Far East Journal of Theoretical Statistics
© 2019 Pushpa Publishing House, Prayagraj, India
http://www.pphmj.com
http://dx.doi.org/10.17654/TS057020143
Volume 57, Number 2, 2019, Pages 143-170 **ISSN: 0972-0863**

MODEL OF MODELS: A NEW PERSPECTIVE TO DEAL WITH MODEL UNCERTAINTY

Silvia Figini[1], Pierpaolo Uberti[2] and Maria Laura Torrente[2]

[1]Department of Political and Social Sciences
University of Pavia and RIDS
Italy
e-mail: silvia.figini@unipv.it

[2]DIEC Department of Economics
University of Genova
Italy
e-mail: uberti@economia.unige.it
marialaura.torrente@economia.unige.it

Abstract

This paper presents a novel methodological approach called the Model of Models (MoM). MoM concerns the selection of the best model for a given partition of the data derived from the realization of the independent variables. Compared to ensemble techniques and model averaging approaches proposed in the literature, MoM does not require a selection of which models to include in the pool of models and it works without resorting to the combination of model predictions.

MoM works on parametric and non parametric predictive models as well as any other dependent or independent variables. In the case of a partition of the data, the theoretical proposal derives the properties of MoM. The implementation of MoM, when no partition of the data is

Received: September 16, 2019; Accepted: October 16, 2019
2010 Mathematics Subject Classification: 00A71, 65C20.
Keywords and phrases: model selection, model choice, model averaging, model uncertainty.

available in advance, is performed using a new algorithm termed as MoMa.

In order to show how MoM works, empirical evidence is provided on simulated data sets.

The proved theoretical results coupled with the empirical evidence gathered from simulated data demonstrate that MoM is a good strategy to deal with model choice and model uncertainty.

1. Introduction

In recent years, a number of multi-model methods have been proposed to account for uncertainties arising from input parameters and the definition of model structure.

As described in Singh et al. [20], the different sources of uncertainty in the modeling process can be categorized as: conceptual uncertainty (i.e., the conceptual model of the underlying system), parametric uncertainty (i.e., uncertainty linked to parameters and absence of data) and stochastic uncertainty (i.e., uncertainty in predictions).

In general, the methods proposed in the literature believe that it is more appropriate to consider multi-model predictive uncertainty than to rely on a single conceptual model. Traditional estimation procedure generally begins with model selection (see e.g., Lin et al. [15] and Klebanov et al. [14]). Once a specific model has been selected, subsequent estimation is conducted using the selected model without taking into consideration the uncertainty from the selection process.

Model averaging estimation which incorporates model uncertainty into the estimation process (see e.g., Ranjan and Gneiting [18]) is an alternative to this procedure. In recent years, there has been rising interest in model averaging from the frequentist (see e.g., Wang et al. [22], Ando and Li [1], Ando and Li [2], Zhang et al. [23]), Bayesian (see e.g. Hoeting et al. [11], Raftery et al. [19]) and ensemble machine learning (see e.g., Breiman [3-5] and, Omer and Lior [17]) perspectives, and some important progress has been made.

Compared to the Frequentist Model Averaging (FMA) approach, there has been an enormous amount of literature on the use of the Bayesian Model Averaging (BMA) approach where the uncertainty of a model is considered by setting a prior probability to each candidate model.

As pointed out by Fragoso et al. [8] the application of BMA is not always straightforward, which could lead to diverse assumptions and situational choices according to its different aspects.

In contrast, the FMA approach requires no priors and the corresponding estimators are entirely determined starting from the data. For this reason, the FMA approach has received much attention over the last decade (see e.g., Hjort and Claeskens [12] and Hjort and Claeskens [13]). The performance of the FMA procedures largely depends on how to choose weights in estimation. Consequently, much of the work focuses on weight choice to achieve stable prediction.

A different strategy to deal with model uncertainty comes from the pooling approach introduced by Stone [21]. Combining predictions from alternative models often improves those forecasts based on a single best model (see e.g., Geweke and Amisano [9]). Furthermore, when single models are subject to structural breaks and miss-identification errors, a pool approach based on many alternative models is expected to outperform methods that try to select the best forecasting model (see e.g., Geweke and Amisano [9], Figini et al. [7], Lv and Liu [16]).

In this paper we propose a completely different approach to deal with model uncertainty and model choice called Model of Models (MoM): MoM concerns the selection of the best model for a given partition of the data derived from the realization of the independent variables.

MoM does not require selecting which models to include in the pool of models and it works without resorting to the combination of model predictions. For this reason MoM can be classified as an objective approach to deal with model selection and model uncertainty. Broadly speaking, for each element of a given partition of the independent variables MoM selects the best model from a model set. The model set is composed of parametric and non parametric predictive models. The competing models in the model set are estimated in advance from the whole data set while the partition of the independent variables is derived independently following the model estimation step. This overcomes the potential over-fitting issues. The results achieved using MoM hold for any partition of the independent variables.

In the second part of the paper, the properties of MoM are derived and proved. Our idea is supported by a strong theoretical framework which is presented in Section 2; Section 3 shows the computational aspects to implement MoM and introduces the algorithm; Section 4 reports the empirical evidence at hand obtained on simulated data. Discussion of the theoretical and computational results is summarized in Section 5.

2. MoM: Theoretical Proposal

MoM is a new approach to deal with model selection and model uncertainty in predictive modeling. In this section we prove that single model selection for each partition of data provides better results in terms of fitting. MoM works with parametric, semi-parametric and non parametric predictive models characterized by quantitative or qualitative dependent and independent variables.

In order to formalize MoM, let $x_1, ..., x_n$ be n independent variables taking values in the real intervals $A_1, ..., A_n$. Let $A = A_1 \times \cdots \times A_n \subseteq \mathbb{R}^n$, let $\mathbb{X} = \{p_1, ..., p_s\} \subset A$ be a set of s input data and $\boldsymbol{y} = (y_1, ..., y_s) \in \mathbb{R}^s$ be the vector of s realizations of the dependent variable.

Let $m \geq 1$ be an integer number and let $f_1, ..., f_m$ be real functions defined over $A \subseteq \mathbb{R}^n$. Each function f_j, $j = 1, ..., m$, is a model that relates the input data $\mathbb{X} \subset A$ to the realizations $\boldsymbol{y}$, and each vector $\hat{\boldsymbol{y}}_{\boldsymbol{j}} = f_j(\mathbb{X}) = (f_j(p_1), ..., f_j(p_s)) \in \mathbb{R}^s$, $j = 1, ..., m$, is the vector of predicted values.

The functions $f_j,\ j = 1, ..., m,$ constitute the model set and as pointed out in Section 1, are assumed to be given. The models $f_j,\ j = 1, ..., m,$ can differ both for the functional form and/or for the subset of the independent variables $x_1, ..., x_n$ used as explanatory variables. We do not assume any further restriction on the models; consequently, we can consider input models with completely different functional forms as well as input models which depend on an increasing number of parameters, as in the classical case of *nested models* as described in Definition 2.1.

