Proceedings of Workshop on Signal Processing, Communications,
Chaos and Systems, Newport, RI, June 2002, pp. 59-92

Model-Set Design, Choice, and Comparison for Multiple-Model Approach to Hybrid Estimation*

X. Rong Li Zhanlue Zhao Peng Zhang Chen He
Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, USA
Phone: 504-280-7416, Fax: 504-280-3950, xli@uno.edu

Abstract – *The most important problem in the application of the multiple-model approach is the design of the model set used. This paper deals with this challenging topic in a general setting, along with model-set choice and comparison. General and representative problems of model-set design, choice, and comparison are considered. Modeling of models as well as true mode as random variables is proposed. Several general methods for design of model sets are presented by minimizing distribution mismatch, minimizing modal distance, and moment matching. The concept of relative efficacy of each model in a set and its two quantitative descriptions are introduced. Optimality criteria and performance measures for model-set design, choice, and comparison based on base-state estimation, mode estimation, mode identification, hybrid-state estimation, information metrics, and hypothesis testing are presented. Several computationally efficient and easily implementable solutions of the model-set choice problems based on sequential hypothesis testing are presented, some of which are optimal. Examples that demonstrate how some of these theoretical results can be used as well as their effectiveness are given. Many of the general results presented in this paper are also useful for performance evaluation of MM algorithms.*

Keywords: Multiple models, model-set design, variable structure, adaptive estimation, target tracking.

1 Introduction

Hybrid estimation is the estimation of a *hybrid process*, such as the state of a hybrid system, which involves two types of components: those varying continuously, known as **base states**, and those that may jump only, known as **modes** or **modal state** [12, 29]. In systems terminology, a base state is the state of a conventional system, while each mode represents a possible system behavior pattern or structure. Hybrid estimation has two major goals: base-state estimation, which is the conventional state estimation, and mode estimation or identification, which actually amounts to decision.

Multiple-model (MM) method is a major adaptive approach to hybrid estimation. It is cost-effective, robust, and has a parallel structure. In the MM method, a set of models is designed to cover modes and the overall estimate is obtained by a certain combination of the estimates from the filters based on these models, respectively. The MM method has received a great deal of attention in recent years due to its unique power and great success in handling problems with both structural and parametric uncertainties and/or changes, and in decomposing a complex problem into simpler subproblems, particularly in target tracking and fault detection and isolation (see, e.g., [12] for a long list of references).

The MM method was initiated in [24]. Many applications (or reinventions) of this MM estimator can be found in the literature under various names (see, e.g., [12]). The first generation of MM algorithms does not consider possible jumps in mode and can be referred to as *autonomous* MM algorithms in that model-based filters do not interact with each other. In the second generation, such as the GPB [1, 10] and IMM [8] algorithms, the mode is assumed to be able to jump among members of a set, usually modeled as Markovian transition. These first two generations have a fixed structure in that they use a *fixed set of models* at all times, although each model in the set could be time-variant or adaptive. They have certain fundamental limitations, which stem from the fundamental assumption that the mode at any time can be represented sufficiently accurately by one of the models in a fixed set that can be determined before measurements are received and its inability to incorporate certain types of prior information. The third generation, known as variable-structure MM (VSMM) [18, 14], overcomes these fundamental limitations by using a variable set of models determined in real time adaptively.

*Research supported in part by ONR grant N00014-00-1-0677, NSF grant ECS-9734285, and NASA/LEQSF grant (2001-4)-01.

For a survey of the MM approach, the reader is referred to [12]. An easily accessible account of the VSMM approach is given in [14], while the IMM algorithm and its variants for target tracking are surveyed in [25].

There are two major directions to improve the MM solution of a given hybrid estimation problem: develop a better MM algorithm in general and design a better model set in particular.

Model-set design is the most important issue in the application of MM estimation. The performance of an MM algorithm for a given problem depends largely on the set of models used and the primary difficulty in the application of the MM method is the design of the model set. Numerous publications have appeared in which ad hoc designs were presented. Unfortunately, very limited theoretical results on this important issue are available. It was shown theoretically in [18] that the use of too many models is as bad as the use of too few models. A circular criterion for model-set choice was presented in [18]. When the mode space is a continuous region, a necessary and sufficient condition was presented in [19] for a convex combination of estimators to be superior to each individual estimators, based on respective model sets. In order to apply the MM method to problems with uncertain *parameters*, two important questions are: (a) which quantity is best selected as the estimatee (i.e., the quantity to be estimated) and (b) how to quantize the parameter space optimally. [13] provides theoretical results on the optimal selection of the estimatee. A procedure to determine the choice of the quantization points was presented in [27] given the number of quantization points. A necessary condition for the effective performance of MM estimation was presented in [9] for a jump linear time-invariant system in terms of its dc gain.

This paper presents theoretical results on model-set design, choice, and comparison. Modeling of models as well as true mode as random variables is proposed. Several general methods for design of model sets, along with the initial model probabilities, are presented. They include distribution approximation, minimizing mismatch between mode and models, and moment matching. As a theoretical basis, criteria and performance measures for model-set choice, comparison, and design are proposed, including those for base-state estimation, mode identification/estimation, and hybrid estimation, as well as by hypothesis testing. A number of solutions to model-set choice problems are presented based on sequential hypothesis testing.

An important question in the model-set design is: How effective a model is when it is used in a model set? Albeit important, we are not aware of any (theoretical) attempt at answering this question. Another contribution of this paper is the introduction of the concept of the relative efficacy of a model in a set and the development of several methods of computing it.

Examples are given that demonstrate how the above theoretical results can be used for model-set design and choice. These results provide insights that are helpful for model-set design, choice, and comparison, such as a better understanding of the function of each model in a set and how to select the parameter values of a model.

Since this paper is quite long, we provide the following table of contents

- Introduction
- Problems of model-set design, choice, comparison, and adaptation in MM estimation
- Probabilistic modeling of modes and models
- Problem formulation for model-set design
- Model-set design by minimizing distribution mismatch
- Model-set design by minimizing modal distance
- Model-set design by moment matching
- Model Efficacy
- Criteria and measures for model-set design, choice, and comparison
- Model-set choice by sequential hypothesis testing
- Illustrative examples of model-set design
- Illustrative examples of model efficacy
- An Illustrative example of model-set choice
- Selection of estimatee for MM estimation
- Concluding remarks
- Appendix

2 The Problems of Model-Set Design, Choice, Comparison, and Adaptation

Model-set comparison and choice deal with the following problem: given a family of candidate model sets, compare these sets and determine which set is the best. **Model-set choice** is more decision oriented in the sense that it simply determines which set is the best without emphasizing how much better it is than the other sets. **Model-set comparison** pays more attention on how much better one model set is than another.

Model-set design differs from the comparison and choice in that it does not necessarily have a given family of candidate model sets. It determines the model set to be used for a given problem. Typical issues in model-set

design include: How many models should be used in the set and how to determine this number? Given the number of models, how to design each model in the set? What is the structure of each model set and each model? What parameter values to use given the structures of models and model sets? Clearly, model-set comparison and choice can be viewed as integral parts of the model-set design.

There are two types of model-set design: offline and online. Offline design is for the total model set or the initial model set in a variable-structure approach, as well as for the fixed-structure approach. In a fixed-structure algorithm, the model set used cannot vary and is determined a priori by model-set design. In a variable-structure algorithm, the model set in effect at any time is determined by an adaptation process, known as **model-set adaptation,** which may be viewed as an online (real-time) design process and will depend on the total model set determined a priori if such a set exists. A natural and promising VSMM approach to estimation is the recursive adaptive model-set (RAMS) approach [15, 14]. It consists of two functional components: model-set adaptation and model-set sequence conditioned estimation. Model-set adaptation is the more difficult component. It decides what model set to use at each time. It can usually be decomposed into two tasks: propose proper candidate model sets and select the best set from these candidates. The whole problem thus amounts to model-set design in real time while the second task is model-set choice. There could have many ways of proposing candidates, which is the primary task of a particular VSMM algorithm. This paper focuses on offline model-set design, along with model-set choice and comparison, while [15, 14] deals with model-set adaptation, in particular, online mode-set choice.

A fundamental assumption of the MM method is that the possible true mode at any time is matched exactly by one of the models used at that time. It is usually the case in realty, however, that none of the models in the set in effect matches exactly the true mode at the time. Many questions thus arises, such as

- "Which model should be deemed the best if the true mode does not match any of the models?"
- "Which set is best if the true mode falls in the common part of the coverage of several model sets?"

Such questions are important for model-set design but are not easy to answer in a general setting. The results of this paper can be used to answer such questions theoretically.

3 Probabilistic Modeling of Models and Modes

In this paper, a **mode** refers to the physical behavior pattern or structure of a system/process (or its precise mathematical model), and a **model** refers to the (possibly simplified) mathematical representation or description of the system or process on which an estimator is based (see [14] for a more detailed explanation). Such a distinction is necessary where mismatch between the model and mode is of concern.

Denote by $\mathbf{S}$ the **mode space**, that is, the set of possible modes under consideration. In general, mode space $\mathbf{S}$ may be either a discrete (finite or countable) set or a continuous region. In the latter case it is assumed that a system mode may only *jump* from a point in $\mathbf{S}$ to another one, rather than vary *continuously*.

A contribution of this paper is the recognition of the need for and introduction of probabilistic modeling of models as well as the true mode.

The need to have a proper description of the true mode is evident: Without such a description, model-set design and performance evaluation of MM algorithms are essentially groundless — we can always find a scenario under which any given realizable "optimal" model set is worse than some other model set. Deterministic descriptions of the true mode in the form of "typical" or "representative" scenarios are prevailing in the literature of MM estimation, particularly for performance evaluation. Such deterministic descriptions have certain drawbacks. For example, the choice of particular scenarios is fairly arbitrary, and thus the corresponding performance evaluation results are less objective or convincing since the performance of MM algorithms is highly dependent on test scenarios. (The **scenario dependence** of the performance of a hybrid estimation algorithm is elaborated in [17].) It is impossible to develop general, systematic methods for model-set design on the basis of such "arbitrary" descriptions of the true mode.

We propose to model the true mode as a random variable $s : \Omega \to \mathbf{S}$, where $\mathbf{S}$ is the mode space and Ω is the sample space. The random variable s may be continuous, discrete, singular, or hybrid. Let $F_s(x)$ and $f_s(x)$ be its cumulative distribution function (cdf) and probability density function (pdf) if exists, respectively. In practice, they can be obtained by past data using statistical techniques or simply from experience. For example, a transposed (i.e., symmetrical) three-phase overhead transmission line in a power system has three simple modes (i.e., normal, single-phase to ground fault, and phase-to-phase fault) and several composite modes (e.g., two-phase to ground fault and three-phase to ground fault). Data of the past operation records (e.g., fault rate and percentage of fault type) provide the required probability distribution of the mode. For a particular application of MM estimation, if $F_s(x)$ is not available at this stage, the benefit of having such a cdf — as presented in this and other papers — suggests that it may be worthwhile to obtain such a cdf. This is a manifestation of guidance of theory to practice. Without such guidance, most practical probabilistic models (e.g., Gaussian models, Poisson mod-

els) would not have been developed and probability theory would have very limited practical value.

Similarly, *we also propose that the problem of designing a model set M (and the corresponding initial model probabilities) be formulated as that of designing a* ***random model*** *m with range M*; that is, design a random variable $m : \Omega \to M$, where Ω is the sample space. As such, the following needs to be determined: (a) cardinality $|M|$ (i.e., number of models); (b) all elements m_i of $M = \{m_1, m_2, \ldots, m_{|M|}\}$ (i.e., model locations/values); (c) prior (or initial) model probabilities $P\{m = x\}$. Note that cdf $F_m(x)$ of m, or equivalently, probability mass function (pmf) $p_m(x) = P\{m = x\}$ summarizes all information needed. While this concept of *random* model may appear alien to a practitioner, we need only to recall that a random variable is (corresponds to) in fact nothing but a properly defined set of *deterministic* numbers. It is exactly in this way that a set of deterministic models used in the MM method, along with the above constraints (a)–(c), defines a random model.

More generally, the second and third generations of MM algorithms require design of (Markovian) laws governing model transitions based on transitions of the true mode. Even more generally, the true mode is better modeled as a random process $s(t) : (\Omega, \mathcal{F}, P) \times T \to \mathbf{S}$; that is, $s(t)$ is a family of random variables, indexed by $t \in T$ and defined on a common probability space $(\Omega, \mathcal{F}, P)$. Similarly, the problem of model-set design is better formulated as the determination of a random process $m(t) : (\Omega, \mathcal{F}, P) \times T \to \mathbf{M}$, where $\mathbf{M}$ is the total model set. These more general formulations are useful for model-set adaptation and design of model transitions. For offline model-set design (the topic of this paper), however, it usually suffices to consider s and m as random variables, which are completely described by their cumulative distribution functions.

For simplicity, we assume that the true mode is continuous in this paper. The same approach works for other cases, although modification is sometimes needed. We always assume that the model is discrete (in fact, finite).

For many applications, the true mode s has a real physical meaning directly and the above probabilistic modeling is clearly reasonable. For many other applications, however, s is an index of underlying structures (or behavior patterns) and it is difficult, if not impossible, to define a proper distance metric directly for $\mathbf{S}$ convincingly with a clear interpretation. In such cases, cdf of s may possibly be defined over an abstract space where the elements of s are arranged such that the neighboring elements correspond to the neighboring structures in the physical world. Then a question is how to define the neighbor concept for structures in the physical world? This question can be answered by using, e.g., Kullback-Leibler distance between the distribution or likelihood functions of any two structures s_i and s_j.

4 Formulation of Model-Set Design

Following the previous section, the true mode (at any time) can be reasonably modeled as a continuous random variable in many cases, while it is better modeled as a discrete (or hybrid) random variable in many other cases. In any case, its sample space $\mathbf{S}$ is usually much larger than the model set M affordable in practice.

From the probabilistic modeling of the true mode and models, it is clear that the *model-set design is essentially a problem of finding a discrete random variable m to approximate a given random variable s*, which can be continuous, discrete, singular, or hybrid, depending on the application. [1] Unfortunately, to our knowledge, there is no generally acceptable solution to this problem in the literature.

We propose three classes of systematic solutions in Sections 5, 6, and 7 below.

5 Minimum-Mismatch Design

The first solution is based on the idea of finding the cdf $F_m(x)$ of a discrete random variable (model) m to approximate the cdf $F_s(x)$ of any given random variable (mode) s. We describe this solution in the scalar case (i.e., for scalar s and m) first and extend it to the vector case later.

5.1 Scalar Case

Assume that the cdf $F_s(x)$ of true mode s is known. Given a tolerance ϵ, we want to construct the cdf $F_m(x)$ of a discrete random variable (i.e., model set) such that $|F_s(x) - F_m(x)| \leq \epsilon$ for all x.

It can be shown that for any given cdf $F_1(x)$ we can find the cdf $F_2(x)$ of some discrete random variable that is arbitrarily close to $F_1(x)$ in terms of the following distance metric

$$d(F_1, F_2) = \max_{x \in R} |F_1(x) - F_2(x)| \tag{1}$$

where $R = [-\infty, \infty]$. In other words, the problem under consideration always has a solution. Further, a general procedure of finding such a cdf $F_2(x)$ is presented in Appendix A.1. What we present below amounts to applying the results therein to model-set design.

What is the minimum number of models needed? The following lemma answers this question.

Lemma 5.1. Given a tolerance ϵ in the above distance metric, the minimum number of models needed is given by

$$|M| = \lceil 1/2\epsilon \rceil = \text{smallest integer not smaller than } 1/2\epsilon$$

A proper tolerance ϵ is not always easy to come by. In some cases, the number of models $|M|$ is predetermined directly from, say, resource for processing or computation.

[1] This probabilistic view also makes it quite intuitive the fundamental finding of [18] that the optimal model set M for the MM approach is $M = \mathbf{S}$ — the performance of MM estimators deteriorates if either extra models are used ($M \supset \mathbf{S}$) or some models are missing ($M \subset \mathbf{S}$) — and the deterioration worsens as M and $\mathbf{S}$ become more mismatched.

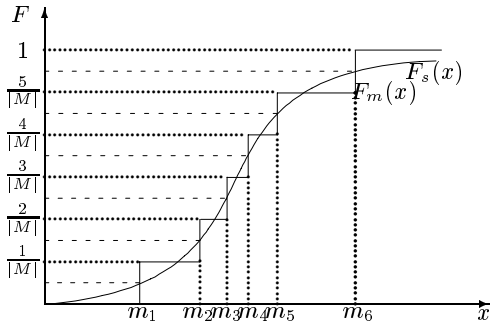


Fig. 1: Approximating a cdf by a stair-case type cdf with given tolerance.

