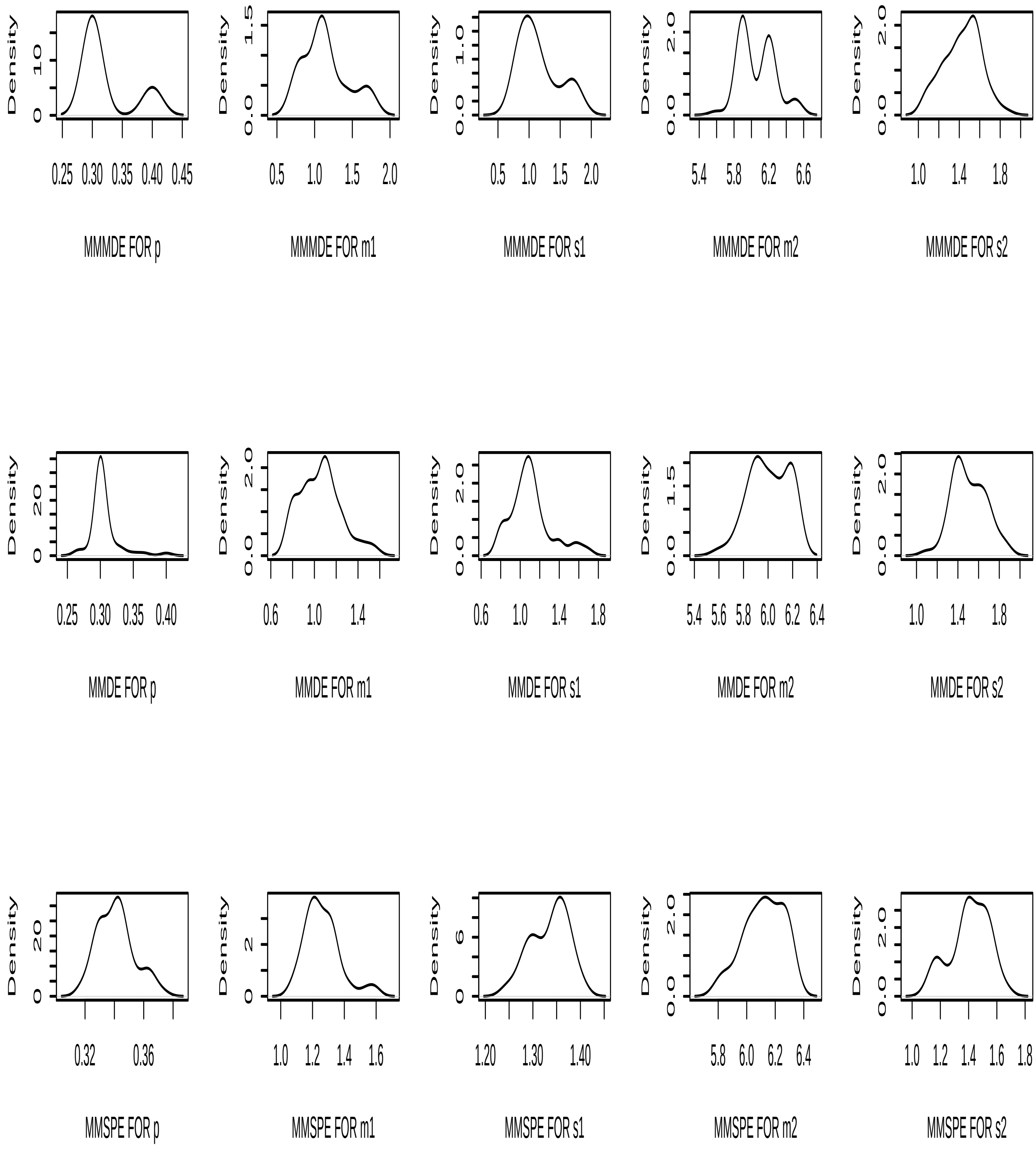

Figure 6: Density plots for the 50 estimates of the normal mixture with independent samples, $n = 200$; the parameters are p=.3, μ_1=m1=1, σ_1=s1=1, μ_2=m2=6, σ_2=s2=1.5.