Definition 2.1 (Nested Models)**.** Let $f_1, ..., f_m$ be real functions defined over $A \subseteq \mathbb{R}^n$ and belonging to the families $\mathcal{F}_1, ..., \mathcal{F}_m$. If $f_j \in \mathcal{F}_{j+1},$ for each $j = 1, ..., m - 1,$ the models are said to be *nested*.

As pointed out in Section 1, in order to deal with model uncertainty, different approaches of *model average* are proposed in the literature and the final results become a linear combination of the models under comparison (see e.g. Hoeting et al. [11], in the Bayesian framework).

Definition 2.2 (Model Average)**.** Let $f_1, ..., f_m$ be real functions defined over $A \subseteq \mathbb{R}^n,$ and let $c_1, ..., c_m$ be positive real numbers such that $\sum_{j=1}^{m} c_j = 1;$ denote $(c_1, ..., c_m) = \boldsymbol{c}.$ The *model average* $f_{\boldsymbol{c}}$ is the linear combination of the models $f_1, ..., f_m$ over *A* defined by:

$$f_{\boldsymbol{c}} = \sum_{j=1}^{m} c_j f_j. \tag{1}$$

The selection of $(c_1, ..., c_m)$ is crucial in model averaging approaches. In Bayesian Model Averaging $(c_1, ..., c_m)$ is replaced by the posterior probability of each model selected in the model space, but in general the choice of $(c_1, ..., c_m)$ is an open point of research.

We noticed that by choosing different vectors of the weights $\boldsymbol{c}$ it is possible to obtain classical model averaging, Bayesian model averaging, ensemble models and, in particular, when $\boldsymbol{c}$ has one unitary entry and $m-1$ null entries, the special case of model selection, in which one of the available models is chosen as the best model based on some given criterion.

An important task in model selection is to derive measures of goodness for a model. A simple approach is to evaluate the error between the measured and the predicted values of the dependent variable (i.e., the difference between the vectors $\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = f(\mathbb{X})$ respectively). This distance function on the vector space $\mathbb{R}^s$ is useful to compare different models to predict $\boldsymbol{y}$. A natural choice for the error is given by distance functions on the vector space $\mathbb{R}^s$. A special case of distance functions is the Brier score, see Brier [6], which measures the model's overall performance by taking into account the calibration and discrimination of each model.

We recall that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^s$ a distance $d_s(\boldsymbol{x}, \boldsymbol{y})$ satisfies the following conditions:

(1) $d_s(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^s$;

(2) $d_s(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$;

(3) $d_s(\boldsymbol{x}, \boldsymbol{y}) = d_s(\boldsymbol{y}, \boldsymbol{x}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^s$;

(4) $d_s(\boldsymbol{x}, \boldsymbol{z}) \leq d_s(\boldsymbol{x}, \boldsymbol{y}) + d_s(\boldsymbol{y}, \boldsymbol{z}) \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^s$.

In this paper we consider distance functions d_s on $\mathbb{R}^s$ of the form $d_s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{s} d(x_i, y_i)$, with d a distance function on $\mathbb{R}$. Note that this choice includes special cases such as the Laplace's distance $\sum_{i=1}^{s} | x_i - y_i |$, the Euclidean distance $\sum_{i=1}^{s} (x_i - y_i)^2$, and distance functions induced by the power of the classical p-norm on $\mathbb{R}^s$, that is, $\sum_{i=1}^{s} | x_i - y_i |^p$, with p being a positive integer.

Definition 2.3. Let $A \subseteq \mathbb{R}^n$, let $\mathbb{X} = \{p_1, ..., p_s\}$ be a set of points of A and $\boldsymbol{y} = (y_1, ..., y_s) \in \mathbb{R}^s$ be the vector of s realizations of the dependent variable. Let f_1, f_2 be real functions defined over A.

(a) If

$$\sum_{i=1}^{s} d(y_i, f_1(p_i)) = \sum_{i=1}^{s} d(y_i, f_2(p_i))$$

then model f_1 is *equivalent* to model f_2, also denoted by $f_1 \equiv f_2$.

(b) If

$$\sum_{i=1}^{s} d(y_i, f_1(p_i)) \leq \sum_{i=1}^{s} d(y_i, f_2(p_i))$$

then model f_1 is *no worse* than model f_2.

Definition 2.3 provides a guideline to compare models in terms of fitting; as a result, the best model is selected.

If different models for the data set $\mathbb{X}$ and realizations $\boldsymbol{y}$ show different local fitting,[1] selecting the best model by using some given performance criterion does not represent the best solution. On the other hand, choosing the best model in each element of a given partition of the data, thus reflecting the realization of the independent variable, could be considered an alternative approach to model selection.

In this paper we prove that MoM has good and desirable properties in terms of fitting, and, in particular, it is no worse than classical model selection procedures. In other words, we suggest not choosing one model among other available models or combining the latter in some optimal average, but rather to use one specific model depending on the realizations of the independent variables.

[1] We do not formally define the concept of local fitting of a model because the intuition to restrict the fitting performance analysis of a model to some subset of the domain is sufficient for the clarity of the paper.

In order to define MoM, the concept of restricted partition of A is required.

Definition 2.4 (Restricted Partition)**.** Let $A \subseteq \mathbb{R}^n$ and $\mathbb{X} = \{p_1, \dots, p_s\}$ be a set of points of A. A *partition* of A is a family $U = \{U_1, \dots, U_r\}$ of sets such that:

(1) $\varnothing \neq U_k \subseteq A \quad \forall k = 1, \dots, r;$

(2) $U_k \cap U_j = \varnothing \quad \forall k, j = 1, \dots, r,$ with $k \neq j;$

(3) $\bigcup_{k=1}^{r} U_k = A.$

Furthermore, if the following additional condition holds:

(4) $U_k \cap \mathbb{X} \neq \varnothing, \quad \forall k = 1, \dots, r,$

then U is called a $\mathbb{X}$-*restricted partition* of A or *restricted partition* of A.

Definition 2.4 differs from the standard definition of partition of a set for 4. The formal definition of MoM is based on restricted partitions of A, i.e., partitions made up of subsets which are not disjoint from the given data set $\mathbb{X} = \{p_1, \dots, p_s\} \subset A.$

Definition 2.4 compares the input models $f_1, \dots, f_m$ based on the distance function $\sum_{i=1}^{s} d(y_i, \hat{y}_i)$ restricted to each subset U_k, $k = 1, \dots, r$, of the restricted partition $U = \{U_1, \dots, U_r\}$.

In order to derive MoM, Definition 2.5 described below, shows how to manage $f_1, \dots, f_m$.

Definition 2.5 (Model of Models - MoM)**.** Let $A \subseteq \mathbb{R}^n$ and $\mathbb{X} = \{p_1, \dots, p_s\}$ be a set of points of A. Let $f_1, \dots, f_m$ be real functions defined over $A \subseteq \mathbb{R}^n$ and let $U = \{U_1, \dots, U_r\}$ be a restricted partition of A (see Definition 2.4). The (MoM) f_U is a real function over A defined by:

$$f_U(\boldsymbol{x}) = \begin{cases} f_{\alpha(U_1)}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_1 \\ \vdots & \vdots \\ f_{\alpha(U_r)}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_r, \end{cases} \tag{2}$$

where $\alpha(U_k) \in \{1, \ldots, m\}$ is:

$$\alpha(U_k) = \arg\min_{j=1,\ldots,m} \left\{ \sum_{p_i \in U_k} d(y_i, f_j(p_i)) \right\} \tag{3}$$

for each $k = 1, \ldots, r$.