Theorem 5.1 (Minimum-set design**)**. Given $|M|$, the model set M^*, along with the pmf p^*, (i.e., the random model) that minimizes the distance metric defined by (1), that is, for $m \in M$,

$$\{M^*, p^*\} = \arg \inf_{\{M,p\} \text{ with given } |M|} \sup_{x \in \mathbf{S}} |F_s(x) - F_m(x)|$$

is given by

$$\begin{aligned} m_i &= \arg \underset{x \in \mathbf{S}}{} \left[F_s(x) = \frac{i - 1/2}{|M|} \right] \\ & \quad i = 1, \ldots, |M| \\ M^* &= \{m_1, m_2, \ldots, m_{|M|}\} \end{aligned} \tag{2}$$

along with the following evenly distributed pmf (i.e., initial model probabilities):

$$\begin{aligned} p_m(x)|_{x=m_i} &= P\{m = m_i | m \in M^*\} \\ &= \frac{1}{|M|}, \; i = 1, \ldots, |M| \end{aligned} \tag{3}$$

Proof. It follows from Appendix A.1 and is in fact self-evident.

This design is depicted in Fig. 1. Note that m_i is chosen to satisfy (2) only from the elements of $\mathbf{S}$, and thus $M \subset \mathbf{S}$.

This approach to model-set design is intuitively appealing. It partitions the mode space into equally probable regions and places a model at the "center" (in fact, median) of each region. As such, all models are equally loaded in that they are equally likely to take effect and cover an region of equal probability. It uses the minimum number of models. It is also perfectly consistent with the common practice of assigning equal initial probability to every model.

Nevertheless, this approach has several weaknesses. First, it is applicable only to cases where cdf of s is available. Second, some or all models m_i may happen to be located in an area of a low probability density. In this case, the models and the mode in effect are likely to have a large mismatch, which implies inferior performance of the corresponding MM algorithm. Third, a few models may have to cover a large region of true mode with a low probability density and thus lead to poor results if the true model turns out to be in this region. Finally, to have a small tolerance in cdf error, the separation between consecutive models may be inevitably small in the areas where s has a high probability density, which is often to be avoided in MM estimation mainly to save computation. In such a case, we may either uphold the tolerance (and thus the separation) or relax it to increase the separation and thus reduce model-set size. The latter may be justified by the fact that a larger error in $|F_s(x) - F_m(x)|$ does not necessarily result in poorer performance of the MM algorithm.

In view of the above and that it is intuitively appealing to have a model at each peak of the pdf $f_s(x)$, we recommend the following. First, place a model at each peak; then use the above approach to obtain the other models; if desirable, adjust the locations of these other models so that models are distributed slightly more uniformly over $\mathbf{S}$.

5.2 Vector Case

When s is not scalar, in general, (2) does not yield a unique solution $M = \{m_1, m_2, \ldots, m_{|M|}\}$ because $F(s) = \frac{i-1/2}{|M|}$ has infinitely many solutions. In this case, the mode space in general can be partitioned into equally probable regions S_i, represented by models m_i. Several ways of determining the location m_i and the regions S_i are currently under investigation. For example, they may be determined such that m_i satisfies (2) and has the smallest expected distance to points in S_i; that is, m_i is the center of probability mass in S_i (see Sec. 6). Applications of set-partitioning results (see, e.g., [2]) are currently being explored.

We now describe a design procedure for the 2D case, which uses a "minimal" number of models given any tolerance on mismatch between the cdfs of the mode and (random) model. It can be easily extended to higher dimensions.

Consider the cdf of a 2D mode s: $F_s(x, y) = F(x, y)$. As explained before, design of a model set along with the initial model probabilities (i.e., model weights) amounts to constructing a random variable m (i.e., a random model) with a certain cdf $F_m(x, y)$. Our goal is to determine locations of a "minimal" number of models along with probability weights such that the resultant cdf $F_m(x, y)$ satisfies the requirement $\max_{x,y} |F_s(x, y) - F_m(x, y)| \leq \epsilon$.

Let $D(x, y) = F_s(x, y) - F_m(x, y)$ be the difference in cdf. Assume for simplicity that $F(x, y)$ is continuous. In Fig. 2, the origin and the upper right corner stand for $(-\infty, -\infty)$ and (∞, ∞), respectively, at which $F(-\infty, -\infty) = 0$ and $F(\infty, \infty) = 1$. Note that $F(x, y)$ is monotonically increasing.

The procedure consists of three steps, as illustrated in Fig. 2.

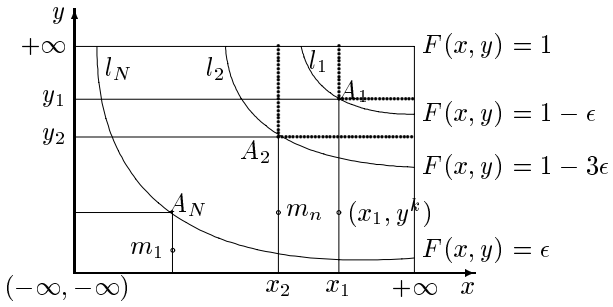


Fig. 2: Illustration of 2D model-set design by minimizing cdf mismatch.

First, determine the equal-height lines $l_{1-\epsilon}$, $l_{1-3\epsilon}$, ..., $l_{1-(2N-1)\epsilon}$, l_ϵ, where N is an integer such that $\epsilon < 1 - (2N-1)\epsilon \leq 3\epsilon$. This means that $F(x_i, y_i) = 1 - i\epsilon$ for any point (x_i, y_i) on the line $l_{1-i\epsilon}$.

Second, determine the points $A_1, A_2, \ldots, A_N$. The location of A_1 is (x_1, y_1). It minimizes $|F(x_1, \infty) - F(\infty, y_1)|$ among all points on $l_{1-\epsilon}$. The point A_1 determines two reference lines for the next point A_2. Its location is (x_2, y_2), which minimizes $|F(x_2, y_1) - F(x_1, y_2)|$ among all points on $l_{1-2\epsilon}$. A_i is determined likewise.

Third, determine the model locations $m_1, m_2, \ldots, m_m$. We place models on the horizontal and vertical lines determined by points $A_1, A_2, \ldots, A_N$. For the line x_2-A_2 it uses the next line x_1-A_1 as a reference. Note that $F(x_1, y)$ is monotonically increasing on the line x_1-A_1. If a point (x_1, y^k) is the lowest point such that $D(x_1, y^k) > \epsilon$, then choose (x_2, y^k) as a model location. This process is done from left to right (i.e., for $x_N, \ldots, x_1$) and from bottom up (i.e., for $y^1, y^2, \ldots$).

The weight of each model is determined at the same time the model location is determined. The weight is "how high" a jump is needed at each model location. The upper bound on the height of a jump of a model at (x_i, y^k) is determined by the difference $D(x, y)$ along the line x_{i-1}-A_{i-1} at or above y^k.

The model locations and weights on the horizontal line y_i-A_i are determined in exactly the same way.

Fig. 5.2 shows an example of the true pdf and the model locations designed, depicted by the sharp peaks. In the design, the tolerance $\epsilon = 0.1$ was chosen. The resultant model locations concentrate around the major peaks of the true density. Fig. 4 shows the error $D(x, y)$. It is bounded by $\epsilon = 0.1$, as required.

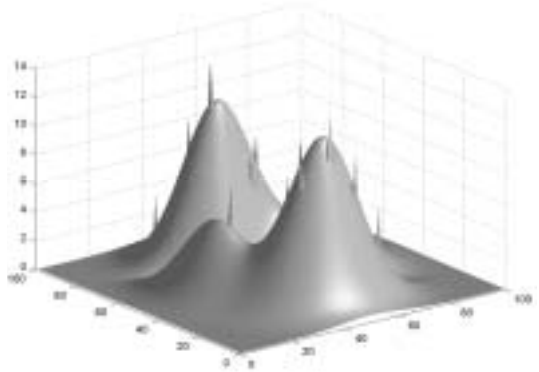


Fig. 3: The true pdf and designed model locations.

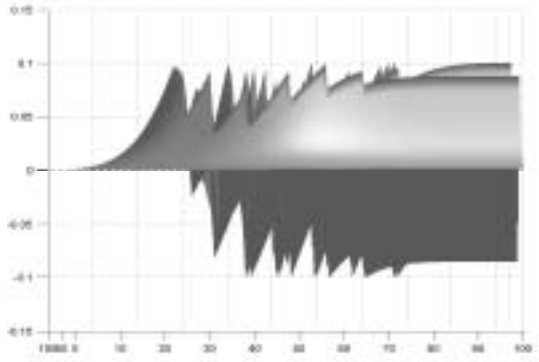


Fig. 4: $D(x, y)$ — the difference in cdf.

6 Minimum-Distance Design

In the previous section, we design a model m to approximate the true mode s by constructing a cdf $F_m(x)$ that is close to the cdf $F_s(x)$ — the design is actually done in the space of the distribution functions. Alternatively, the design can also be done in the vector space of random variables; that is, find an m that is close to s in their vector space directly. In order to do this, a metric of the closeness between model and mode is needed.

Closeness metric between model and mode. The distance metric in the vector space of random variables is most often defined as the square root of the mean-square value $\bar{d}(s, m) = (E[d(s, m)])^{1/2}$, where $d(s, m) = (s - m)'(s - m)$. Of course, other metrics can also be defined, such as $\bar{d}(s, m) = (E[d(s, m)])^{1/p}$, where $d(s, m) = [(s - m)'(s - m)]^{p/2}$. When s and m are vectors, $d(s, m)$ is actually a scalar metric of the families of vectors, since a random vector actually corresponds to a family of vectors in a linear space. We will consider the more general metric with an arbitrary p but we are more interested in the case $p = 1, 2$.

For $s \in \mathbf{S}$ and $m \in M$, we have

$$\begin{aligned} & E[d(s,m)] = E[E[d(s,m)|m]] \\ = & \sum_{m_i \in M} P\{m = m_i\} \int_{\mathbf{S}} d(s,m_i) f(s|m = m_i) ds \\ = & E[E[d(s,m)|s]] \\ = & \int_{\mathbf{S}} \sum_{m_i \in M} d(s,m_i) P\{m = m_i|s\} f(s) ds \end{aligned} \quad (4)$$

It is thus seen that the closeness of m and s depends on $w_i(s) = P\{m = m_i|s\}$, the model probability conditioned on the true mode s. A study of the conditional probability $P\{m = m_i|s\}$ will be reported later.

In this paper, for simplicity, we assume

$$P\{m = m_i|s\} = 1(s; S_i) = \begin{cases} 1 & s \in S_i \\ 0 & s \notin S_i \end{cases} \quad (5)$$

This is equivalent to assuming that

$$\{\{m = m_1\}, \ldots, \{m = m_N\}\} = \{\{s \in S_1\}, \ldots, \{s \in S_N\}\}$$

is a partition of the mode space $\mathbf{S}$; that is, each model covers a subset (region) of the mode space *exclusively*, which is often so perceived in practice. With this assumption, (4) becomes

$$E[d(s,m)] = \sum_i \int_{S_i} d(s,m_i) f(s) ds \quad (6)$$

We now present several general results under this assumption.

Theorem 6.1 (Optimality conditions of model set). Assume that $\mathcal{S} = \{S_1, \ldots, S_N\}$ is a partition of the mode space $\mathbf{S}$, where S_i is covered by model m_i *exclusively* in the sense $\{s \in S_i\} = \{m = m_i\}$. Then, the following conditions hold for the optimality in the sense of minimizing distance metric $\bar{d}(s,m)$ defined above.

A. Given any partition $\mathcal{S} = \{S_1, \ldots, S_N\}$ of mode space $\mathbf{S}$, a model set $M = \{m_1, \ldots, m_N\}$ is optimal if each model m_i is a (generalized) centroid of the corresponding partition member S_i:

$$m_i = s_i^* \triangleq \arg\min_m E[d(s,m)|s \in S_i] \quad (7)$$

B. Given any model set M, a partition is optimal if and only if points in any partition member S_i are closer to m_i than to any other $m_j \in M$ almost surely:

$$\begin{aligned} S_i = \{ & s : d(s,m_i) < d(s,m_j), \\ & \forall m_j \neq m_i, m_i, m_j \in M\} \end{aligned}$$

that is, a point s must be assigned to its nearest neighbor m_i among all $m \in M$; further, the set of equal-distance points may be assigned to either S_i or S_j:

$$S_{ij} = \{s : d(s,m_i) = d(s,m_j) \leq d(s,m_k), \forall m_k \in M\}$$

Remarks. (a) This theorem basically states that under the stated assumption, if exists, the optimal model set is within the class in which models are located at the (generalized) centroids of members of a nearest-neighbor partition of the mode space. (b) The generalized centroid reduces to the conditional mean (i.e., the centroid (mean) of S_i) $s_i^* = E[s|s \in S_i]$ if $d(s,m) = (s-m)'(s-m)$ or the conditional median (i.e., the median of S_i) if $d(s,m) = |s-m|$. (c) Both conditions are quite intuitive. (d) This theorem does not address the issue whether an optimal model set that minimizes the above metric is existent or unique, or whether a solution that meets conditions A and B is existent or unique. (e) The optimality conditions of this theorem actually hold for closeness metrics more general than defined above.

Most importantly, this theorem provides a theoretical basis for iteration procedures to find an optimal model set under the stated assumption. For example, we may start with an initial partition of mode space; find a candidate of the model set as the (generalized) centroid of each partition member; use the nearest-neighbor rule to obtain the corresponding (updated) partition; and repeat this process until convergence. Alternatively, we may start with an initial model set; use the nearest-neighbor rule to obtain the corresponding partition; obtain an update of the model set as the (generalized) centroid of each partition member; and repeat this process until convergence.

The above centroid model set has several nice and intuitive properties, as presented in the next theorem.

Theorem 6.2 (Properties of optimal model set). Any model set that covers each S_i by its centroid $m_i = E[s|s \in S_i]$ exclusively (i.e., $\{s \in S_i\} = \{m = m_i\}$) has the following properties:

(a) The (random) model and mode have the same mean: $E[m] = E[s]$.

(b) The modeling error is orthogonal to model: $E[m(s-m)'] = 0$.

(c) $E[ms'] = E[sm'] = E[mm']$ and thus $E[m's] = E[s'm] = E[m'm]$, meaning that cross power of the mode and model is equal to the power of the model.

(d) $E[(s-m)(s-m)'] = E[ss'] - E[mm']$ and thus $E[(s-m)'(s-m)] = E[s's] - E[m'm]$, meaning that minimum MSE is the power of the mode minus the power of the (optimal) model.

(e) $E[s(s-m)'] = E[(s-m)(s-m)']$ and thus $E[s'(s-m)] = E[(s-m)'(s-m)]$.

Remarks. (a) It follows from Theorem 6.1 that given a partition of the mode space, a model set that covers S_i by $m_i = E[s|s \in S_i]$ exclusively is optimal in the sense of minimizing MSE matrix $E[(s-m)(s-m)']$ and thus minimizing MSE scalar $E[(s-m)'(s-m)]$. (b) Property (b) is actually orthogonality principle for optimal linear estimation. The model m as an estimator of mode s appears to be not necessarily linear, but it turns out to be linear under the stated assumption [see proof of (b)].

7 Moment-Matching Design

In some practical situations, some moments, but not the complete distribution, of the true mode s are known. In some other situations, we do not have a good knowledge of a proper tolerance $|F_s(x) - F_m(x)| \le \epsilon$, but only want to match the moments of m to the known moments of s.

Given up to the qth moments of s, we want to find a discrete random variable m (i.e., the number and locations of points m_i with the associate probability mass p_i) such that

$$E[m^n] = E[s^n], \quad n = 1, \ldots, q$$

Several questions arise immediately. For example, what is the minimum number of models such that up to the qth moments of s and m are matched? How to design the corresponding pmf (locations m_i and probability masses p_i) of m? Given the number of models $|M|$, how to design pmf of m that matches as many as possible the lowest moments of s? For simplicity, we will consider only matching mean and covariance in this section since it is the common practice.

Let the pmf of m be

$$\begin{aligned} p_i &= \{m = m_i | m \in M\} > 0, \\ \forall i &\in J = \{1, \ldots, |M|\}, \quad M = \{m_1, \ldots, m_{|M|}\} \end{aligned}$$

Then, the mean and covariance of m are

$$\bar{m} = \sum_{i \in J} m_i p_i, \quad C_m = \sum_{i \in J} (m_i - \bar{m})(m_i - \bar{m})' p_i$$

7.1 Minimum Model-Set Design

The following theorem answers the first question above for $q = 2$.

Theorem 7.1 (Minimum models). The minimum number of models needed for m to match the *mean $\bar{s}$ and covariance C_s* of the true mode s is rank of C_s plus one: minimum number of models = $\text{rank}(C_s) + 1$.