In Proposition 2.6 we prove that for any restricted partition U of A the MoM is no worse than the original models $f_1, \ldots, f_m$ (see Definition 2.4).

Proposition 2.6. *Let* $A \subseteq \mathbb{R}^n$, *let* $\mathbb{X} = \{p_1, \ldots, p_s\} \subset A$ *be a set of s input data and* $\boldsymbol{y} = (y_1, \ldots, y_s) \in \mathbb{R}^s$ *be the vector of the realizations of the dependent variable. Let U be a restricted partition of A* (*see Definition* 2.4). *Let* $f_1, \ldots, f_m$, *with* $m \geq 1$, *be real functions defined over A and* f_U *be the MoM* (*see Definition* 2.5). *Then*, *the model* f_U *is no worse than each model* f_j, $j = 1, \ldots, m$.

Proof. Let $U = \{U_1, \ldots, U_r\}$ and let $\alpha(U_1), \ldots, \alpha(U_r)$ be the indexes defined by formula (3). By using the properties of the restricted partition U (see Definition 2.4), the expression $\sum_{i=1}^{s} d(y_i, f_U(p_i))$ can be rewritten as follows:

$$\sum_{i=1}^{s} d(y_i, f_U(p_i)) = \sum_{k=1}^{r} \sum_{p_i \in U_k} d(y_i, f_U(p_i))$$

$$= \sum_{k=1}^{r} \sum_{p_i \in U_k} d(y_i, f_{\alpha(U_k)}(p_i)). \tag{4}$$

For each $j = 1, \ldots, m,$ by formula (3), it follows that:

$$\sum_{p_i \in U_k} d(y_i, f_{\alpha(U_k)}(p_i)) \leq \sum_{p_i \in U_k} d(y_i, f_j(p_i)). \tag{5}$$

Combining (4) and (5) we get

$$\sum_{i=1}^{s} d(y_i, f_U(p_i)) \leq \sum_{k=1}^{r} \sum_{p_i \in U_k} d(y_i, f_j(p_i))$$

$$= \sum_{i=1}^{s} d(y_i, f_j(p_i)).$$

Therefore, by using Definition 2.3-(b) the Proposition is proved.

We observed that different restricted partitions U and V of A could lead to equivalent models, that is $f_U \equiv f_V$ (see Definition 2.3-(a)). This happens when U and V have the same number r of subsets and $U_k \cap \mathbb{X} = V_k \cap \mathbb{X} \neq \varnothing$ holds for each $k = 1, \ldots, r.$

Furthermore, we remark that Proposition 2.6 holds whatever restricted partition U is chosen. Despite this evidence related to the fitting performances, it is clear that the MoM f_U depends on the choice of the restricted partition U but, in practice, some partitions of the data will perform better than others. This leads us to introduce the concept of refinement of a partition as described in Definition 2.7.

Definition 2.7 (Refinement of a Partition)**.** Let U and V be two different restricted partitions of A (see Definition 2.4). The partition U is a *refinement* of V, denoted by $U \leq V,$ if every element of U is a subset of an element of V.

When we compare the MoM corresponding to the two restricted partitions U and V, with $U \leq V,$ the model f_U corresponding to the finer partition is no worse than the model $f_V,$ as proved in Proposition 2.8. Proposition 2.8 links the selection of the partition with the fitting performance of the MoM.

Proposition 2.8. *Let* $A \subseteq \mathbb{R}^n$, *let* $\mathbb{X} = \{p_1, ..., p_s\} \subset A$ *be a set of s input data and* $\boldsymbol{y} = (y_1, ..., y_s) \in \mathbb{R}^s$ *be the vector of realizations of the dependent variable. Let U and V be restricted partitions of A* (*see Definition* 2.4) *and suppose that U is a refinement of V. Let* $f_1, ..., f_m$, *with* $m \geq 1$, *be real functions defined over A*, *and* f_U *and* f_V *be MoM* (*see Definition* 2.5). *Then*, *the model* f_U *is no worse than the model* f_V.

Proof. Let $V = \{V_1, ..., V_r\}$ and $U = \{U_1, ..., U_{r'}\}$; let $\alpha(V_1), ..., \alpha(V_r)$ and $\alpha(U_1), ..., \alpha(U_{r'})$ be the indexes defined by formula (3). Using Definition 2.4 and the properties of partition *U*, expression $\sum_{i=1}^{s} d(y_i, f_U(p_i))$ can be rewritten as follows:

$$\sum_{i=1}^{s} d(y_i, f_U(p_i)) = \sum_{k=1}^{r'} \sum_{p_i \in U_k} d(y_i, f_U(p_i))$$

$$= \sum_{k=1}^{r'} \sum_{p_i \in U_k} d(y_i, f_{\alpha(U_k)}(p_i)). \tag{6}$$

Since U is a refinement of V, $r' \geq r \geq 1$ and there exists a partition $I = \{I_1, ..., I_r\}$ of the set $\{1, ..., r'\}$ such that the sets $\{U_j \mid j \in I_h\}$ are subsets of V_h, for each $h = 1, ..., r$. Obviously, for each $h = 1, ..., r$, the family of sets $\{U_j \mid j \in I_h\}$ is a partition of V_h. Consequently, expression (6) can be rewritten as follows:

$$\sum_{i=1}^{s} d(y_i, f_U(p_i)) = \sum_{k=1}^{r'} \sum_{p_i \in U_k} d(y_i, f_{\alpha(U_k)}(p_i))$$

$$= \sum_{h=1}^{r} \sum_{j \in I_h} \sum_{p_i \in U_j} d(y_i, f_{\alpha(U_j)}(p_i)). \tag{7}$$

By formula (3), for each $j \in I_h$, $h = 1, \ldots, r$, it easily follows that:

$$\sum_{p_i \in U_j} d(y_i, f_{\alpha(U_j)}(p_i)) \leq \sum_{p_i \in U_j} d(y_i, f_{\alpha(V_h)}(p_i)). \tag{8}$$

Combining (7) and (8) we obtain

$$\begin{aligned}\sum_{i=1}^{s} d(y_i, f_U(p_i)) &\leq \sum_{h=1}^{r} \sum_{j \in I_h} \sum_{p_i \in U_j} d(y_i, f_{\alpha(V_h)}(p_i)) \\ &= \sum_{h=1}^{r} \sum_{p_i \in V_h} d(y_i, f_{\alpha(V_h)}(p_i)) \\ &= \sum_{i=1}^{s} d(y_i, f_V(p_i)).\end{aligned}$$

Therefore, using Definition 2.3-(b) the proposition is proved.

As a consequence of Proposition 2.8, we can derive the following corollary.

Corollary 2.9. *Assume that the hypotheses of Proposition* 2.8 *hold. Let* $U_t \leq U_{t-1} \leq \cdots \leq U_1$ *be a finite sequence of restricted partition refinements of the set A and* $f_{U_1}, \ldots, f_{U_t}$ *be the corresponding MoMs* (*see Definition* 2.5). *Then, for each* $k = 1, \ldots, t-1,$ *the model* f_{k+1} *is no worse than the model* f_k, *that is*

$$\sum_{i=1}^{s} d(y_i, f_{U_t}) \leq \sum_{i=1}^{s} d(y_i, f_{U_{t-1}}) \leq \cdots \leq \sum_{i=1}^{s} d(y_i, f_{U_1}).$$

Proof. The proof is trivial and directly follows from Proposition 2.8 and Definition 2.3-(b).

Corollary 2.9 shows how the goodness of fit of the corresponding MoM $f_{U_1}, \ldots, f_{U_t}$ weakly increases when the sequence of restricted partitions is composed by successive refinements.

The restricted partitions U of A such that each subset contains a single point of $\mathbb{X}$ generate the MoM f_U with minimum value of the distance $\sum_{i=1}^{s} d(y_i, f_U(p_i))$. We noticed that these special cases are only interesting from a theoretical point of view. In practice, the MoM works better on restricted partitions of A with a small number of elements.

Proposition 2.6 proves that, compared to classical techniques, the MoM generally improves the fitting performances. Unfortunately, this improvement has a drawback in terms of continuity of the MoM. Generally speaking, even if we assume that the input models $f_1, \ldots, f_m$ are continuous functions on A, most times the MoM f_U is *not continuous* on A. There exists at least one point $x_0 \in A$ such that:

$$\lim_{x \to x_0} f_U(x) \neq f_U(x_0).$$

Using Definition 2.4, it is intuitive that the points lying on the border of different subsets of U are points of potential discontinuity for the MoM.

Supposing that the (topological) borders of two subsets of U, say U_i and U_j, with $i \neq j$, are not disjoint, the intersection of the borders of U_i and U_j contains at least one point of A, denoted by x_0. In order to check the continuity/discontinuity of the MoM at x_0, we compute $\lim_{x \to x_0} f_U(x)$. From the definition of the topological border, each neighborhood of x_0 contains both points of U_i and U_j, so we can compute the following two limits:

$$\lim_{x \to x_0 \,|\, x \in U_i} f_U = f_{\alpha(U_i)}(x_0)$$

$$\lim_{x \to x_0 \,|\, x \in U_j} f_U = f_{\alpha(U_j)}(x_0).$$

In general, $f_{\alpha(U_i)}(x_0) \neq f_{\alpha(U_j)}(x_0)$, thus the MoM is not continuous in x_0.

The size of the potential discontinuity points of f_U depends on the number of subsets in the restricted partition U of A. We must point out that, from a practical point of view, the discontinuity points of f_U may lead, depending on the context, lead to unstable forecasts. In particular, in the neighborhood of each discontinuity point, an infinitesimal variation of the input variable could imply a discrete jump of the value of the dependent variable. In practical applications, in order to reduce the border regions between partition elements and the potential discontinuities of the MoM, we need to minimize the number of subsets of the restricted partition. In order to clarify how MoM works, let us look at Example 2.10.

Example 2.10. Let $n = 1$ and $\mathbb{X}$ be a set of $s = 40$ points in $\mathbb{R}$, as depicted in Figure 1-(a). The independent variable $x_1 = x$ takes values in the interval $[0, 6.5]$. Let us consider the Euclidean distance $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$ and $m = 3$ polynomial models f_1, f_2, f_3 of degree 1, 2, 3, respectively:

$$f_1 = -0.13802x + 0.38043$$

$$f_2 = 0.0056368x^2 - 0.17331x + 0.4166$$

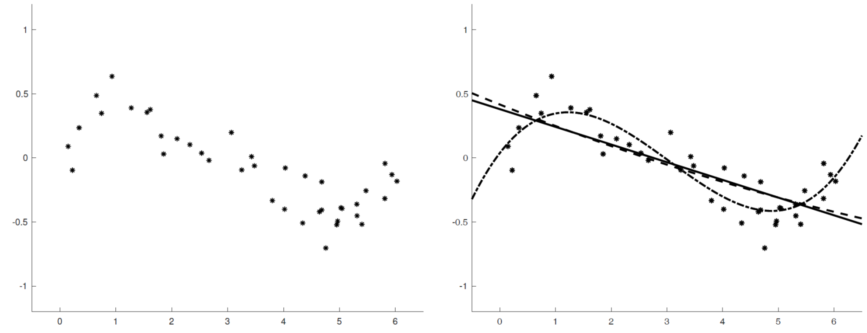
$$f_3 = 0.031524x^3 - 0.28876x^2 + 0.5659x + 0.037973$$

as shown in Figure 1-(b). Note that f_1 and f_2 are nested in f_3 (see Definition 2.1). The goodness of each model is:

$$\sum_{i=1}^{40}(y_i - f_1(p_i))^2 = 1.5646$$

$$\sum_{i=1}^{40}(y_i - f_2(p_i))^2 = 1.5553$$

$$\sum_{i=1}^{40}(y_i - f_3(p_i))^2 = 0.7874.$$



(a) (b)

Figure 1. (a) The data set $\mathbb{X}$ and the vector of responses $\boldsymbol{y}$; (b) the graphs of the three models: f_1 (solid line), f_2 (dashed line) and f_3 (dash-dot line).

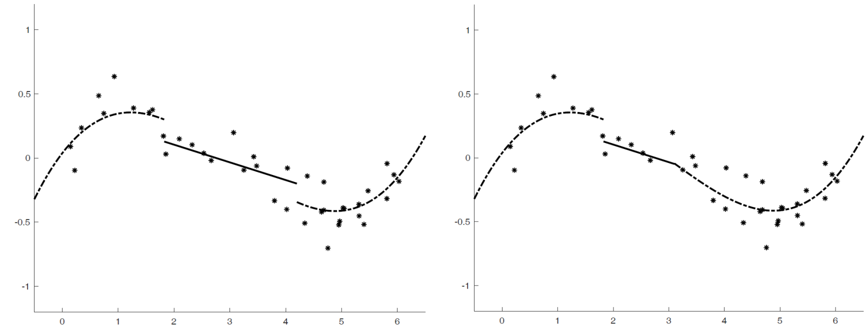
The restricted partition $U = \{U_1, U_2, U_3\}$ of $A = [0, 6.5]$ is:

$$U_1 = [0, 1.82), \quad U_2 = [1.82, 4.2), \quad U_3 = [4.2, 6.5].$$

For each $k = 1, 2, 3$ the index $\alpha(U_k)$ defined in (3) is:

$$\alpha(U_1) = 3, \quad \alpha(U_2) = 1, \quad \alpha(U_3) = 3.$$

The MoM f_U is shown in Figure 2-(a). The goodness of the MoM is $\sum_{i=1}^{40}(y_i - f_U(p_i))^2 = 0.7059$, which is strictly smaller than $\sum_{i=1}^{40}(y_i - f_j(p_i))^2$, $j = 1, 2, 3$ as Proposition 2.6 states.