Now consider the problem of design $\{m_i^*, p_i, i \in J\}$ such that

$$\begin{aligned} &\sum_{i \in J} p_i = 1, \quad \sum_{i \in J} m_i^* p_i = \bar{s}, \\ &\sum_{i \in J} (m_i^* - \bar{s})(m_i^* - \bar{s})' p_i = C_s \end{aligned} \tag{8}$$

In fact, we only need to design $\{m_i, p_i, i \in J\}$ such that

$$\sum_{i \in J} p_i = 1, \quad \sum_{i \in J} m_i p_i = \mathbf{0}, \quad \sum_{i \in J} m_i m_i' p_i = I_{n \times n} \tag{9}$$

where $n = \text{rank}(C_s)$. All designs presented below are for this standard problem. Given a problem with known mean $\bar{s}$ and covariance C_s, the design $\{m_i, p_i, i \in J\}$ can be converted to design $\{m_i^*, p_i, i \in J\}$ by $m_i^* = A[m_i', \mathbf{0}]' + \bar{s}$, which satisfies (8), where $C_s = A\,\text{diag}(I_{n \times n}, \mathbf{0})A'$.

Theorem 7.2 (Minimal-set design). The design $\{m_i, p_i\}_{i=0}^{n+1}$ with

$$\begin{aligned} 0 &\le p_0 < 1, \quad p_0^1 = p_0, \quad p_1^1 = p_2^1 = (1-p_0)/2 \\ m_0^1 &= 0, \; m_1^1 = (1-p_0)^{-1/2}, \; m_2^1 = -(1-p_0)^{-1/2} \\ &\vdots \\ p_0^j &= p_0, \quad p_i^j = p_i^{j-1}/2, \quad i = 1, \ldots, j, \\ p_{j+1}^j &= (1-p_0)/2 \\ m_0^j &= \mathbf{0}, \quad m_i^j = \left[(m_i^{j-1})', (1-p_0)^{-1/2}\right]', \\ &\qquad i = 1, \ldots, j, \; m_{j+1}^j = \left[\mathbf{0}, -(1-p_0)^{-1/2}\right]' \end{aligned}$$

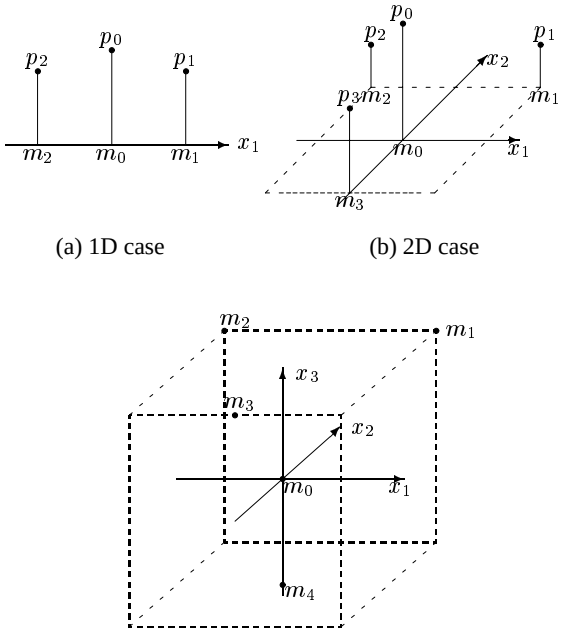
satisfies (9), where $m_i = m_i^n, p_i = p_i^n, i = 0, 1, \ldots, n+1$, and the superscript denotes dimension of a vector.

Fig. 5 illustrates this design with a minimal model set for $n = 1$, 2, and $n = 3$, respectively. For $n = 3$, m_0 is at the center of the cube, while all other models are on the surface of the cube; m_4 is at the center of the bottom square. Note that the coordinates of every model are either 0 or $\pm(1-p_0)^{-1/2}$. The mean and covariance are matched by the probability mass: $\sum_{i=1}^{j} p_i^j = p_{j+1}^j$.

Corollary. In this design, $0 \le p_0 < 1$ is a free parameter for us to choose. If we choose $p_0 = 0$ (i.e., delete m_0), we actually have $n+1$ models, which by Theorem 7.1 is the smallest possible number of models to match mean and covariance.

Remarks. (a) Although the model m_0 is not needed to match mean and covariance, in practice, such a model located at the expected true mode is usually very beneficial for MM estimation. (b) The value of p_0 affects higher order moments — a greater p_0 implies that the distribution of m is more concentrated around the mean. (c) This minimal-set design depends very much on not only the choice of the coordinate system but also the artificial labeling of each coordinate (e.g., the locations and the probability masses of the models would vary if x_1 and x_3 of Fig. 5(c) were interchanged). The latter dependence is entirely artificial and is better eliminated, while the former dependence is inevitable because the coordinate directions (after transformation from m^* to m) are actually eigenvector directions.

Minimal-set designs are not unique. Fig. 6 illustrates another simple minimal-set design in the 3D case. Its extension to a higher dimension is straightforward. In



(a) 1D case (b) 2D case

(c) 3D case

Fig. 5: Illustration of a minimal-set design.

this design, a model with probability mass p is placed on each positive semi-axis of equal distance α from the origin (i.e., $m_i = \alpha e_i$, $\forall i$, where $e_i = [\mathbf{0}_{1\times(i-1)}, 1, \mathbf{0}_{1\times(n-i)}]'$ is the ith coordinate vector); the last model is $m_{n+1} = \beta[-1, -1, \ldots, -1]'$ with probability mass q. It is clear that the mean and covariance are $\bar{m} = 0$ and $C_m = I_{n\times n}$ if $q = p$ and $\alpha = \beta = 1/\sqrt{p}$. As for the design of Fig. 5, if desirable, an additional model may be placed at the origin with probability p_0 without affecting mean and covariance. Then $p = (1 - p_0)/(n + 1)$. In Sec. 7.3, $q \neq p$ and $\alpha \neq \beta$ are chosen to obtain a minimal set with an equal distance between models.

The minimal-set design of Fig. 5 has attractive features that the model locations and probability mass are determined recursively as dimension increases and that all self skewnesses are equal to zero in the design of Fig. 5: $E\{[m(j) - \bar{s}(j)]^3\} = 0, \forall j \leq n$.

7.2 Symmetric Model-Set Design

All the above minimal-set designs clearly have an asymmetrical distribution spatially and possibly probabilistically. For many applications in practice, it is appealing that the models are symmetrically distributed and invariant to the artificial labeling of coordinates. For this reason, we present the following theorem.

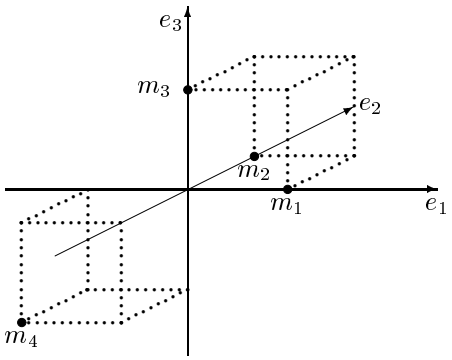


Fig. 6: Illustration of another minimal-set design.

Theorem 7.3 (Minimal symmetric-set design). The design $\{m_i, p_i\}_{i=0}^{2n}$ with the following *symmetric* distribution

$$\begin{aligned} 0 \quad &\leq \quad p_0 < 1, \quad p_i = (1 - p_0)/(2n), \quad i = 1, \ldots, 2n \\ m_0 \quad &= \quad \mathbf{0},\ m_i = -m_{n+i} = e_i \sqrt{\frac{n}{1 - p_0}}, \quad i = 1, \ldots, n \end{aligned}$$

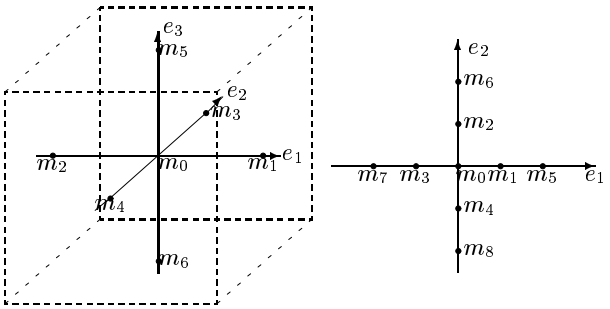
satisfies (9), where $e_i = [\mathbf{0}_{1\times(i-1)}, 1, \mathbf{0}_{1\times(n-i)}]'$ is the ith coordinate vector.

As for the design of Theorem 7.1, $0 \leq p_0 < 1$ is a free parameter for us to choose whose value affects higher-order moments. If we choose $p_0 = 0$ (i.e., delete m_0), we actually have $2n$ models. In practice, however, the use of model m_0 is usually very beneficial for MM estimation.

Fig. 7(a) illustrates this symmetric-set design for $n = 3$, where m_0 is at the center of the cube, while all other models are at the center of a boundary square of the cube. Note that if m_0 is not used, all models are located symmetrically on an axis (representing an eigenvector direction) with an equal distance from the origin; thus, the mean is matched provided an equal probability mass is assigned to all models and the covariance is matched by such a special assignment that all models on each axis have a total contribution of 1 to the covariance.

In this design, there are only two models along each axis direction, excluding m_0. In many applications, more models are needed for an MM estimator to perform well. Therefore, we present the following extension of Theorem 7.3.

Theorem 7.4 (Symmetric-set design). The design



(a) 3D case (b) 2D case

Fig. 7: Illustration of a symmetric-set design.

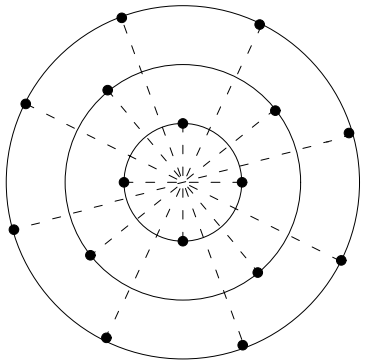


Fig. 8: Illustration of a more evenly distributed model-set design.

$\{m_i, p_i\}_{i=0}^{2kn}$ with the following *symmetric* distribution

$$\begin{aligned} 0 \;\le\; & p_0 < 1, \\ & p_{2(j-1)n+i} = p_{(2j-1)n+i} = (1-p_0)/(2\alpha_j\beta_j n) \\ & \quad i = 1,\ldots,n, \quad j = 1,\ldots,k \\ m_0 \;=\; & \mathbf{0},\; m_{2(j-1)n+i} = -m_{(2j-1)n+i} = e_i\sqrt{\frac{\alpha_j n}{1-p_0}}, \\ & i = 1,\ldots,n, \quad j = 1,\ldots,k \end{aligned}$$

satisfies (9), where $\alpha_k \ge \alpha_{k-1} \ge \cdots \ge \alpha_1 > 0$ and $\beta_j > 0$ satisfy

$$\sum_{j=1}^{k} \frac{1}{\beta_j} = 1, \quad \sum_{j=1}^{k} \frac{1}{\alpha_j\beta_j} = 1$$

A simple and meaningful choice for α_j and β_j is

$$\alpha_j = j\alpha_1, \quad \beta_j = k, \quad j = 1,\ldots,k$$

which yields $\alpha_1 = 3/4$, $\alpha_2 = 3/2$ for $k = 2$, and $\alpha_1 = 11/18$, $\alpha_2 = 22/18$, $\alpha_3 = 33/18$ for $k = 3$. Fig. 7(b) illustrates this symmetric-set design for $n = 2$ and $k = 2$.

A possible drawback of this symmetric design is that the models are distributed highly unevenly in space, albeit symmetrically. We now present a design that is much more evenly distributed. This can be accomplished by rotating models $m_{2(j-1)n+i}$ and $m_{(2j-1)n+i}$ for $j \ge 2$ such that they are more evenly distributed. We only consider $n = 2$ and $k = 4$ with $\alpha_4 = \alpha_3$, as shown in Fig. 8. It can be extended to the general case.

Let

$$\begin{aligned} e_i^1 \;=\; & e_i, \quad e_i^2 = (e_i^1 + e_{i+1}^1)/\sqrt{2},\; e_i^3 = (e_i^1 + e_i^2)/\sqrt{2}, \\ e_i^4 \;=\; & (e_{i+1}^1 + e_i^2)/\sqrt{2},\; e_{2+i}^j = -e_i^j \end{aligned}$$

for $i = 1, 2$ and $j = 1, 2, 3, 4$. Note first that a key to the design of Theorem 7.4 is

$$\begin{aligned} \mathrm{cov}(m) &= \sum_{i=1}^{n} \mathrm{diag}\left(\mathbf{0}_{(i-1)\times(i-1)}, \sum_{j=1}^{k}\frac{1}{\beta_j}, \mathbf{0}_{(n-i)\times(n-i)}\right) \\ &= I\sum_{j=1}^{k}\frac{1}{\beta_j} = \sum_{j=1}^{k}\frac{1}{\beta_j}[e_1,\ldots,e_n] \\ &= \left[\sum_{j=1}^{k}\frac{1}{\beta_j}e_1,\ldots,\sum_{j=1}^{k}\frac{1}{\beta_j}e_n\right] \end{aligned}$$

Similarly, we may use

$$\mathrm{cov}(m) = \left[\sum_{j=1}^{k}\frac{1}{\beta_j}e_1^j,\ldots,\sum_{j=1}^{k}\frac{1}{\beta_j}e_n^j\right]$$

In our simple case with $n = 2$ and $k = 4$, it becomes

$$\mathrm{cov}(m) = \left[\frac{e_1^1}{\beta_1} + \frac{e_1^2}{\beta_2} + \frac{e_1^3}{\beta_3} + \frac{e_1^4}{\beta_4},\; \frac{e_2^1}{\beta_1} + \frac{e_2^2}{\beta_2} + \frac{e_2^3}{\beta_3} + \frac{e_2^4}{\beta_4}\right]$$

We may choose

$$\begin{aligned} & \alpha_1 = 13/18,\; \alpha_2 = 2\alpha_1,\; \alpha_4 = \alpha_3 = 3\alpha_1, \beta_j = 4, \\ & \qquad j = 1,\ldots,4 \\ & 0 \le p_0 < 1,\, p_{4(j-1)+i} = p_{2(2j-1)+i} = (1-p_0)/(16\alpha_j) \\ & \qquad i = 1,2, \quad j = 1,\ldots,4 \end{aligned}$$

Note, however, that while this design has zero mean, its covariance is no longer equal to the identity matrix.

7.3 Equal-Distance Model-Set Design

The above symmetric-set designs do not have an even model distribution in space. In practice, it is sometimes desirable to have a set of models that are evenly distributed.

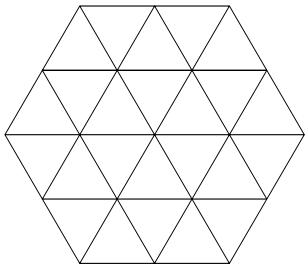


Fig. 9: Illustration of a diamond model-set design.

For instance, this may be the case when each model is considered to be able to cover a region of the same size.

Diamond set. For this purpose, consider the diamond-set design illustrated in Fig. 9 for the 2D case. Note that the set of models on the whole diamond consists of hexagonal layers of models: 1 at the center (0th layer), 6 on the first layer (i.e., those on the unit circle), 12 on the second layer (6 of them are on the circle of radius 2), 18 on the third layer (6 of them are on the circle of radius 3), and so on. Alternatively, the model set may also be viewed as consisting of even finer (circle) layers of models: Models on each layer have equal distance from the origin (i.e., are on a circle of radius $0, 1, \sqrt{3}, 2, \sqrt{7}, 3, 2\sqrt{3}, \sqrt{13}, 4$, and so on, respectively). In general (the square of) the radii of these circles are given by

$$r_{ij}^2 = \begin{cases} (i\sqrt{3}/2)^2 + (\frac{2j-1}{2})^2 & i \text{ odd},\ 1 \le j \le \frac{i+1}{2} \\ (i\sqrt{3}/2)^2 + (j-1)^2 & i \text{ even},\ 1 \le j \le \frac{i}{2}+1 \end{cases}$$

where the double subscript ij stands for the jth circle that passes through the models on the ith hexagonal layer, for example:

$$\begin{aligned} r_{11}^2 &= (1\sqrt{3}/2)^2 + (1/2)^2 = 1 \\ r_{21}^2 &= (2\sqrt{3}/2)^2 + 0^2 = 3, \\ r_{22}^2 &= (2\sqrt{3}/2)^2 + 1^2 = 2^2 \\ r_{31}^2 &= (3\sqrt{3}/2)^2 + (1/2)^2 = 7, \\ r_{32}^2 &= (3\sqrt{3}/2)^2 + (3/2)^2 = 3^2 \\ r_{41}^2 &= (4\sqrt{3}/2)^2 + 0^2 = 12, \\ r_{42}^2 &= (4\sqrt{3}/2)^2 + 1^2 = 13, \\ r_{43}^2 &= (4\sqrt{3}/2)^2 + 2^2 = 4^2 \end{aligned}$$

Clearly, this diamond set is symmetric and has equal distance between any two adjacent models. Furthermore, the following theorem states that this diamond-set design can also be used to match arbitrarily given mean and covariance of the mode by simply assigning each model on the same (hexagonal or circle) layer equal probability.