(a) (b)

Figure 2. The data set $\mathbb{X}$, the vector of responses $\boldsymbol{y}$ and the graph of the MoM f_U (figure (a)) and the MoM f_V (figure (b)).

Figure 2-(a) shows the discontinuity points introduced with the MoM, $x_0 = 1.82$ and $x_1 = 4.2$. Let us consider the restricted partition $V = \{V_1, V_2, V_3, V_4, V_5, V_6\}$ of $A = [0, 6.5]$, where:

$$V_1 = [0, 0.8) \quad V_2 = [0.8, 1.82) \quad V_3 = [1.82, 3.1)$$

$$V_4 = [3.1, 4.2) \quad V_5 = [4.2, 5.03) \quad V_6 = [5.03, 6.5].$$

Note that V is a refinement of U. For each $k = 1, \ldots, 6$, $\alpha(V_k)$ as defined in (3) is:

$$\alpha(V_1) = 3 \quad \alpha(V_2) = 3 \quad \alpha(V_3) = 1$$

$$\alpha(V_4) = 3 \quad \alpha(V_5) = 3 \quad \alpha(V_6) = 3.$$

The MoM f_V is shown in Figure 2-(b). The goodness of the model is $\sum_{i=1}^{40}(y_i - f_V(p_i))^2 = 0.6994$ which is not greater than $\sum_{i=1}^{40}(y_i - f_U(p_i))^2 = 0.7059$, as Proposition 2.8 states.

According to Proposition 2.6, the MoM is no worse than selecting a single model.

In Proposition 2.11 we prove that for nested models (see Definition 2.1) the MoM outperforms single model selection and model averaging techniques.

Proposition 2.11. *Let* $A \subseteq \mathbb{R}^n$, *let* $\mathbb{X} = \{p_1, \ldots, p_s\} \subset A$ *be a set of s input data and* $\boldsymbol{y} = (y_1, \ldots, y_s) \in \mathbb{R}^s$ *be the dependent variable. Let* $f_1, \ldots, f_m$, *with* $m \geq 1$, *be m nested models belonging to the families* $\mathcal{F}_1, \ldots, \mathcal{F}_m$. *Suppose that* $\mathcal{F}_m$ *is a vector space and that* f_m *is the best model of* $\mathcal{F}_m$, *such that*:

$$f_m = \arg\min\left\{\sum_{i=1}^{s} d(y_i, f(p_i)) | f \in \mathcal{F}_m\right\}. \tag{9}$$

For each restricted partition U of A and any $\boldsymbol{c} = (c_1, \ldots, c_m)$ *such that* $\sum_{j=1}^{m} c_j = 1$ *the MoM* f_U *is no worse than the average model* $f_{\boldsymbol{c}}$.

Proof. Since $f_{\boldsymbol{c}} = \sum_{j=1}^{m} c_j f_j$, with $f_j \in \mathcal{F}_j \subseteq \mathcal{F}_m$, $j = 1, \ldots, m$, and $\mathcal{F}_m$ is a vector space, it follows that $f_{\boldsymbol{c}} \in \mathcal{F}_m$, for any $\boldsymbol{c} = (c_1, \ldots, c_m)$ with $\sum_{j=1}^{m} c_j = 1$. Therefore, from (9), it follows that:

$$\sum_{i=1}^{s} d(y_i, f_m(p_i)) \leq \sum_{i=1}^{s} d(y_i, f_{\boldsymbol{c}}(p_i)).$$

Combining the previous inequality with Proposition 2.6, we get:

$$\sum_{i=1}^{s} d(y_i, f_U(p_i)) \leq \sum_{i=1}^{s} d(y_i, f_m(p_i)) \leq \sum_{i=1}^{s} d(y_i, f_{\boldsymbol{c}}(p_i)).$$

Using Definition 2.3-(b) the proposition is proved.

3. MoMa: Model of Models Algorithm

The implementation of the MoM requires the following inputs: $f_1, \ldots, f_m$ alternative models to predict the dependent variable and a restricted partition of A.

The restricted partition of A can usually be derived using unsupervised techniques based on clustering approaches (see e.g., Hastie et al. [10]). In Section 2 the existence of a restricted partition of A is assumed to be given and the MoM is defined on that partition.

In practical applications, the choice of the restricted partition is independent and clustering algorithms could be used to randomly partitioning the data available.

This section, starting from $f_1, \ldots, f_m$ and a point in A, proposes an algorithm that iteratively constructs a restricted partition of A to obtain the

MoM. This approach allows us to define MoM even when no partition of the independent variables is provided in advance. The algorithm, which the authors call MoMa, works as follows.

Model of Models Algorithm (MoMa)

Given a set of s input data $\mathbb{X} = \{p_1, ..., p_s\} \subset A = A_1 \times \cdots \times A_n \subseteq \mathbb{R}^n$, with $A_1, ..., A_n$ real intervals, a vector $\boldsymbol{y} = (y_1, ..., y_s) \in \mathbb{R}^s$ of the realizations of the dependent variable, a vector $\boldsymbol{f} = (f_1, ..., f_m)$ of m real functions defined over A, $0 < p < 1$ and thresholds $0 < d_1 < d_2$, the algorithm returns (V, α), where $V = \{V_1, ..., V_r\}$ and $\alpha = (\alpha_1, ..., \alpha_r)$.

I. Set $k = 0$ and $Z_2 = \mathbb{X}$ and set d as follows

$$d = \begin{cases} d_1 & \text{if } \lfloor p \cdot s \rfloor < d_1 \\ d_2 & \text{if } \lfloor p \cdot s \rfloor > d_2 \\ \lfloor p \cdot s \rfloor & \text{otherwise.} \end{cases}$$

II. While $|Z_2| \geq d$ do

(1) set $k = k + 1$ and $Z_1 = \varnothing$;

(2) randomly choose an element $p_{i^*} \in Z_2$ and move it from Z_2 to Z_1;

(3) select the $d - 1$ points of Z_2 closest to p_{i^*}; move these points from Z_2 to Z_1;

(4) compute the index j^* such that

$$j^* = \arg\min_{j=1,...,m} \sum_{p_i \in Z_1} d(y_i, f_j(p_i))$$

(5) set $\bar{j} = j^*$;

(6) while $\bar{j} = j^*$ and $Z_2 \neq \varnothing$ do

```
    (a) let $p_{\bar{i}}$ be the point of $Z_2$ closest to $p_{i^*}$;
    (b) compute the index $\bar{j}$ such that
        $\bar{j} = \arg\min_{j=1,\dots,m} \sum_{p_i \in Z_1 \cup p_{\bar{i}}} d(y_i, f_j(p_i))$
    (c) if $\bar{j} = j^*$ then move $p_{\bar{i}}$ from $Z_2$ to $Z_1$;
(7) set $V_k = Z_1$, $\alpha_k = \bar{j}$ and $r = k$.
III. If $| Z_2 | > 0$ then set $k = k + 1$, $r = k$, $V_k = Z_2$ and
        $\alpha_k = \arg\min_{j=1,\dots,m} \sum_{p_i \in Z_2} d(y_i, f_j(p_i))$.
IV. Return $V = \{V_1, \dots, V_r\}$ and $\alpha = (\alpha_1, \dots, \alpha_r)$.
```

In order to obtain a significant frequency of observations for each element of the partition, (expressed in terms of number of data points in the corresponding subset of the partition), p, d_1 and d_2 are fixed in advance.