Theorem 7.5 (Diamond-set design). Consider a diamond-set design as illustrated in Fig. 9. Assign each model on the lth (hexagonal or circle) layer an equal probability p_l such that all probability masses sum up to unity. Let the total contribution to the covariance from the models on the lth layer be C_l. Then this diamond-set design satisfies (9) if $\sum_{l=1}^{k} C_l = I$, where k is the number of layers.

Remark. In particular, p_l and C_l can be chosen so that $C_l = C_r = I/k$ and every model has the same probability or the total probability mass of models on different layers are equal.

The simplest possible diamond-set design (with one at the center and six on the first layer) was implemented in [19] for an example of maneuvering target tracking using MM algorithms.

There are many equal-distance sets. In 3D for example, the well-known regular tetrahedron, cube, regular octahedron, regular dodecahedron, and regular icosahedron each leads to an equal-distance set design by placing a model at every vertex. However, the above diamond-set design is, on top of its regularity, attractive for several other nice properties, such as the ease for design (as stated in Theorem 7.5) and its economy in the sense of using a small number of models to cover a large region.

In reality, each model is effective only over a finite region. Call this region the *effective coverage region* of the model. Two natural questions are: Given the mode space $\mathbf{S}$ and the effective coverage region R_m of each model, what is the minimum number of models needed and where should the models be placed? Clearly, a lower bound on the number of models needed is $|M| \ge V_{\mathbf{S}}/V_m$, where $V_{\mathbf{S}}$ and V_m are the volumes of $\mathbf{S}$ and R_m, respectively. Assume that $\mathbf{S}$ and R_m are (n-dimensional) balls of radii $r_{\mathbf{S}}$ and r_m, respectively. Consider a diamond set in which for every diamond cell, each cell vertex to the cell center is r_m. Then every point in the inscribed ball B of the union of all models' R_m is covered by at least one R_m. It appears that this diamond set covers B using model coverage regions with the smallest number of models in general.

More generally, the diamond set has a small Hausdorff distance to the mode space relative to other (equal-distance) sets of the same number of models (vertices). For two (finite) sets A and B with a distance metric $d(x, y)$, $x \in A$ and $y \in B$, the Hausdorff distance between A and B is defined as $d(A, B) = \max\{\rho(A, B), \rho(B, A)\}$, where $\rho(A, B) = \sup_{x \in A} \inf_{y \in B} d(x, y)$. Note that the use of Hausdorff distance here — which corresponds to the worst case in distance between model and mode — is more reasonable than the more popular distance between two sets: $d(A, B) = \inf_{x \in A, y \in B} d(x, y)$ (which corresponds to the best case and is often zero for model-set design).

Equal-distance minimal-set design. The diamond set has many nice features, but it is not a minimal set. A minimal set with equal distance between models can be obtained by the minimal-set design of Fig. 6 with a special choice of $\{p, q, \alpha, \beta\}$ such that all models are separated by

an equal distance. Clearly, $m_1, \ldots, m_n$ have an equal distance of $\sqrt{2}\alpha$. So, we need only to place m_{n+1} in a place such that its distance to every model in $\{m_1, \ldots, m_n\}$ is $\sqrt{2}\alpha$. Specifically, choose the set $\{p, q, \alpha, \beta\}$ of nonnegative numbers to satisfy

$$\begin{aligned} np + q + p_0 &= 1 \quad \text{(unity probability)} \\ \alpha p - \beta q &= 0 \quad \text{(zero mean)} \\ \alpha^2 p + \beta^2 q &= 1 \quad \text{(identity covariance)} \\ (\alpha+\beta)^2 + \beta^2(n-1) &= 2\alpha^2 \quad \text{(equal distance)} \end{aligned}$$

(the last equation above follows from setting $\|m_i - m_{n+1}\|^2 = \|m_i - m_j\|^2, \forall i, j \leq n$), which yields

$$\begin{aligned} q &= \frac{1-p_0}{\sqrt{n+1}}, \quad p = \frac{q+p_0-1}{n}, \\ \alpha &= \frac{1}{\sqrt{p(1+p/q)}}, \quad \beta = \frac{p}{q}\alpha \end{aligned}$$

Then, the design of Fig. 6 has a minimal set that satisfies (9) and has equal distance between models. As such, this design places a model at each vertex of a convex $(n+1)$-hedron with equal edge length $\sqrt{2}\alpha$ (e.g., an equilateral triangle in 2D and a regular tetrahedron in 3D). Note, however, that m_{n+1} is closer to the origin than m_i $(i \leq n)$ (i.e., the polyhedron is not centered at the origin) because $\|m_{n+1} - 0\|^2 = n\beta^2 < \alpha^2 = \|m_i - 0\|^2$.

8 Model Efficacy

Each model has a certain *relative effective coverage* region of the true mode within the model set in use. In this section, we introduce the concept of **relative efficacy** of the coverage of a model, along with its quantitative measures. Specifically, we introduce a window function $w_i(s)$ to quantify the efficacy of model m_i in covering the true mode s relative to other models in the set. The larger the $w_i(s)$ is, the more effective (i.e., perform better) the model m_i is (relative to other models in the set) given s. Knowledge about such relative efficacy is quite useful in model-set design.

The prior (unconditional) model probabilities $P\{m = m_i | m \in M\}$ are used as the initial model probabilities in an MM algorithm. We would like to determine these probability from a probabilistic description (pdf or pmf) of the true mode s given the model set M. This is essentially a problem of how to define the pmf of a discrete random variable with a given sample space so that it best approximate a given random variable with a larger sample space. To our knowledge, however, there is no generally acceptable solution to this problem, although the theoretical results of Secs. 5, 6, and 7 are indeed applicable. As a by-product, the prior model probabilities $P\{m = m_i | m \in M\}$ can be obtained from the relative efficacy.

8.1 Probability-Based Efficacy

Given a model set $M = \{m_1, m_2, \ldots, m_{|M|}\}$, the corresponding pmf can be determined using the above window function $w_i(s)$:

$$P\{m = m_i | m \in M\} = \int_{\mathbf{S}} w_i(s) f(s) ds \qquad (10)$$

where $w_i(s)$ are in general functions of $m_1, m_2, \ldots, m_{|M|}$ and satisfy

$$1 = \sum_{i=1}^{|M|} P\{m = m_i | m \in M\} = \sum_{i=1}^{|M|} \int_{\mathbf{S}} w_i(s) f(s) ds$$

which is guaranteed by the following requirements:

$$\begin{aligned} w_i(s) &\geq 0, \ \sum_{i=1}^{|M|} w_i(s) = 1, \qquad (11) \\ \forall s &\in \mathbf{S}, \ i = 1, \ldots, |M| \end{aligned}$$

Additionally, it is desirable to have $w_i(m_i) = \sup_{s \in \mathbf{S}} w_i(s)$, except possibly for the models on the boundary. Fig. 10 depicts such a family of window functions.

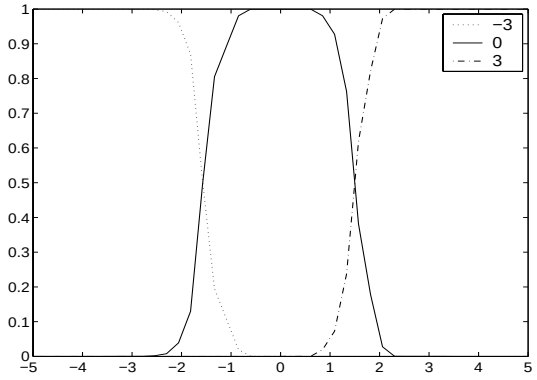


Fig. 10: Relative efficacy $w_i(s)$ of each model in set $M = \{0, \pm 3^\circ/\text{s}\}$.

By Bayes' rule, we have the following conditional pdf

$$f(s | m = m_i) = \frac{f(s) w_i(s)}{\int_{\mathbf{S}} f(s) w_i(s) ds}, \quad \forall s \in \mathbf{S}$$

This leads to, dropping conditioning on $m \in M$,

$$\begin{aligned} P\{m = m_i | s\} &= \frac{1}{f(s)} f(s | m = m_i) P\{m = m_i\} \\ &= w_i(s) \qquad (12) \end{aligned}$$

This result also follows from (10) and the fact

$$P\{m = m_i | m \in M\} = \int_{\mathbf{S}} P\{m = m_i | s, m \in M\} f(s) ds$$

It is thus seen that the *relative efficacy* of the coverage by model m_i given the true mode s [i.e., the window function $w_i(s)$] can be defined to be equal to the probability of model m_i (in the model set) given s. Also, it follows from (10) that the initial probability $P\{m = m_i | m \in M\}$ is then in fact the average (expected) efficacy of the coverage by model m_i relative to other models in the set.

This equation also makes it explicit that the initial model probabilities depend on the distribution of the true mode.

8.2 Test-Based Efficacy

Alternatively, relative efficacy of model coverage can be defined based on hypothesis testing as follows. Consider testing (optimally)

$$H_1 : m = m_1 \quad \text{vs.} \quad \cdots \quad \text{vs.} \quad H_{|M|} : m = m_{|M|} \quad (13)$$

using all available data $z(s)$, which is a function of the true mode s. For a given s, the probability that H_i is not rejected by an (optimal) test is taken to be the relative efficacy of m_i. More precisely,

$$w_i(s) = P\{H_i \text{ not rejected} | s\}/L$$

where L is the number of hypotheses that are not rejected at the end of the test, and

$$\begin{aligned} & P\{H_i \text{ not rejected} | s\} \triangleq E[P\{H_i \text{ not rejected} | z(s)\} | s] \\ = & \int P\{H_i \text{ not rejected} | z(s)\} f(z|s) dz \end{aligned}$$

where $P\{H_i \text{ not rejected} | z(s)\}$ is the probability that H_i is not rejected given data $z(s)$. In practice, the probability $P\{H_i \text{ not rejected} | s\}$ is approximated by relative frequency. Even more precisely, we define

$$w_i(s) = \int \frac{1}{L(z)} P\{H_i \text{ not rejected} | z(s)\} f(z|s) dz \quad (14)$$

because the number L of hypotheses not rejected at the end of the test actually depends on the observation z.

Note that although theoretically equivalent, we do not define

$$w_i(s) = P\{\text{accept } H_i | s\}$$

mainly for implementation considerations. First, compared with a test of a fixed sample size, a sequential test, such as those proposed in [15], appears more appropriate for several reasons, for example: (a) superiority in efficiency, (b) symmetry in the roles the decision errors played[2], and (c) ease at determining (approximate) decision thresholds. However, a sequential test will not necessarily accept a hypothesis in all cases (e.g., when two models are located symmetrical about the true mode). Second, $w_i(s)$ is quite small for most values of $s \in \mathbf{S}$. This means that $P\{\text{reject } H_i | s\} \gg P\{\text{accept } H_i | s\}$ for most $s \in \mathbf{S}$ and thus the former definition is superior in terms of efficiency and accuracy if a sequential test is used since $P\{H_i \text{ not rejected} | s\} = 1 - P\{H_i \text{ rejected} | s\} \neq P\{\text{accept } H_i | s\}$.

The rationale for division of L in (14) is that we do not rank the hypotheses that are not rejected at the end of the test using finite samples and thus we may think they are equally likely to be true. Alternatively, we may perform a test among the hypotheses not rejected. The drawback is that the results of such a test tends to be not reliable.

Note that the model efficacy so defined also has the properties of (11).

8.3 Determination of Model Efficacy

We will use $w_i^P(s)$ and $w_i^T(s)$ to denote probability-based and test-based model efficacies, respectively. Clearly, the above definitions of model efficacy are valid for vector-valued s. They are in fact also applicable to the cases where s is not defined over a metric space. In these cases, the model efficacies $w_i(s)$ for different s are simply a set of unordered (isolated) probabilities.

The test-based model efficacy $w_i^T(s)$ can be computed as follows. For each given value of s, generate N samples of the corresponding data z by randomly generating process and measurement noise; for each sample of data z, run a sequential procedure (e.g., those of [15]) to test hypotheses (13); finally, $w_i^T(s)$ is computed by (14) using, for $i = 1, \ldots, |M|$,

$$w_i^T(s) \approx \frac{1}{N} \sum_{j=1}^{N} 1_{i,j}/L_j \quad (15)$$

where L_j is the number of hypotheses not rejected at the end of the test in the jth run, and $1_{i,j} = 0$ if H_i is rejected in the jth run and 1 otherwise. Note that s is fixed for all times when generating z. The error probability (common to all hypotheses) used in the test can be viewed as the error probability (or $1-$ confidence) of the model efficacy.

We now consider the determination of the probability-based model efficacy. Note first that

$$\begin{aligned} w_i^P(s) &= P\{m = m_i | s, m \in M\} \\ &= E\left[P\{m = m_i | z, s, m \in M\} | s\right] \\ &= E\left[P\{m = m_i | z(s), m \in M\} | s\right] \\ &= \int P\{m = m_i | z(s), m \in M\} f(z|s) dz \end{aligned}$$

This provides a theoretical basis for the following general method of obtaining the relative efficacy (i.e., window function) $w_i(s)$ via Monte-Carlo simulation: For every fixed true mode s, generate a random sample of measurement $z_1, \ldots, z_N$ to compute the model probabilities

[2] E.g., for a binary test, type I and type II errors are symmetric in some sequential tests and Bayes tests, but not in Neyman-Pearson type tests. However, it is often too subjective in assigning cost associated with a decision error in a Bayes test.

$P\{m = m_i|z_j, m \in M\}, j = 1, \ldots, N$ using model set $M = \{m_1, m_2, \ldots, m_{|M|}\}$ with the initial model probabilities given by (10). Then, we have

$$w_i^P(s) \approx \frac{1}{N}\sum_{j=1}^{N} P\{m = m_i|z_j, m \in M\}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \frac{f(z_j|m = m_i)}{f(z_j)} P\{m = m_i|m \in M\} \quad (16)$$

where $f(z_j|m = m_i)$ is the mode m_i likelihood. Note, however, that $w_i(s)$ so determined depends on the initial model probabilities $P\{m = m_i|m \in M\}$, whose computation by (10) presumes knowledge of $w_i(s)$. In view of this, to be more accurate, an iteration can be used: Once $w_i(s)$ is obtained as above, the initial model probabilities is updated by the use of (10) via numerical integration; then $w_i(s)$ is updated again and the process is repeated. This process can be started with some set of initial model probabilities, such as the one given below using a rectangular window. Fortunately, it is our experience that the model efficacy is not sensitive to the initial model probabilities because the likelihoods dominate.

We emphasize that all z in the above determination of $w_i^T(s)$ and $w_i^P(s)$ are observations at a fixed time (i.e., not time sequences) at which s is the true mode of the system in effect. If time sequences of observations z^k are used, the procedure would lead to $w_i^T(s^k)$ and $w_i^P(s^k)$, which are not addressed in this paper even if s^k is only a sequence of constant s.

For process (state or signal) estimation, the above results still hold for the case where s_k and m_k are not allowed to vary because

$$\begin{aligned} w_i(s_{k-1}) &= P\{m_{k-1} = m_i|s_{k-1}, m_{k-1} \in M_{k-1}\} \\ &= P\{m_k = m_i|s_k, m_k \in M_k\} \\ &= w_i(s_k) \end{aligned}$$

Note, however, that $w_i(s_k = a) \neq w_i(s^k)|_{s^k=(a,a,\ldots,a)}$. In the case where s_k or m_k may vary, the (marginal) model efficacy is actually a function of time k:

$$\begin{aligned} w_i(s_k|s^{k-1}) &= P\{m_k = m_i|s^k, m_k \in M_k\} \\ &\approx \frac{1}{N}\sum_{j=1}^{N} P\{m_k = m_i|z_j^k, m_k \in M_k\} \end{aligned}$$

where $P\{m = m_i|z_j^k, m \in M_k\}$ are the model m_i probabilities at time k obtained in an AMM estimator given the jth measurement sequence z_j^k; s_k and m_k stand for the true mode and model in effect at time k and s^k and z^k denote the true mode and measurement sequences through time k. Note that $w_i(s_k|s^{k-1})$ depends on the past true mode s^{k-1} in that z_j^k is generated by a fixed s^{k-1} but varying s_k. This time-varying efficacy has the following backward recursion:

$$\begin{aligned} & w_j(s_{k-1}|s^{k-2}) \\ &= P\{m_{k-1} = m_j|s^{k-1}, m_{k-1} \in M_{k-1}\} \\ &= \sum_{m_i \in M_k} P\{m_{k-1} = m_j|m_k = m_i, \\ &\quad s^{k-1}, m_{k-1} \in M_{k-1}\} P\{m_k = m_i|s^{k-1}, m_k \in M_k\} \\ &= \sum_{m_i \in M_k} P\{m_{k-1} = m_j|m_k = m_i, s^{k-1}, m_{k-1} \in M_{k-1}\} \\ &\quad \cdot \int_{\mathrm{S}} f(s_k|s^{k-1}) w_i(s_k|s^{k-1}) ds_k \end{aligned}$$

where the last equation above follows from a time-varying version of (10); $f(s_k|s^{k-1})$ is the pdf of s_k conditioned on s^{k-1}, which governs the transition of the true mode; and $P\{m_{k-1} = m_j|m_k = m_i, s^{k-1}, m_{k-1} \in M_{k-1}\}$ is the (backward) transition probability of model m_i to model m_j given true mode sequence s^{k-1}. Given these transition probabilities and pdf, the time-varying relative efficacy of each model $w_i(s_k|s^{k-1})$ can be computed at least in principle backward in time.

8.4 Simple Windows

One of the simplest classes of windows in the scalar case consists of rectangular windows, assuming m_i are arranged in an increasing order,

$$w_i(s) = \begin{cases} 1[s; (-\infty, (m_1 + m_2)/2]] & i = 1 \\ 1[s; ((m_{|M|-1} + m_{|M|})/2, \infty)] & i = |M| \\ 1[s; \frac{m_{i-1}+m_i}{2}, \frac{m_i+m_{i+1}}{2}] & \text{otherwise} \end{cases} \quad (17)$$

where $1[x; R]$ is the indicator function, defined by

$$1(x; R) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases}$$

These rectangular windows amount to assuming the following coverage regions S_i of models m_i:

$$\begin{aligned} S_1 &= \{s : -\infty < s \leq (m_1 + m_2)/2\} \\ S_{|M|} &= \{s : (m_{|M|-1} + m_{|M|})/2 \leq s < \infty\} \\ S_i &= \{s : (m_{i-1} + m_i)/2 < s \leq (m_i + m_{i+1})/2\}, \\ i &= 2, \ldots, |M| - 1 \end{aligned}$$

With these rectangular windows, we clearly have

$$P\{m = m_i|m \in M\} \quad (18)$$

$$= \begin{cases} \int_{-\infty}^{(m_1+m_2)/2} f(s)ds & i = 1 \\ \int_{(m_{|M|-1}+m_{|M|})/2}^{\infty} f(s)ds & i = |M| \\ \int_{(m_{i-1}+m_i)/2}^{(m_i+m_{i+1})/2} f(s)ds & \text{otherwise} \end{cases} \quad (19)$$

In particular, if m_i are determined by (2), these rectangular windows give the discrete uniform pmf given by (3).

This provides a justification of the model-set design (2): Its prior model probabilities is (discrete) uniformly distributed such that the mode space is partitioned into equally probable regions, each represented by a model.

The above results clearly can be extended to multi-dimensional cases straightforwardly.

The discrete uniform pmf (3) is widely used in practice. It is clear from the discussion so far in this paper that this practice is justified only when the models m_i are determined by (2), given a distribution of the true mode.

We may want to use some other simple window functions in practice. Often, $w_i(s)$ could be chosen to have a bell or trapezoidal shape centered at m_i, except for the end window functions (see results later for an empirical support). The use of such window functions increases the probabilities of the models located at high probability density areas, as compared with the rectangular windows. Although this windowing technique does resemble those or the design of digital FIR filters or spectral estimators, it should be noted that a window is chosen here to quantify the efficacy of a model. This differs vastly from the underlying criteria there.

If the model set $M = \{m_1, m_2, \ldots, m_{|M|}\}$ is quite dense (i.e., model separation is small), the model probabilities can be computed approximately by

$$P\{m = m_i | m \in M\} = f_s(m_i)/D, \;\; D = \sum_i f_s(m_i)$$

that is, the model probabilities are proportional to the pdf values. This follows from (19) as the model separation approaches zero.

9 Criteria and Measures for Model-Set Comparison, Choice, and Design

In this section, it is assumed that the mode space is actually a region (not necessarily a subspace) in a metric space in which the distance measure $\sqrt{E[(\cdot)'(\cdot)]}$, denoted by $\|\cdot\|$, is defined, where the expectation may be conditional; each model is a point in this metric space; and a model set M is a discrete set of models (modal points). Note that in general M need not be a subset of $\mathbf{S}$ since a model may be a *simplified* representation of a mode[3].

Since model sets differ from the mode space, the superiority of one model set to another one should be properly defined. The appropriateness of a definition should be judged by the ultimate goal of the MM estimation for the particular application. In other words, a variety of criteria and measures is reasonable; which one is more appropriate depends on what the ultimate goal is. The following measures and optimality criteria are introduced in view of the fact that MM estimation is usually used for (i) base-state estimation, (ii) mode identification, (iii) mode estimation, which amounts to *soft* identification of the mode, and/or (iv) hybrid estimation (i.e., simultaneous base-state/mode estimation and identification).

In practice, the size of a model set, that is, the number of models in the set, is of a major concern. Also, the more cluster-like the model set is, the better usually. This is, however, considered at most indirectly in most of the following measures and criteria.

The criteria and measures are presented below in a form that is not conditioned on data, which is applicable directly to an offline design. For an online model-set design or adaptation, their data-conditioned versions may be used.

9.1 For Base-State Estimation

A model m_i is said to *match* a mode better (or be a *better representation* of a mode) than model m_j if its model-based optimal estimator is a better estimator of the base state of the hybrid system given the system mode. A model set A is said to be better for base-state estimation than set B at a given time if

$$\|x - \hat{x}_A\|^2 < \|x - \hat{x}_B\|^2 \tag{20}$$

where x is the base state; $\hat{x}_A$ and $\hat{x}_B$ are the (optimal) estimators based on sets A and B, respectively. For a given mode s, $\|x - \hat{x}_A\|^2 = E[(x - \hat{x}_A)'(x - \hat{x}_A)|A, s]$, where E is over both x and measurement z at the time. In the case of an unknown mode s, $\|x - \hat{x}_A\|^2 = E[(x - \hat{x}_A)'(x - \hat{x}_A)|A]$, where E is over x, z, and s. A model set A may be deemed *uniformly better* than set B if (20) holds uniformly with respect to time. Note that, in general, there is no *uniformly optimal* model set if the exact set of all possible system modes (more rigorously, the exact digraph with the exact transitions) is not known perfectly. That is part of the reason why a variable-structure algorithm may be superior to a fixed-structure algorithm.

It is sometimes more tractable to replace x in (20) with some optimal estimator of x, such as the following:

$$\|\hat{x}_S - \hat{x}_A\|^2 < \|\hat{x}_S - \hat{x}_B\|^2 \tag{21}$$

where $\hat{x}_S$ is the **optimal base-state estimator** based on the mode space $S = \mathbf{S}$, (more precisely, $D = \mathbf{D}$, the digraph used is the exact total digraph with the exact transitions that corresponds to $\mathbf{S}$ [18]), that is

$$\hat{x}_S = \arg \inf_{\hat{x}_M} \|x - \hat{x}_M\|^2$$

and $\|\hat{x}_S - \hat{x}_M\|^2 = E[(\hat{x}_S - \hat{x}_M)'(\hat{x}_S - \hat{x}_M)|M, S]$. Such a definition was used in [18] to obtain a circular criterion for comparison of two model-sets: Set B is better than set A if and only if $\hat{x}_B$ falls into a circle determined by $\hat{x}_S$ and $\hat{x}_A$, A, and B. An example of model-set choice using this criterion is given later.

[3] In some cases a model may correspond to a point in the metric space outside $\mathbf{S}$.

Definitions based on other estimation criteria (e.g., maximum likelihood, maximum a posteriori or some other Bayesian costs) are also possible, but the above definition of $\|\cdot\|^2$ should be replaced by an appropriate one.

The above definitions are reasonable if the main purpose of the application is to obtain a base-state estimator that is as accurate as possible, which is often the case. It is, however, not convenient for some other purposes, such as mode identification, modal-state estimation and for model-set choice by hypothesis testing. For this reason, the following definitions are introduced, which are not entirely equivalent to the above.

9.2 For Mode Estimation

A model m_c is said to be *closest* to a mode s if their distance squared $\|s-m_c\|^2$ is the shortest among the set of models under consideration.

Given a collection of model sets *not necessarily disjoint*, the one with a model that is closest to a given mode could be deemed the best model set for that mode. With this definition, however,

- It is quite likely that two (or more) model sets have the best model for the mode and thus are both deemed the best. If this is indeed the case, the set with a higher correct-model probability (defined later) may be deemed better. This is still not good enough for many applications since the individual effects of the other models in the set are not accounted for, which may be important, especially when the correct-model probabilities for both sets are low.
- It is impossible that M_1 is better than M_2 if M_1 is a subset of M_2. This is not good since it does not encourage the use of a smaller but good model set, which is important in practice.

The following definition does not suffer from these two drawbacks: In a family of model sets $\mathcal{M}$, the set M^* with the smallest average modal distance squared between its models and the true mode is deemed the best, that is,

$$\|s-m\|^2_{m\in M^*, s\in\mathbf{S}} = \min_{M\in\mathcal{M}} \|s-m\|^2_{m\in M, s\in\mathbf{S}} \tag{22}$$

The *mean modal distance squared* of a model set M to a (random) mode s is defined by (4) as (dropping conditioning on $s\in\mathbf{S}$ and $m\in M$)

$$\begin{aligned} \|s-m\|^2 &= E[(s-m)'(s-m)] && (23)\\ &= E[E[(s-m)'(s-m)|s]]\\ &= E\left[\sum_{m_i\in M}(s-m_i)'(s-m_i)P\{m=m_i|s\}\right] && (24)\end{aligned}$$

Sec. 6 presents a method for model-set design that is particularly suitable for this measure. An example of model-set design by minimizing this average modal distance squared is given later.

For performance evaluation, it is more convenient to use the following equivalent formula

$$\begin{aligned}&\|s-m\|^2 = E[E[(s-m)'(s-m)|z,m]]\\ &= E\left[\int \sum_{m_i\in M}(s-m_i)'(s-m_i)P\{m=m_i|z\}f(z|s)dz\right]\end{aligned}$$

where z is measurement and the expectation is over s, and thus depends on the probability distribution of s. Its finite-sample approximation, called **average modal distance** *squared*, can be computed via Monte Carlo simulation over L samples of true mode $s_n\in\mathbf{S}$, $n=1,\ldots,L$, each with N samples of measurement z_{nj}, $j=1,\ldots,N$:

$$\begin{aligned}&\|s-m\|^2\\ &\approx \frac{1}{NL}\sum_{n=1}^{L}\sum_{j=1}^{N}\sum_{m_i\in M}(s_n-m_i)'(s_n-m_i)P\{m=m_i|z_{nj}\}\end{aligned} \tag{25}$$

where $P\{m=m_i|z_{nj}, m\in M\}$ is the posterior model probability, as obtained in an MM algorithm.

Alternatively, a model set M^* is said to induce the **optimal mode estimator** in a family of model sets $\mathcal{M}$ if its optimal estimator provides the most accurate **mode estimation** in the sense of having the smallest mean-square error:

$$\|s-\hat{s}_{M^*}\|^2 = \min_{M\in\mathcal{M}}\|s-\hat{s}_M\|^2 \tag{26}$$

The mode estimate using model-set M is defined by

$$\begin{aligned}\hat{s}_M &= E[s|s\in M, z]\\ &= \sum_{m_i\in M} m_i P\{s=m_i|s\in M,z\}\end{aligned} \tag{27}$$

where z depends on the true mode s to be estimated and the posterior model probability $P\{s=m_i|s\in M,z\}$ is available from an MM estimator using model set M. Here $\|s-\hat{s}_M\|^2 = E[(s-\hat{s}_M)'(s-\hat{s}_M)|s\in\mathbf{S}]$, where the expectation is over both s and z. A finite-sample approximation, called **average mode estimation error squared**, is the following based on Monte Carlo simulation over L samples of true mode $s_i\in\mathbf{S}$, $i=1,\ldots,L$, each with N samples of measurement z_{ij}, $j=1,\ldots,N$:

$$\|s-\hat{s}\|^2 \approx \frac{1}{NL}\sum_{i=1}^{L}\sum_{j=1}^{N}(s_i-\hat{s}_{ij})'(s_i-\hat{s}_{ij}) \tag{28}$$

where $\hat{s}_{ij}$ is the mode estimate (27) of s_i from the MM estimator using measurement z_{ij}.

Note that the replacement of the distance squared in (22) with the distance may lead to a different conclusion,

but the conclusion based on (26) is invariant with respect to such a replacement.

It may be inconvenient, difficult, or impossible to define $\|s - m_i\|^2$, modal distance, or mode estimation error reasonably for some practical problems if different models are characterized by different quantities, rather than different values of the same quantity. Mapping of s and m onto the same space will work only if the mappings are distance preservative.

Note that the best possible mode estimate $\hat{s}_{\mathbf{S}} = E[s|s \in \mathbf{S}, z] = \int_{\mathbf{S}} s dF(s|s \in \mathbf{S}, z)$ is usually infeasible because $\mathbf{S}$ is usually either too large or unknown; otherwise we may choose $M = \mathbf{S}$.

(25) and (28) have been used in a number of publications [22, 21, 20] for performance evaluation with deterministic scenarios (with $N = 1$) and random scenarios (with $L = 1$).

Definitions based on other estimation criteria (e.g., ML or MAP) are also possible, which are, however, usually less convenient.

9.3 For Mode Identification

In the sequel, whenever a set M is given, it is implicitly assumed that its model probabilities $P\{m = m_i|m \in M\}, \forall m_i \in M$, are also given by the optimal estimator based on set M.

A model m_p is said to be a **most probable** one in the set M if its probability is the largest in the set:

$$P\{m = m_p|m \in M\} = \max_{m_i \in M} P\{m = m_i|m \in M\} \tag{29}$$

Note that it is possible that more than one model may be most probable or closest to the mode since $m \in M$ is assumed for a finite M. Given a model set M, the probability $P\{m = m_c|m \in M\}$ may be called the **correct-model probability** of the set, where m_c is the model (in the set) closest to the mode s (defined in Sec. 9.2 or Sec. 9.6). Similarly, $1 - P\{m = m_c|m \in M\}$ may be referred to as the incorrect-model probability.

A comparison of two MM estimators *using the same model set* can also be made based on the probabilities of correct, incorrect, and no mode identification. These mode identification events are defined as follows:

- **Correct mode identification (CID)**: The model closest to the true mode has the highest probability and the ratios of its probability to the other model probabilities all exceed a threshold for mode identification.

- **Incorrect mode identification (IID)**: A model not closest to the true mode has the highest probability and the ratios of its probability to the other model probabilities all exceed the threshold.

- **No mode identification (NID)**: The ratio of the highest model probability to the second highest model probability does not exceed the threshold.

A simplified version is the following.

- Correct mode identification (CID): The model closest to the true mode has the highest probability that exceeds a threshold (say 0.5).

- Incorrect mode identification (IID): The model with the highest probability that exceeds the threshold is not the one closest to the true mode.

- No mode identification (NID): No model has a probability above the threshold.

For MM estimators, the one with the highest ratio of CID/IID, under the (approximately) same level of NID, may be deemed the best, where the same NID may be achieved by adjusting the identification threshold of each estimator. It should be emphasized that such a comparison is meaningful only for MM estimators using the same (total) model set because the relation of the model set to the mode space is not accounted for. For example, it is almost always the case that a 2-model MM estimator has better CID, IID, and NID than a 100-model MM estimator for the same problem because these percentages do not consider how fine the mode space is quantized by the model set.

All the criteria and measures defined in Secs. 9.2 and 9.3, including their extensions and variants for problems in a particular application area, have been adopted in the evaluation of variable-structure MM algorithms in [22, 21, 32] for a comparison with the fixed-structure algorithms.

9.4 For Hybrid State Estimation

A model set A is said to be better than set B for a given mode if

$$\|\xi - \hat{\xi}_A\|^2 < \|\xi - \hat{\xi}_B\|^2 \tag{30}$$

or (not equivalently)

$$\|\hat{\xi}_S - \hat{\xi}_A\|^2 < \|\hat{\xi}_S - \hat{\xi}_B\|^2 \tag{31}$$

where $\xi = (x, s)$ is the hybrid state of the hybrid system; $\hat{\xi}_S$, $\hat{\xi}_A$, and $\hat{\xi}_B$ are the (optimal) estimators based on the optimal set $S = \mathbf{S}$, set A, and set B, respectively. This definition appears to be good theoretically. It has, however, a major difficulty: It is not easy to define the measure $\|\cdot\|^2$ reasonably for the hybrid state in many situations.

It is also possible to define a Bayesian risk (cost) function and then the model (or model set) based on which the optimal estimator minimizes the function may be deemed the optimal.