Of course, the value selection for p, d_1 and d_2 is crucial in real application and it represents a task for future research for the authors.

We should point out that the MoMa algorithm respects several properties as pointed out in the following Propositions.

Proposition 3.1. *The MoMa algorithm stops in a finite number of steps and returns a pair* (V, α), *where* $V = \{V_1, \dots, V_r\}$ *is a partition of* $\mathbb{X}$ *and* $\alpha = (\alpha_1, \dots, \alpha_r) \in \{1, \dots, m\}^r$.

Proof. The stopping criterion of the MoMa is given in steps II and III; since in step I, Z_2 starts with $Z_2 = \mathbb{X}$ and at each round of the algorithm at least one element p_{i^*} of Z_2 is removed from the set (step II.3, step II.6, step III), the condition $Z_2 = \emptyset$ is (possibly) reached after many iterations.

Regarding the correctness of the output, we simply prove that by induction that $V = \{V_1, ..., V_r\}$ is a partition of $\mathbb{X}$. Note that, during each iteration, by construction the set $Z_1 \cup Z_2$ does not change. In particular, during the first iteration, which is $k = 1$, we have $Z_1 \cup Z_2 = \mathbb{X}$, therefore running the loop, $\{V_1, Z_2\}$ is a partition of $\mathbb{X}$. We suppose that, at the end of the k-th iteration, $\{V_1, ..., V_k, Z_2\}$ is a partition of $\mathbb{X}$, that is $Z_2 = \mathbb{X} \backslash \bigcup_{j=1}^{k} V_j$. Then, at the $(k+1)$-th iteration, we obtain $V_{k+1} \cup Z_2 = \mathbb{X} \backslash \bigcup_{j=1}^{k} V_j$, so $\{V_1, ..., V_k, V_{k+1}, Z_2\}$ is a partition of $\mathbb{X}$.

Proposition 3.2. *Let* $A = A_1 \times \cdots \times A_n \subseteq \mathbb{R}^n$, *with* $A_1, ..., A_n$ *real intervals*, $\mathbb{X} = \{p_1, ..., p_s\} \subset A$ *be a set of s input data*, $\boldsymbol{y} = (y_1, ..., y_s) \in \mathbb{R}^s$ *be the vector of realizations of the dependent variable and* $\boldsymbol{f} = (f_1, ..., f_m)$ *be a vector of m real functions defined over A. Let* (V, α) *be the output of MoM Algorithm applied to* $(\mathbb{X}, \boldsymbol{y}, \boldsymbol{f})$. *Let r be the number of elements of V. For any partition* $U = \{U_1, ..., U_r\}$ *of A such that* $U_k \cap \mathbb{X} = V_k$, *for* $k = 1, ..., r$, *the MoM function* f_U *is given by*:

$$f_U(\boldsymbol{x}) = \begin{cases} f_{\alpha_1}(\boldsymbol{x}) & \textit{if } \boldsymbol{x} \in U_1 \\ \vdots & \vdots \\ f_{\alpha_r}(\boldsymbol{x}) & \textit{if } \boldsymbol{x} \in U_r. \end{cases}$$

Proof. Our first observation is that the partition U is a restricted partition of A. Since $V = \{V_1, ..., V_r\}$ is a partition of $\mathbb{X}$ (see Proposition 3.1), using hypothesis $U_k \cap \mathbb{X} = V_k$, for $k = 1, ..., r$, it follows that property (4) of Definition 2.4 is satisfied.

We consider the generic k-th iteration of the MoMa: when step II.7 is executed (at the end of the internal loop starting at step II.6) the index α_k satisfies:

$$\alpha_k = \arg\min\left\{ \sum_{p_i \in V_k} d(y_i,\ f_j(p_i)) \mid j = 1, \ldots, m \right\}$$

$$= \arg\min\left\{ \sum_{p_i \in U_k} d(y_i,\ f_j(p_i)) \mid j = 1, \ldots, m \right\}$$

and the last equality is derived using the hypothesis $U_k \cap \mathbb{X} = V_k$. Analogously, if step III is executed, we have:

$$\alpha_r = \arg\min\left\{ \sum_{p_i \in V_r} d(y_i,\ f_j(p_i)) \mid j = 1, \ldots, m \right\}$$

$$= \arg\min\left\{ \sum_{p_i \in U_r} d(y_i,\ f_j(p_i)) \mid j = 1, \ldots, m \right\}.$$

From formula (3) it follows that $\alpha_k = \alpha(U_k)$, hence the proposition is proved.

4. Empirical Evidence

This section shows the empirical evidence achieved on the simulated data set, by using the MoMa algorithm on two different data examples.

Example 4.1. Example 4.1 considers the data set introduced in Example 2.10. Let $\mathbb{X}$ be the set of 40 points in $\mathbb{R}$, $\boldsymbol{y}$ be the vector of 40 realizations (see Figure 1-(a)) and f_1, f_2, f_3 be polynomial models of degree 1, 2, 3 (see Figure 1-(b)).

Let us consider the Euclidean distance $\sum_{i=1}^{s} (y_i - \hat{y}_i)^2$ and run the MoMa Algorithm (with parameters $p = 0.1$, $d_1 = 6$ and $d_2 = 10$) on $\mathbb{X}$ and on the vector of the real functions $\boldsymbol{f} = (f_1, f_2, f_3)$. As a result, MoMa returns (V, α), where V is a partition of $\mathbb{X}$ composed of 2 subsets made up of 10 and 30 points respectively.

Figure 3 depicts the sets $(V_1, V_2) \in V$ using different symbols according to the associated model (the symbol + for f_1, the symbol $*$ for f_2 and the symbol $\circ$ for f_3).

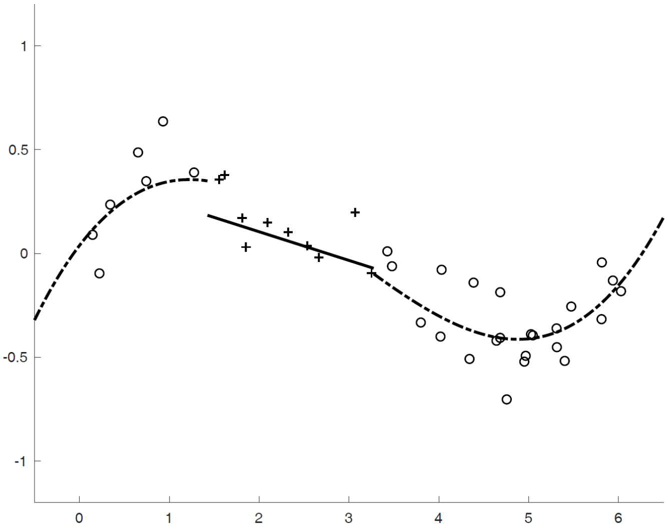


Figure 3. The points of the partition $V = (V_1, V_2)$ of the set $\mathbb{X}$ represented with different symbols, according to the corresponding model: + for f_1 and $\circ$ for f_3.