9.5 Probabilistic Measures for Model-Set Comparison and Choice

A definition based on a pure probabilistic metric is the following: Set M_1 is **more probable** than set M_2 if

$$P\{m \in M_1 | m \in M\} > P\{m \in M_2 | m \in M\}$$

where M_1 and M_2 are subsets of M. It, however, also suffers from the second drawback mentioned in Sec. 9.2. Most probability-based definitions have a common shortcoming: The probability depends significantly on the event $\{m \in M\}$, namely, the choice (assumption) of the total *model* set M if $M \neq \mathbf{S}$. One may consider replacing M with the mode space $\mathbf{S}$. However, it may be difficult or infeasible to calculate the associated probability $P\{s \in M_2 | s \in \mathbf{S}\}$ for most practical problems, while the probabilities $P\{m \in M_i | m \in M\}$ are available from an MM algorithm using model set M.

It has been shown [18] that the optimal model set M for the MM approach is $M = \mathbf{S}$ and the deterioration worsens as M and $\mathbf{S}$ become more mismatched. Given the same degree of mismatch, however, the case of missing models is (somewhat) worse than the case of extra models. With these effects, given a collection of not necessarily disjoint model sets $M_1, \ldots, M_N$, the *best model set* may be defined as the one with the smallest number of models among the model sets with the largest probability of including the true mode s; that is, the best of $M_1, \ldots, M_N$ is the set M_j with the cardinality $|M_j| = \min_{q \in Q} |M_q|$, where

$$\begin{aligned} P\{s \in M_q | s \in \mathbf{S}\} &= \max_{i \in \{1,\ldots,N\}} P\{s \in M_i | s \in \mathbf{S}\}, \\ \forall q \in Q &\subset \{1, \ldots, N\} \end{aligned}$$

9.6 Information Measures

We now present several information measures for modes, models, and model sets. They are probably most thorough, fundamental, and general of all measures proposed in this paper.

Kullback-Leibler information number between mode and model. Compared with measures proposed above, a more thorough but more abstract measure of the closeness (or discrepancy) between a model and a mode is based on some information metrics. The Kullback-Leibler information number, also known as **relative entropy**, is probably the most suitable, which has been applied to such closely related areas as system identification, model validation, and performance evaluation [6, 5, 31, 11]. It measures how close a model is to a mode in terms of information contained. We provide the following definitions.

Let $f_s(x)$ and $f_m(x)$ be the pdfs of a random mode s and random model m, respectively. Then the Kullback-Leibler information number is defined by

$$\begin{aligned} I(f_s; f_m) &= E[\log f_s(x) - \log f_m(x)] \\ &= E[\log[f_s(x)/f_m(x)]] \\ &= \int \log\left(\frac{f_s(x)}{f_m(x)}\right) f_s(x) dx \end{aligned}$$

Given a value of a pair of mode s and model m, the corresponding Kullback-Leibler information number can be defined by

$$\begin{aligned} I[f_s(z); f_m(z)] &= E[\log[f_s(z)/f_m(z)]] \\ &= \int \log\left(\frac{f_s(z)}{f_m(z)}\right) f_s(z) dz \end{aligned}$$

through the likelihood functions of the given, deterministic mode s and model m, that is, $f_s(z)$ and $f_m(z)$, respectively. More generally, the **Kullback-Leibler information number** between a random mode s and a random model m in terms of data z can be defined by

$$\begin{aligned} I(f_{z,s}; f_{z,m}) &= E[\log[f_{z,s}(z,x)/f_{z,m}(z,x)]] \\ &= \int \log\left(\frac{f_{z,s}(z,x)}{f_{z,m}(z,x)}\right) f_{z,s}(z,x) dz dx \end{aligned}$$

where $f_{z,s}(z,x)$ is the joint pdf of the data and mode s, and $f_{z,m}(z,x)$ is the joint pdf of the data z and model m.

While these Kullback-Leibler information numbers are positive definite, none of them are true metrics (distance functions) as defined by positive definiteness, symmetry, and triangle inequality.

With these definitions, we say mode m_* is the **best model** in the set M if

$$I(s; m_*) = \min_{m \in M} I(s; m)$$

where $I(s;m) = I(f_{z,s}; f_{z,m})$, $I[f_s(z); f_m(z)]$, or $I(f_s; f_m)$, defined above, depending on the case.

For discrete s and m, the above definitions are valid after replacing the pdfs with the corresponding probability mass functions.

Information number between models. Similarly, we define (Jeffreys) **information number** between two models m_i and m_j to measure how close they are in terms of information contained as

$$I(m_i, m_j) = I(m_i; m_j) + I(m_j; m_i)$$

where $I(m_i; m_j) = I(s; m_j)|_{s=m_i}$ and $I(m_j; m_i) = I(s; m_i)|_{s=m_j}$. This information number is positive definite and symmetric, but still not a true metric.

Information distances between models (model sets) relative to mode distribution. Similarly, we can define **information distances** for two models (or model sets) to measure how close they are, given mode distribution.

Specifically, we define the information distance between two *deterministic* models m_i and m_j given mode distribution as

$$\begin{aligned} d(m_i, m_j) &= \left| E\left[\log\left(\frac{f_{m_i}(z)}{f_{m_j}(z)}\right)\right]\right| \\ &= \left|\int \log\left(\frac{f_{m_i}(z)}{f_{m_j}(z)}\right) f(z,s)dzds\right| \end{aligned}$$

where $f_{m_i}(z)$ is the likelihood function of model m_i and the expectation is over both observation z and mode s. It is the expected value (over all s) of their information distance for a given value of s, given by

$$d_s(m_i, m_j) = \left|\int \log\left(\frac{f_{m_i}(z)}{f_{m_j}(z)}\right) f(z|s)dz\right|$$

We define the information distance between two *random* models m_i and m_j (or essentially equivalently, between two model sets M_i and M_j, see Sec. 3) as

$$\begin{aligned} d(M_i, M_j) &= D(m_i, m_j) \\ &= \left| E\left[\log\left(\frac{f_{z,m_i}(z,x)}{f_{z,m_j}(z,x)}\right)\right]\right| \\ &= \left|\int \log\left(\frac{f_{z,m_i}(z,x)}{f_{z,m_j}(z,x)}\right) f_{z,s}(z,x)dzdx\right| \end{aligned}$$

For models as discrete random variables, replace the density parts with the corresponding probability mass functions.

It can be easily shown that the information distances so defined between two models (model sets) are indeed distance metrics.

Mutual information between mode and model. Closely related with relative entropy is mutual information. The mutual information between a random mode and a random model is the relative entropy between their joint distribution and their distribution product:

$$\begin{aligned} I(s:m) &= I(f_{s,m}; f_s f_m) \\ &= E\left[\log\left(\frac{f(s,m)}{f(s)f(m)}\right)\right] \\ &= \int \log\left(\frac{f(s,m)}{f(s)f(m)}\right) f(s,m)dsdm \end{aligned}$$

It measures the dependence between s and m; in other words, it quantifies information contained in m for predicting s (or the other way round). Similarly, we define the mutual information between two random models through their probability mass functions $p(m_i, m_j), p(m_i), p(m_j)$ as

$$\begin{aligned} I(m_i, m_j) &= E\left[\log\left(\frac{p(m_i, m_j)}{p(m_i)p(m_j)}\right)\right] \\ &= \sum \log\left(\frac{p(m_i, m_j)}{p(m_i)p(m_j)}\right) p(m_i, m_j) \end{aligned}$$

It measures the dependence between m_i and m_j.

10 Model-Set Choice by Hypothesis Testing

10.1 Formulation of Model-Set Choice as Hypothesis Testing Problems

Model-set choice particularly suitable for model-set *adaptation* has been studied extensively in a general setting based on hypothesis testing in [15] and thus will not be repeated here.

In this section, we consider only offline model-set choice problems that are general in nature. A major difference between the (online) model-set adaptation and the (offline) model-set design is that a special model set (i.e., the current set) is present in the former, but usually not in the latter.

The following general and representative problems are considered:

Problem 1: "Which model is the best in a given set?" or "Select the best model in the set $M = \{m_1, m_2, \ldots, m_n\}$."

Problem 2: "Which of the two model sets M_1 and M_2 is better?" or "Choose one of the two model sets M_1 and M_2."

Problem 3: "Which of the model sets $M_1, \ldots, M_N$ is the best?" or "Choose one of the model sets $M_1, \ldots, M_N$."

Problem 4: "Is any of the model sets $M_1, \ldots, M_N$ better than the set M?" or "Determine if any of the model sets $M_1, \ldots, M_N$ is better than a given set M."

Clearly, Problems 1 and 2 are special cases of Problem 3.

Solutions of these problems presented below are adapted from those presented in [15]. Denoting by s the true mode, Problem 1 can clearly be formulated as the following hypothesis testing problem

$$H_1 : s = m_1 \text{ vs. } H_2 : s = m_2 \ \cdots \ \text{vs.} \quad H_n : s = m_n$$

This is in general a multihypothesis testing problem. For $n = 2$, it reduces to a binary hypothesis problem, which can be solved by, e.g., the sequential probability ratio test (SPRT), the (non-sequential) Neyman-Pearson test, or a Bayes test. For multihypothesis problems, however, there is in general no optimal Neyman-Pearson or SPRT-type test without additional constraints. Several tests proposed in [15] can be applied to solve this problem, of which the sequential ranking test appears most attractive.

Problem 2 can be formulated as the following hypothesis testing problem

$$H_1 : s \in M_1 \quad \text{vs.} \quad H_2 : s \in M_2$$

Both hypotheses are in general composite, for which there is in general no optimal test without additional constraints. What is even worse is that the model sets are quite often non-disjoint. What if the true mode lies in the common part of the two model sets? Such problems are almost never dealt with in statistics. Nevertheless, the sequential model-set likelihood (or probability) ratio test proposed in [15]

can be modified to solve this problem fairly satisfactorily, which is optimal in some meaningful sense.

Problem 3 can be formulated as the following hypothesis testing problem

$$H_1 : s \in M_1 \text{ vs. } H_2 : s \in M_2 \ \cdots \ \text{vs. } H_N : s \in M_N$$

It is a generalization of both Problems 1 and 2. It differs from Problem 1 in that each hypothesis is in general composite rather than simple. Naturally, all difficulties associated with either multiple, composite, or non-disjoint hypothesis testing problems are present here. Nevertheless, several tests proposed in [15] can be modified to solve this problem, of which the most attractive is the sequential ranking test.

Problem 4 can be formulated as the following hypothesis testing problem

$$H : s \in M \text{ vs. } H_1 : s \in M_1 \ \cdots \ \text{vs. } H_N : s \in M_N$$

It differs from Problem 3 in that model set M is special. It also differs from model-set adaptation problems in the implication of an indecision as well as each hypothesis. A modified version of the multiple model-set SPRT proposed in [15] is particularly suitable for this problem.

10.2 Model-Set Probability and Likelihood

Since hypothesis testing is driven by data, offline mode-set choice by hypothesis testing also requires the use of data. This can be done using past data or via simulation. Let z_k be the observation at time k and $z^k = \langle z_\kappa \rangle_{\kappa \le k}$ be the observation sequence through time k. Let $\tilde{z}_k = z_k - \hat{z}_{k|k-1}$ be the observation residual at time k; that is, the part of z_k that is unpredictable from the past. It is available from an MM estimator. Let $\tilde{z}^k = \langle \tilde{z}_\kappa \rangle_{\kappa \le k}$ be the sequence of the observation residuals through time k.

Since the task is to decide on a model set, the probabilities and likelihoods of a model set are naturally of major interest.

The *marginal* likelihood of a model-set M_j at time k is the sum of the predicted probabilities $P\{m = m_i | m \in M_j, z^{k-1}\}$ times the marginal likelihoods $p[\tilde{z}_k | m = m_i, z^{k-1}]$, both for all the *models* in M_j:

$$\begin{aligned} & L_k^{M_j} \triangleq [\tilde{z}_k | m \in M_j, z^{k-1}] \\ = & \sum_{m_i \in M_j} p[\tilde{z}_k | m = m_i, z^{k-1}] P\{m = m_i | m \in M_j, z^{k-1}\} \end{aligned}$$

The *joint* likelihood of the model-set M_j is defined as $L_{M_j}^k = p[\tilde{z}^k | m \in M_j]$. Note that a subscript k and a superscript k are used for quantities at k and through k, respectively. The *joint* likelihood ratio $\Lambda^k = L_{M_1}^k / L_{M_2}^k$ of *model-set* M_1 to M_2 can often[4] be (approximately) computed as the product of model-set marginal likelihood ratios:

$$\Lambda^k = \prod_{k_0 \le \kappa \le k} \frac{L_\kappa^{M_1}}{L_\kappa^{M_2}} \tag{32}$$

which simplifies computation greatly, where k_0 is the test starting time. This is important: The observation sequence itself is usually not independent; however, the sequence of marginal likelihood ratios is usually approximately independent since the observation residual sequence is at most weakly correlated and the likelihood ratio is approximately Gaussian distributed under some mild conditions [26].

The (posterior) probability that the true mode is in a model-set M_j at time k is defined as

$$\begin{aligned} \mu_k^{M_j} &= P\{m \in M_j | m \in \mathbf{M}_k, z^k\} \\ &= \sum_{m_i \in M_j} P\{m = m_i | m \in \mathbf{M}_k, z^k\} \end{aligned} \tag{33}$$

which is the sum of the probabilities of all models in M_j, where $\mathbf{M}_k$ is the total model-set in effect at time k, which includes M_j as a subset and is problem dependent. The model probability $P\{m = m_i | m \in \mathbf{M}_k, z^k\}$ for each model m_i is available from an MM estimator.

10.3 Solutions of Hypothesis Testing Problems for Model-Set Choice

There are two types of hypothesis test: sequential and non-sequential. Sequential tests are more attractive for model-set choice than nonsequential tests primarily for the following reasons: For any given decision error rates, sequential tests make decisions only if the evidence is sufficient; they are more efficient — need a smaller sample size, which does not need to be determined in advance; and (approximate) thresholds of sequential tests are easy to determine.

Our proposed hypothesis tests for model-set choice are based on the following tests, developed in [15] for model-set adaptation (where $M_1 \subset M$, $M_2 \subset M$) and modified here for offline model-set choice (where there is no M).

Theorem 10.1 (MS-SLRT): Mode-set sequential likelihood ratio test. For Problem 2 with the following specifications for the *expected* (weighted sum of) decision error probabilities

$$\begin{aligned} & P\{H_2 | m \in M_1\} \\ = & \sum_{m_i \in M_1} P\{H_2 | m = m_i\} P\{m = m_i | m \in M_1\} \le \alpha \end{aligned} \tag{34}$$

$$\begin{aligned} & P\{H_1 | m \in M_2\} \\ = & \sum_{m_i \in M_2} P\{H_1 | m = m_i\} P\{m = m_i | m \in M_2\} \le \beta \end{aligned} \tag{35}$$

the following SPRT-based test is uniformly most efficient:

[4] For example, this is exact if $\log(\Lambda^k)$ sequence is a random walk.

- Choose M_1 if $\Lambda^k \geq B$
- Choose M_2 if $\Lambda^k \leq A$
- Use $M_1 \cup M_2$ and continue to test with more observations if $A < \Lambda^k < B$

where Λ^k is the joint likelihood ratio, and A and B are two constants to be determined.

Proof. It follows similarly as that of Theorem 1 in [15].

Remarks.

- Since $p[\tilde{z}^k | m \in M_l]$ and $p[\tilde{z}_\kappa | m \in M_l, z^{\kappa-1}]$ are nothing but the joint and marginal likelihoods of the model set M_l, the test is actually the SPRT using the *model-set* joint likelihood ratio. $P\{m = m_i | m \in M_l, z^{\kappa-1}\}$ is the predicted probability of mode m_i in model set M_l. The model probabilities and likelihoods are given by an MM estimator.
- In practice, the following constants A and B can be used
$$A = \frac{\beta}{1-\alpha}, \qquad B = \frac{1-\beta}{\alpha} \tag{36}$$
which are exact if L^k can only be either in (A, B) or on the boundary A or B (i.e., there is no overshoot — excess over the boundaries) and are accurate if the amount of overshoot is small. In the case there is overshoot, the test based on the above values of A and B is (slightly) more conservative than the optimal one: It is (slightly) less efficient than the optimal one but the error probability specifications are still satisfied.
- The composite hypothesis testing problem is effectively converted into a simple one with a clear physical justification [i.e., the error probabilities (34)–(35)] by the technique of the weight function $P\{m = m_i | m \in M_l\}$.
- SPRT is usually very efficient in the case where the type I error probability α is not very small even if the truth is in between the two hypotheses.
- A regularity condition for the optimality of this test is that the sequence of model-set log-likelihood ratios $\log(\Lambda^k)$ is a random walk.

Theorem 10.2 (MS-SPRT): Mode-set sequential probability ratio test. With the following specifications for the expected decision error probabilities

$$\sum_{m_i \in M_1} P\{H_2 | m = m_i\} P\{m = m_i\} \quad \leq \quad \alpha' \tag{37}$$

$$\sum_{m_i \in M_2} P\{H_1 | m = m_i\} P\{m = m_i\} \quad \leq \quad \beta' \tag{38}$$

the above model-set sequential likelihood ratio test (Theorem 1) is uniformly most efficient for Problem 2 if the joint likelihood ratio Λ^k is replaced by the probability ratio $P^k = \frac{P\{m \in M_1 | z^k\}}{P\{m \in M_2 | z^k\}}$.

Proof. It follows similarly as that of Theorem 2 in [15].

Remarks.

- The test statistic in Theorem 10.1 is the *product* of ratios of model-set marginal *likelihoods* (for the current and past times), while in Theorem 10.2 it is the ratio of posterior mode-set probabilities *at the current time only*. This makes sense since a marginal likelihood does not carry historical information, while the posterior probability depends on past as well as current information.
- The weight functions $P\{m = m_i | m \in M_l\}$ used in Theorem 10.1 satisfy the normalization requirements (i.e., sum up to one), but those in Theorem 10.2 do not. (34)–(35) and (37)–(38) are different. The latter uses the same weight for the two error probabilities under the same true mode, which seems more reasonable than using different weights as in the former. This is achieved at the price of losing the normalization property. Their error probabilities are distinct and thus the corresponding optimal tests are different.
- The model-set probability ratio and likelihood ratio are related via the ratio of prior probabilities:
$$\begin{aligned} P^k &= \frac{P\{m \in M_1 | z^k\}}{P\{m \in M_2 | z^k\}} \\ &= \frac{p[\tilde{z}_k | m \in M_1, z^{k-1}]}{p[\tilde{z}_k | m \in M_2, z^{k-1}]} \frac{P\{m \in M_1 | z^{k-1}\}}{P\{m \in M_2 | z^{k-1}\}} \\ &= \Lambda^k \frac{P\{m \in M_1\}}{P\{m \in M_2\}} \end{aligned}$$
This relationship can be used to simplify the calculation of P^k from Λ^k.

In summary, the use of the weight functions $P\{m = m_i | m \in M_l\}$ or $P\{m = m_i\}$ converts the composite hypothesis testing problem into a simple one, and thus the model-set likelihood or probability acts exactly the same as the likelihood for a simple hypothesis testing problem. Probably more importantly, overlap between model sets M_1 and M_2 poses no problem after this conversion. This is the key to Theorems 10.1 and 10.2.

Problem 3 may be solved by the following **sequential ranking test**: At each time k, rank all N_k of the model sets $M_1, \ldots, M_N$ that have not yet been rejected as $M_{(1)}, \ldots, M_{(N_k)}$ such that their model-set probabilities $\mu_k^{M_{(i)}} = P\{m \in M_{(i)} | z^k, m \in \mathbf{M}_k\}$ are in a decreasing order:

$$\mu_k^{M_{(1)}} \geq \mu_k^{M_{(2)}} \geq \cdots \geq \mu_k^{M_{(N_k)}}$$

where $\mathbf{M}_k$ is the union of all models not yet rejected. Then,

- Accept $M_{(1)}$ if $\frac{\mu_k^{M_{(1)}}}{\mu_k^{M_{(2)}}} \geq B$
- Reject $M_{(j)}, \ldots, M_{(N_k)}$ if $\frac{\mu_k^{M_{(j)}}}{\mu_k^{M_{(1)}}} \leq A$

  Continue to test until one model set has been accepted or not rejected, and thus are chosen. Here A and B are design parameters, which control the error probabilities.

The model-set probabilities can be replaced by the model-set *joint* likelihoods in the above test.

Problem 4 may be solved by the following **multiple model-set sequential likelihood ratio test** (**MMS-SLRT**):

S1. Perform N one-sided MS-SLRTs simultaneously for N pairs of hypotheses ($H : m \in M$ vs. $H_1 : m \in M_1$), $\ldots$, ($H : m \in M$ vs. $H_N : m \in M_N$). These tests are *one-sided* in the sense that H is never rejected, which is implemented by using thresholds B_i and $A = -\infty$. This step ends when only one of the hypotheses $H_1, \ldots, H_N$ is not rejected yet:

  – Reject all M_i for which $\Lambda_i^k = L_M^k / L_{M_i}^k \geq B_i$;
  – Continue to test for the remaining pairs until only one of the hypotheses $H_1, \ldots, H_N$ is not rejected.

  Specifically, let K be the smallest sample size (time) by which some $(N-1)$ of the N alternative hypotheses are rejected by the one-sided MS-SLRTs

$$\begin{aligned} k_i &= \min\left\{k : \Lambda_i^k \geq B_i, i = 1, \ldots, N, i \neq j\right\} \\ K &= \min\left\{k : k \geq k_i, i = 1, \ldots, N, i \neq j\right\} \end{aligned}$$

  Then, the test accepts H_j if $\Lambda_j^K < B_j$, where $B_1, \ldots, B_N \geq 1$ are chosen such that the type I and type II error probabilities for all binary problems are α and β, respectively. If more than one hypothesis is rejected at last simultaneously, the one with the largest model-set likelihood is accepted.

S2. Perform a (two-sided) MS-SLRT to test $H : m \in M$ vs. $H_j : m \in M_j$, where H_j is the winning hypothesis in Step 1.

Remarks.

- The union of all remaining model sets is used until the final decision in Step 2 is made.
- If MS-SLRTs with B_i of (36) are used for all binary tests in Step 1, then the thresholds are all equal: $B_j = B_i = \frac{\alpha}{1-\beta}$. Alternatively, the threshold suggested in [23] may be used.
- Since the expected sample size is asymptotically minimized in all the one-sided MS-SPRTs of Step 1 [23] and the MS-SPRT of Step 2 when $H : m \in M$ is true, it seems reasonable to expect that the above MMS-SLRT has a minimum expected sample size asymptotically.
- The model-set joint likelihood ratio may be replaced by the model-set probability ratio $P_j^k = \frac{P\{m \in M \,|\, z^k\}}{P\{m \in M_j \,|\, z^k\}}$, leading to multiple model-set sequential probability ratio test (MMS-SPRT).

Another solution of Problem 4 is the following **multiple-level test** (**MLT**):

S1 Test all pairs of hypotheses *separately*:

$$H : m \in M \quad \text{vs.} \quad H_i : m \in M_i$$

using, e.g., MS-SLRT or MS-SPRT. Complete all these tests. Let $\mathcal{H}$ be the set of accepted hypotheses from all these tests. Delete H from $\mathcal{H}$ unless $\mathcal{H}$ contains no other hypothesis or H is deemed much more important than the other model sets.

S2 Let H be the best hypothesis in $\mathcal{H}$[5]. Go to Step 1 to test H against all the other hypotheses in $\mathcal{H}$ pairwise.

Repeat this process until only one hypothesis remains and the corresponding model set is then chosen. Before the final decision is made, the union of all remaining model sets is used.

Remarks. (a) The error probability pair α_i and β_i used in the binary tests of Step 1 is better chosen in such a way that these tests have about the same expected sample size since the sample size of Step 1 is equal to the largest sample size of the component problems. (b) Clearly, Step 1 here is more efficient than the corresponding MMS-SLRT or MMS-SPRT given the same error probabilities. The weakness of this test is the possibility of multiple iterations, which depends on how good the set M is. In model-set choice, there is no need to consider a mode jump. If the model sets do not have large overlap, a good model set M may be inferior to at most a few other model sets. Thus, statistically speaking, quite often only one, occasionally two, and unlikely more iterations are needed in the MLT.

We emphasize that all these tests are computationally extremely efficient and easy to implement — little extra computation is required other than what is already available in an MM estimator.

[5]The best hypothesis is one that is not rejected in any other binary tests, and the first hypothesis accepted (if the same error probabilities are used for all binary tests) or the one accepted with the smallest error probability (if the same expected sample size is used). Model-set likelihoods or probabilities are used to break ties if any.

As a degenerated case, it is clear that the above tests using model sets are applicable to Problem 1 for testing among models in a single set by simply replacing the model-set likelihood and probability with model likelihood and probability.

Examples of the sequential tests presented here can be found in [15, 14].

11 Examples for Model-Set Design

11.1 The Design Problem

Most examples presented in this paper deal with the following simple system with position-only measurements

$$x_{k+1} = Fx_k + w_k, \;\; z_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x_k + v_k \quad (39)$$

where $x = [x_1, \dot{x}_1, x_2, \dot{x}_2]'$, and process and measurement noises have constant means $\bar{w} = 0$, $\bar{v} = 0$ and covariances $Q = 0.1^2 I$, $R = rI$, respectively, with $r = 100^2$ unless stated otherwise. Assume that the linear-Gaussian assumption of the Kalman filter is valid. Two generic types of model are considered: nearly constant-velocity (CV) model and coordinated-turn (CT) (with a known turn rate) model, given by [16]

$$F_{\rm CV} = \text{diag}[F_2, F_2], \quad F_2 = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \quad (40)$$

$$F_{\rm CT}(\omega) = \begin{bmatrix} 1 & \frac{\sin\omega T}{\omega} & 0 & -\frac{1-\cos\omega T}{\omega} \\ 0 & \cos\omega T & 0 & -\sin\omega T \\ 0 & \frac{1-\cos\omega T}{\omega} & 1 & \frac{\sin\omega T}{\omega} \\ 0 & \sin\omega T & 0 & \cos\omega T \end{bmatrix} \quad (41)$$

where $T = 5$s is sampling period. Denote by CT(3°/s) a CT model with turn rate $\omega = 3^\circ$/s. Note that CV = CT(0°/s).

All numerical examples presented in this paper are based on the use of the above CT and CV models in one or more sets. These models come primarily from the consideration in maneuvering target tracking for Air Traffic Control (ATC) surveillance [4, 3].

Consider a problem of surveillance for an ATC system. Suppose we decide to use an MM algorithm with a set M of three CT models: $m_1 = \omega_1$, $m_2 = \omega_2$, $m_3 = \omega_3$. Clearly, to design this set M optimally in some sense, a probability density function (pdf) $f(s)$ of the true turn rate s is needed. We propose the following Gaussian-mixture model:

$$f(s) = c_0 N_0(s) + cN_1(s) + cN_{-1}(s), \; c_0 + 2c = 1 \quad (42)$$

where

$$N_0(s) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{s^2}{2\sigma_0^2}}, \;\; N_{\pm 1}(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s\pm\omega_s)^2}{2\sigma^2}}$$

and ω_s is a known standard turn rate. This model can be well justified if most flights will either approach/departure the runway directly without a turn, or with a turn of an extremely slow rate or a rate close to the standard one ($\pm\omega_s$). Note that $f(s)$ may or may not have three peaks, depending on the parameters ω_s, c_0, c, σ_0, and σ.

Consider a specific example with

$$c_0 = c = 1/3, \quad \sigma_0 = \sigma = 1^\circ/\text{s}, \quad \omega_s = 3^\circ/\text{s} \quad (43)$$

Its pdf is plotted in Fig. 11.

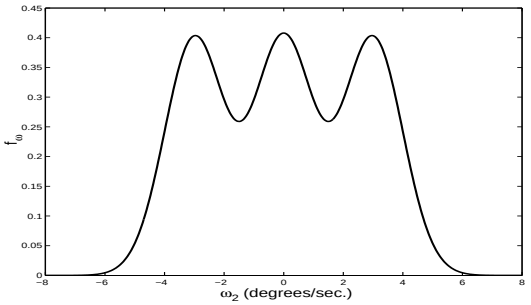


Fig. 11: The probability density function of the true turn rate ω.

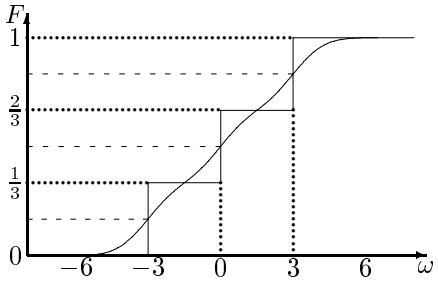


Fig. 12: Model-set design via cdf of turn rate ω.

11.2 Design by Minimizing Distribution Mismatch

Clearly, the distribution-based approach of Part I can be applied to solve this problem for the specific example with (43). Assume that we want to use only three models. First, plot the cdf $F_s(x)$ of the true mode (its pdf is plotted in Fig. 11) as in Fig. 12. Then, since $F_s(-3^\circ/\text{s}) \approx 1/6$, $F_s(0) = 3/6$, and $F_s(3^\circ/\text{s}) \approx 5/6$, the simple distribution-based design yields the model set $M = \{\omega_1, \omega_2, \omega_3\} = \{0, \pm 3^\circ/\text{s}\}$, with the corresponding initial model probabilities $\{1/3, 1/3, 1/3\}$. Note that it turns out that we have a model at each peak of the pdf $f_s(x)$. This would not be the case if the three peaks are closer or σ is larger. Clearly, this approach is more beneficial when we want to have more models.

11.3 Design by Minimizing Modal Distance

We now demonstrate how to design M by minimizing an optimality criterion for mode estimation. Consider minimizing the average modal distance squared (24), dropping $s \in \mathbf{S}$ for simplicity,

$$\begin{aligned} & \|s-m\|^2_{m\in M} = E[(s-\omega)^2 | m \in M] \\ = \;& E[E[(s-\omega)^2 | s, m \in M] | m \in M] \\ = \;& \int_{-\infty}^{\infty} \sum_i (s-\omega_i)^2 P\{m=\omega_i | m \in M, s\} f(s) ds \end{aligned}$$

In fact, it turns out that for this example mode estimation error squared $\|s-\hat{s}\|^2$ is equal to average modal distance squared.

By symmetry of $f(s)$, $\omega_1 = 0$, $\omega_2 = \omega$, $\omega_3 = -\omega$ and we need only to determine the optimal ω. Thus, we need only to solve the following optimization problem

$$\begin{aligned} \min_\omega J \;=\; & \int [s^2 P\{m=0|s\} + (s-\omega)^2 P\{m=\omega|s\} \\ & + (s+\omega)^2 P\{m=-\omega|s\}] f(s) ds \end{aligned}$$

We emphasize that conditional model probabilities $w_i(s) = P\{m=\omega_i | m \in M, s\}$ given true mode in the above is not to be confused with the unconditional model probabilities $P\{m=\omega_i | m \in M\}$. For example, compare (44) with (46). The significance of this probability and how to determine it were discussed in great detail in Sec. 8. As explained there, $w_i(s)$ can be interpreted as a window function. Thus, the problem becomes

$$\min_\omega J = \int [s^2 w_1(s) + (s-\omega)^2 w_2(s) + (s+\omega)^2 w_3(s)] f(s) ds$$

For simplicity, however, we use the rectangular window (see Sec. 8.4), which leads to the following pmf:

$$\begin{aligned} w_i(s) \;&=\; P\{m=\omega_i|s\} \\ &=\; \begin{cases} 1[s;(-\omega/2,\omega/2)], & i=1 \\ 1[s;(\omega/2,\infty)], & i=2 \\ 1[s;(-\infty,-\omega/2)], & i=3 \end{cases} \end{aligned} \tag{44}$$

(44) indicates that models $m_1 = 0$, $m_2 = \omega$, or $m_3 = -\omega$ is in effect if and only if the true turn rate falls inside the intervals $(-\omega/2, \omega/2)$, $(\omega/2, \infty)$, or $(-\infty, -\omega/2)$, respectively. In other words, these intervals are the effective coverage of the three models. This use of the rectangular window is well justified by Theorem 6.1 of Part I since it satisfies the nearest-neighbor condition (Condition B).