Let $U = \{U_1, U_2\}$ be any partition of $A = [0, 6.5]$ such that $U_i \cap \mathbb{X} = V_i$, $i = 1, 2$ and let f_U be the corresponding MoM (see Proposition 3.2)):

$$f_U(\boldsymbol{x}) = \begin{cases} f_1(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_1 \\ f_3(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_2. \end{cases}$$

The goodness of fit of the MoM f_U is:

$$\sum_{i=1}^{s} (y_i - f_U(p_i))^2 = 0.7663, \tag{10}$$

which is strictly smaller than $\sum_{i=1}^{s}(y_i - f_j(p_i))^2$, for $j = 1, 2, 3$ (as proved in Proposition 2.6) and comparable with the values $\sum_{i=1}^{s}(y_i - f_U(p_i))^2 = 0.7059$. The result obtained in 10 is comparable with $\sum_{i=1}^{s}(y_i - f_V(p_i))^2 = 0.6994$ of the MoMs f_U and f_V as reported in Example 2.10.

Example 4.2. Let $n = 2$ and $\mathbb{X}$ be a set of $s = 300$ points in $\mathbb{R}^2$, as depicted in Figure 4, and $\boldsymbol{y}$ be the vector of 300 realizations. The independent variables take values in $A = [-5, 5] \times [-5, 5]$. Let us consider the models f_1, f_2, f_3 and the Euclidean distance $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$. The following values give a measure of the goodness of fit of the models:

$$\sum_{i=1}^{s}(y_i - f_1(p_i))^2 = 21.1296$$

$$\sum_{i=1}^{s}(y_i - f_2(p_i))^2 = 6.8069$$

$$\sum_{i=1}^{s}(y_i - f_3(p_i))^2 = 29.9841. \tag{11}$$

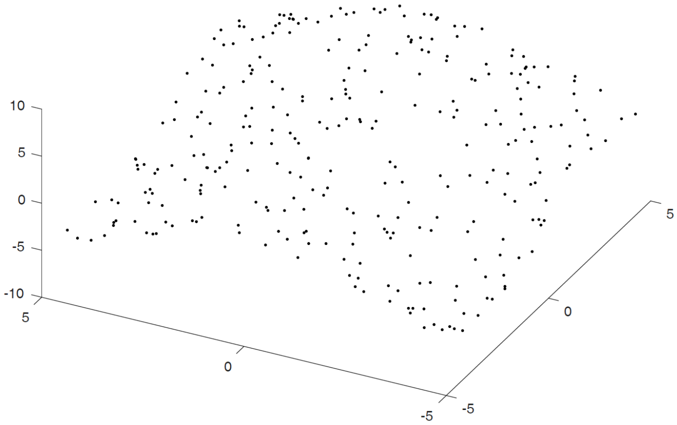


Figure 4. The data set $\mathbb{X}$ and the vector $\boldsymbol{y}$ of realizations.

Run MoMa (with parameters $p = 0.02$, $d_1 = 20$ and $d_2 = 30$) on the data set $\mathbb{X}$, using the realizations $\boldsymbol{y}$ and the vector of real functions $\boldsymbol{f} = (f_1, f_2, f_3)$. MoMa returns the pair (V, α), where V is a partition of $\mathbb{X}$ made up of 5 subsets and $\alpha = (2, 1, 3, 3, 2)$. The sets of the partition $V = \{V_1, V_2, V_3, V_4, V_5\}$ have the following cardinalities:

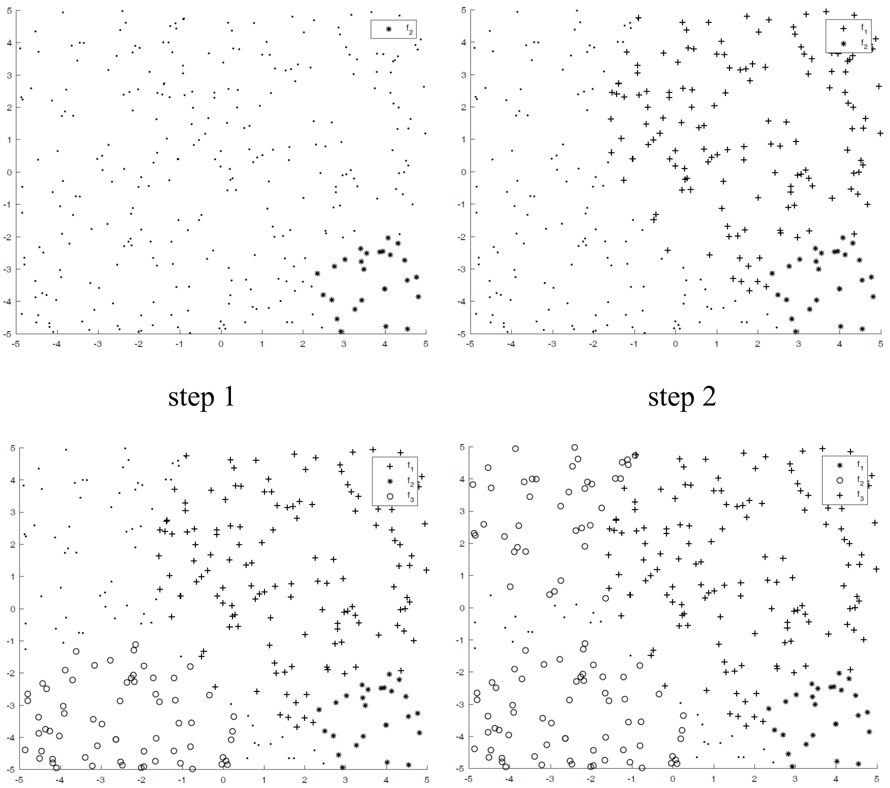
$$|V_1| = 25 \quad |V_2| = 145 \quad |V_3| = 63$$

$$|V_4| = 40 \quad |V_5| = 27.$$

Figure 5 shows the partitions obtained in each iteration of MoMa. During the first iteration the subset V_1 is derived and the corresponding model is f_2 (see subfigure (step 1)). In this case the goodness of fit corresponding to f_2 is given by the value 6.8069 as in (11). Then, during the second, the third and the fourth iterations, the remaining points of $\mathbb{X}$ are split into the subsets V_2, V_3 and V_4 connected to the models f_1, f_3 and f_3 respectively (see subfigures (step 2), (step 3) and (step 4)). In these cases the numerical values representing a measure of the goodness of fit are 6.6927, 6.6518 and 6.6515. We observe that, according to Proposition 2.8, the three values are decreasing, meaning that the goodness of fit of the model under construction is improved. Finally, during the last iteration, the points still lying in $\mathbb{X}$ are gathered in the subset V_5 and associated to the model f_2 (see Figure 6).

Let $U = \{U_1, U_2, U_3, U_4, U_5\}$ be any partition of $A = [-5, 5] \times [-5, 5]$ such that $U_i \bigcap \mathbb{X} = V_i$, $i = 1, \ldots, 5$ and let f_U be the corresponding MoM (see Proposition 3.2)):

$$f_U(\boldsymbol{x}) = \begin{cases} f_2(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_1 \\ f_1(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_2 \\ f_3(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_3 \\ f_3(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_4 \\ f_2(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in U_5. \end{cases}$$



step 1

step 2

step 3

step 4

Figure 5. The partitions of $\mathbb{X}$ after the first 4 iterations of MoMa. In subfigure (step k), $k = 1, ..., 4,$ the points of the subsets $V_1, ..., V_k$ are shown according to the associated model $(+ \text{ for } f_1, * \text{ for } f_2 \text{ and } \circ \text{ for } f_3),$ while the points still lying in $\mathbb{X}$ are represented with the dot symbol.

The goodness of fit of the MoM f_U is:

$$\sum_{i=1}^{s} (y_i - f_U(p_i))^2 = 6.6515,$$

which is strictly smaller than $\sum_{i=1}^{s} (y_i - f_j(p_i))^2,$ for $j = 1, 2, 3,$ as proved in Proposition 2.6.

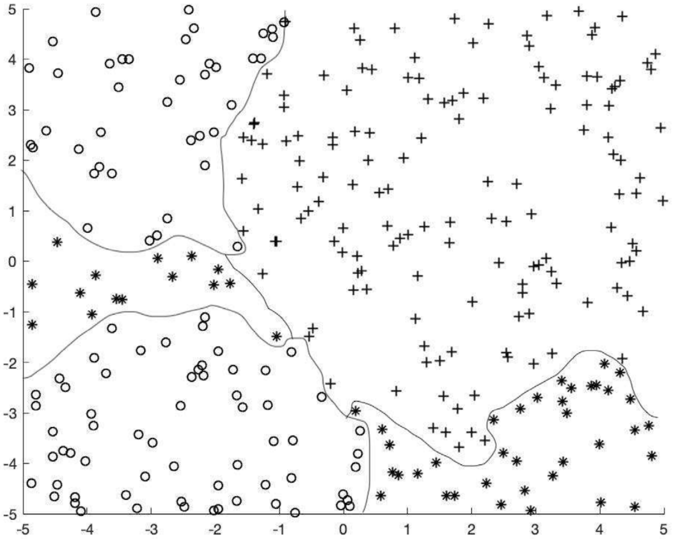


Figure 6. The partition of the set $\mathbb{X}$: the points of the subsets $V_1, ..., V_5$ are represented according to the associated model: + for f_1, $*$ for f_2 and $\circ$ for f_3.

5. Conclusions

This paper presents a novel approach called Model of Models (MoM). MoM concerns the selection of the best model for a given partition of the data derived from the realization of the independent variables. Compared to model averaging approaches proposed in the literature to deal with model uncertainty, MoM does not require the selection of models to include in the pool of models and it works without resorting to the combination of model predictions.

MoM works on parametric and non parametric predictive models. The selection of the partition of the independent variables is derived following the model estimation step. This helps to overcome the issue related to overfitting. Assuming a partition of the data, the authors implement the methodological proposal introducing a new algorithm which they call MoMa.

The theoretical results at hand, coupled with the empirical evidence achieved on simulated data, underline that MoM is a good strategy to deal with model choice and model uncertainty.

In terms of practical application, future points of research can be summarized in: elicitation of the parameters involved in the MoMa algorithm and unsupervised techniques to derive optimal partition of the data. Further research will consider testing the forecasting ability of MoM within a cross validation framework.

Acknowledgements

This work has been supported by RIDS (RES Institute for Data Science). We thank the anonymous referees for their comments, which have improved the quality of the paper. (Authors contributions: Section 1 and Section 5 Silvia Figini; Section 2 and Section 3 Pierpaola Uberti; Section 4: Maria Laura Torrente).

References

[1] T. Ando and K. C. Li, A model-averaging approach for high-dimensional regression, Journal of the American Statistical Association 109 (2014), 254-265.

[2] T. Ando and K. C. Li, A weight-relaxed model averaging approach for high-dimensional generalized linear models, Annals of Statistics 45 (2017), 2654-2679.

[3] L. Breiman, Bagging predictors, Machine Learning 26 (1996), 123-140.

[4] L. Breiman, Random forests, Machine Learning 45(1) (2001), 5-32.

[5] L. Breiman, Statistical modeling: the two cultures, Statistical Science 16(3) (2001), 199-231.

[6] G. W. Brier, Verification of forecasts expresses in terms of probability, Monthly Weather Review 78 (1950), 1-3.

[7] S. Figini, R. Savona and M. Vezzoli, Corporate default prediction model averaging: A normative linear pooling approach intelligent systems in accounting, Finance and Management 23 (2016), 6-20.

[8] T. Fragoso, W. Bertoli and F. L. Neto, Bayesian model averaging: a systematic

review and conceptual classification, International Statistical Review 86 (2018), 1-28.

[9] J. Geweke and G. Amisano, Prediction with misspecified models, American Economic Review 102(3) (2012), 130-141.

[10] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, 2009.

[11] J. A. Hoeting, D. Madigan, A. Raftery and C. T. Volinsky, Bayesian model averaging: a tutorial, Statistical Science (1999), 382-401.

[12] N. L. Hjort and G. Claeskens, Frequestist model average estimators, Journal of the American Statistical Association 98 (2003), 879-899.

[13] N. L. Hjort and G. Claeskens, Model Selection and Model Averaging, Cambridge University Press, 2012.

[14] L. B. Klebanov, S. B. Rachev and F. J. Fabozzi, Robust and Non-robust Models in Statistics, Nova Science Publishers, 2009.

[15] B. Lin, Q. Wang, J. Zhang and Z. Pang, Stable prediction in high-dimensional linear models, Statistics and Computing 27 (2017), 1401-1412.

[16] J. Lv and J. S. Liu, Model selection principles in mis-specified models, Journal of the Royal Statistical Society Series B 76 (2014), 141-167.

[17] S. Omer and R. Lior, Ensemble learning: a survey, Wires Data Mining and Knowledge Discovery 8(4) (2018).

[18] R. Ranjan and T. Gneiting, Combining probability forecasts, Journal of the Royal Statistical Society Series B (2010), 71-91.

[19] A. Raftery, T. Gneiting, F. Balabdaoui and M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, Monthly Weather Review (2005), 1155-1174.

[20] A. Singh, S. Mishra and G. Ruskauff, Model averaging techniques for quantifying conceptual model uncertainty, Ground Water 48(5) (2010).

[21] M. Stone, The opinion pool, Annals of Mathematical Statistics 32 (1961), 1339-1342.

[22] H. Wang, X. Zhang and G. Zou, Frequentist model averaging estimation: a review, Journal of Systems Science and Complexity 22 (2009), 732-748.

[23] X. Zhang, G. Zou and H. Liang, Model averaging and weight choice in linear mixed-effects models, Biometrika 101 (2014), 205-218.