With (44), the cost function is

$$\begin{aligned} & J = \int \sum_{i=1}^{3} (s-\omega_i)^2 P\{m=\omega_i | m \in M, s\} f(s) ds \\ = \;& \int_{-\omega/2}^{\omega/2} s^2 f(s) ds + \int_{\omega/2}^{\infty} (s-\omega)^2 f(s) ds \\ & + \int_{-\infty}^{-\omega/2} (s+\omega)^2 f(s) ds \\ = \;& \int_{-\infty}^{\infty} s^2 f(s) ds + 2\omega^2 \int_{\omega/2}^{\infty} f(s) ds - 4\omega \int_{\omega/2}^{\infty} s f(s) ds \\ = \;& c_0 \sigma_0^2 + 2c(\sigma^2 + \omega_s^2) + 2\omega^2 P_2 - 4\omega J_4 \end{aligned} \tag{45}$$

where $P_2 = \int_{\omega/2}^{\infty} f(s) ds$, given by

$$\begin{aligned} & P_i = P\{m=\omega_i | m \in M\} \\ = \;& \begin{cases} 2\int_0^{\omega/2} f(s) ds, & i=1 \\ \int_{\omega/2}^{\infty} f(s) ds, & i=2,3 \end{cases} \\ = \;& \begin{cases} c_0 \text{ erf}\left(\frac{\omega/2}{\sqrt{2}\sigma_0}\right) + c \text{ erf}\left(\frac{\omega/2-3}{\sqrt{2}\sigma}\right) \\ \quad + c \text{ erf}\left(\frac{\omega/2+3}{\sqrt{2}\sigma}\right), & i=1 \\ \frac{1}{2}(1-P_1^{(2)}), & i=2,3 \end{cases} \end{aligned} \tag{46}$$

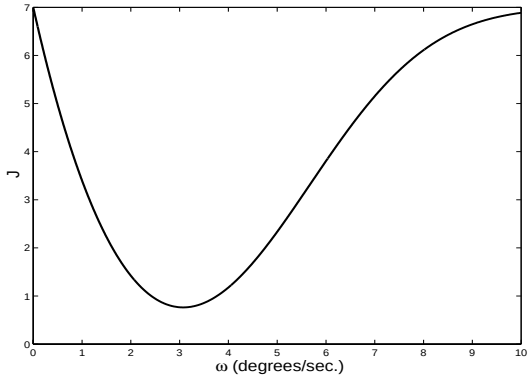
and J_4 can be obtained by direct integration and using the error function $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$ as

$$\begin{aligned} & J_4 = \int_{\omega/2}^{\infty} s f(s) ds \\ = \;& \frac{c_0 \sigma_0}{\sqrt{2\pi}} e^{-\frac{\omega^2}{8\sigma_0^2}} + \frac{c\sigma}{\sqrt{2\pi}} \left[e^{-\frac{(\omega/2+\omega_s)^2}{2\sigma^2}} + e^{-\frac{(\omega/2-\omega_s)^2}{2\sigma^2}} \right] \\ & + \frac{c\omega_s}{2} \left[\text{erf}\left(\frac{\omega/2+\omega_s}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{\omega/2-\omega_s}{\sqrt{2}\sigma}\right) \right] \end{aligned}$$

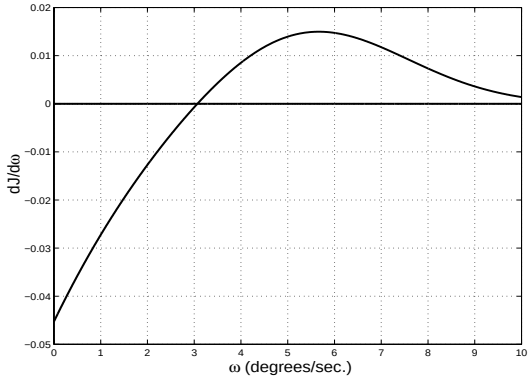
For the specific example with (43), the above optimization problem can be solved by plotting the above integrals J vs. ω (Fig. 13) and identifying the minimum within the interval of interest, which is around 3.05. Fig. 13(b) plots the derivative $dJ/d\omega$ obtained from Mathematica, which verifies the minimum. Therefore, the corresponding best model set is approximately $M = \{0, \pm 3^\circ/\text{s}\}$ and the (initial) model probabilities are $P_1 = 1/3, P_3 = P_2 = 1/3$

A design based on modal distance is particularly suitable for such applications as fault detection and isolation where the primary objective of hybrid estimation is mode estimation.

Note that the above procedures are still applicable even if a more sophisticated probabilistic model than (42) is used. However, if more than one parameter is to be determined, then the resulting optimization problem is in general multi-dimensional.



(a) Modal distance squared



(b) Derivative of modal distance squared

Fig. 13: The average modal distance squared (and its derivative) versus the turn rate ω.

11.4 Effectiveness of Designs

Monte-Carlo simulations were conducted to verify the above model-set designs. In the simulation, 500 true modes were generated randomly with the distribution given by (42). They were unknown to the estimators and not allowed to jump. The initial state of the system was $x_0 = [1000, 100, 200, 120]'$ and for simplicity each estimator used x_0 as the initial state estimate. For each true mode generated, 100 samples of base state trajectories and the corresponding measurement sequences were generated. These measurements were used by an autonomous MM (AMM, also known as static MM) estimator based on a model set — assuming that the true mode belongs to the model set with the corresponding pmf — to estimate the base state and the true mode. Three AMM estimators $\hat{x}_M$, $\hat{x}_{M_1}$, and $\hat{x}_{M_2}$ were obtained, over the time steps $k = 1, 2, \ldots, 10$, based on the model sets $M = \{0, \pm 3^\circ/\text{s}\}$, $M_1 = \{0, \pm 2^\circ/\text{s}\}$, and $M_2 = \{0, \pm 4^\circ/\text{s}\}$, with the initial model probabilities $\{1/3, 1/3, 1/3\}$, $\{0.3786, 0.2427, 0.3786\}$, and $\{0.288, 0.424, 0.288\}$, respectively, calculated by (46).

Fig. 14 shows the RMS errors, average modal distances, and mode estimation errors, as defined by (25) and (28), with $L = 500$ and $N = 100$. These results support the above designs.

11.5 Design by Hypothesis Testing Given Scenarios

The hypothesis-testing based approach to model-set choice of Sec. 10 can also be used for model-set design. We now give one such example.

An important question for model-set design is the following: Given a number of interested (or representative) scenarios, how to design a model set with fewer models than the number of scenarios?

Suppose for simplicity that the scenarios of interest are: true turn rates are 0°/s, 1°/s, 2°/s, 3°/s and 4°/s, respectively, and the model set to be designed is $\{0, \pm\omega\}$. In other words, the task is to determine ω such that the model set can cover the five possible true turn rates effectively. Varying ω, Fig. 15 shows (over 100 Monte-Carlo runs) the percent of correct decision of model selection for the five true turn rates of interest using MMS-SPRT with CV model as the special model. It seems reasonable from Fig. 15 to choose $\omega = 3^\circ$/s so that the correct decision is still above 80% even in the worst case where the true turn rate is 1°/s or 2°/s.

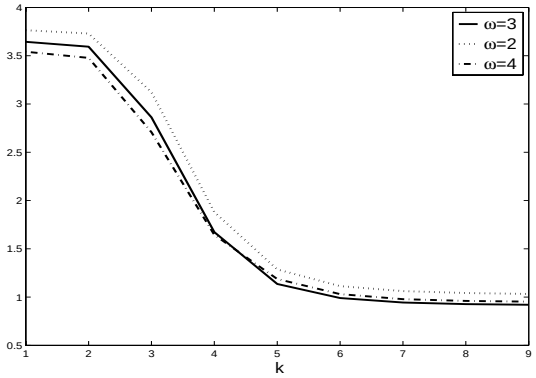
12 Examples of Model Efficacy

For the above ATC example, plots of the efficacies $w_i^P(s)$ and $w_i^T(s)$ of each of the three models in $M = \{0, \pm 3^\circ/\text{s}\}$ are given in Fig. 16 and Fig. 17, generated by (16) and (15), respectively. The variance of the measurement noise used is $r = 10^2$ and $\alpha = \beta = 0.1$ were used in the test to obtain $w_i^T(s)$. Two separate SPRTs were used to test two pairs of hypotheses ($H_1 : \omega = 0$ vs. $H_2 : \omega = 3^\circ/\text{s}$) and ($H_1 : \omega = 0$ vs. $H_2 : \omega = -3^\circ/\text{s}$). The final winner is clear from the winners of these two pairs.

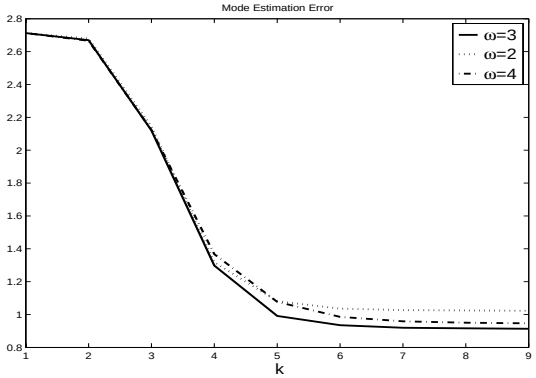
A comparison of Fig. 16 and Fig. 10, where the initial model probabilities are $[P_1, P_2, P_3] = [1/3, 1/3, 1/3]$ and $[0.7, 0.15, 0.15]$, respectively, verifies that probability-based model efficacy is insensitive to the initial model probabilities used. Also, it is clear that the probability-based and test-based model efficacies are close.

We now demonstrate another way of using hypothesis tests to determine model efficacy.

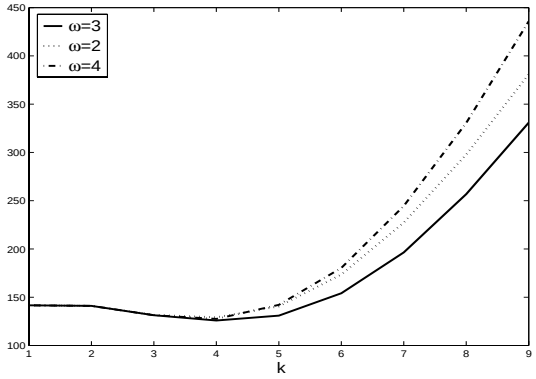
Consider model set $M = \{0, \pm 3^\circ/\text{s}\}$. Varying the turn rate of the true mode from 0 to 5°/s, the application of the MMS-SPRT of Sec. 10 (with CV model chosen as the special model) to the degenerated case of models (rather than model sets) yielded the results (over 100 Monte-Carlo runs) shown in Fig. 18 with 0% of no decision. All plots are averaged over the thresholds corresponding to the type I and type II error probabilities $\alpha = \beta \in (0.01, 0.1)$. Note that when the true turn rate is smaller than 1.5°/s or so, the correct decision is to choose the CV model, while it



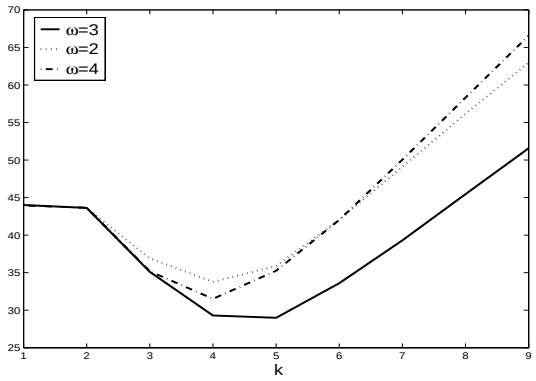
(a) Modal distances



(b) Mode estimation errors



(c) RMS position errors



(d) RMS velocity errors

Fig. 14: RMS errors, modal distances, and mode estimation errors of four MM estimators.

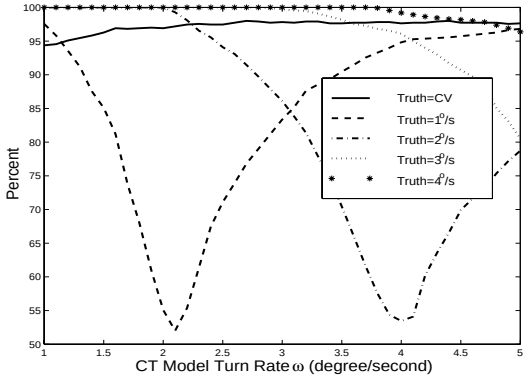


Fig. 15: Percent of correct model selection.

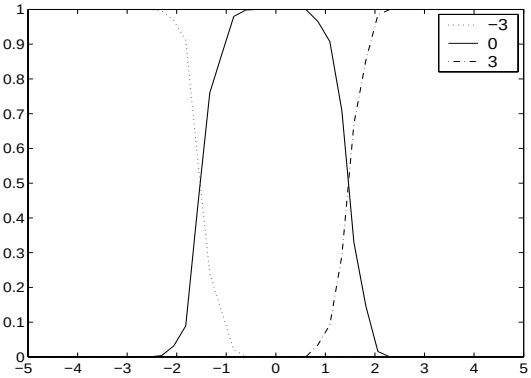


Fig. 16: Probability-based model efficacies $w_i^P(s)$ for set $M = \{0, \pm 3^\circ/\text{s}\}$.

is a correct decision to choose the CT(3°/s) model when the true turn rate is above 1.5°/s or so. It can be reasonably concluded from Fig. 18 that the effective coverage regions of the three models in the set are approximately $(-1^\circ/s, 1^\circ/\text{s})$, $(2^\circ/\text{s}, 5^\circ/\text{s})$, $(-5^\circ/\text{s}, -2^\circ/\text{s})$, respectively. If the true turn rate (say, 1.7°/s) is not in any of the three intervals, it is covered by the two neighboring models jointly, which can be verified by a test between the two model sets $M_1 = \{0, 3^\circ/\text{s}\}$ and $M_2 = \{-3^\circ/\text{s}\}$.

It should be emphasized that the efficacy of a model depends also on the other models in the set.

Fig. 19 shows (over 100 Monte-Carlo runs) that the use of two model sets $M_1 = \{0, 3^\circ/\text{s}\}$ and $M_2 = \{0, -3^\circ/\text{s}\}$ is not effective: It has a large ambiguous (i.e., no decision) region, which is roughly $(-1.3^\circ/\text{s}, 1.3^\circ/\text{s})$. Nevertheless, the test performed quite well — the percent of incorrect decision was very small (below 2.5°/s) for a true turn rate lower than 1.3°/s and zero for a higher rate. In the figures, the results from MMS-SPRT and MMS-SLRT are labeled as "probability" and "likelihood," respectively. Note the similarity between Fig. 19 and the left part of Fig. 18.

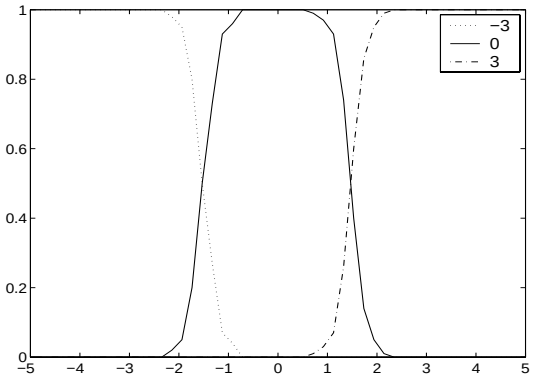


Fig. 17: Test-based model efficacies $w_i^T(s)$ for set $M = \{0, \pm 3^\circ/\text{s}\}$.

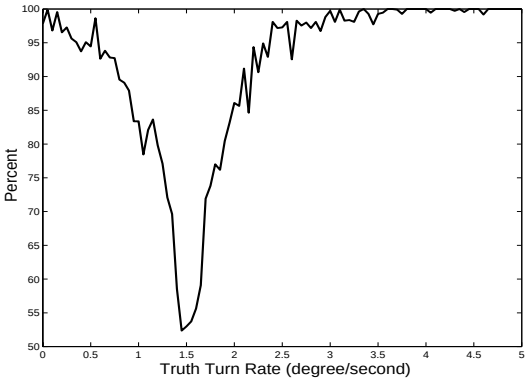


Fig. 18: Percent of correct model selection.

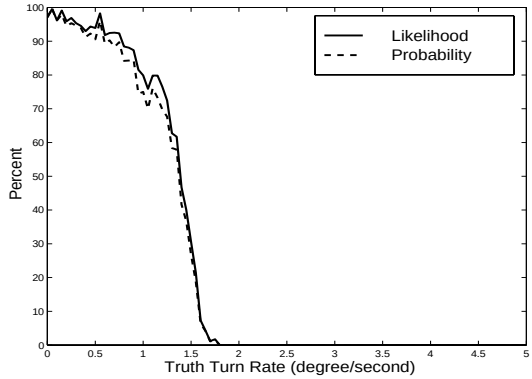


Fig. 19: Percent of no decision.

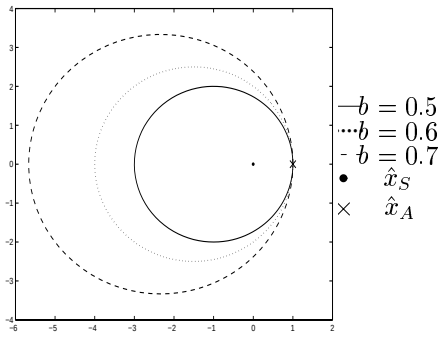


Fig. 20: Illustration of circular criterion for model-set choice.

This similarity makes good sense: The ambiguous region $(-1.3^\circ/\text{s}, 1.3^\circ/\text{s})$ is roughly the effective region of the CV model, which belongs to both M_1 and M_2.

13 Model-Set Choice for Base-State Estimation

It was shown in [18] that for two arbitrary model sets A and B with $A \subset B$, $\|\hat{x}_S - \hat{x}_B\| \leq \|\hat{x}_S - \hat{x}_A\|$ holds if and only if

$$r \leq r_t = \frac{\sqrt{b^2 \cos^2\theta + 1 - b^2} - b\cos\theta}{1-b} \qquad (47)$$

where

$$r = \frac{\|\hat{x}_S - \hat{x}_C\|}{\|\hat{x}_S - \hat{x}_A\|}, \quad \cos\theta = \frac{(\hat{x}_S - \hat{x}_A)'\,(\hat{x}_S - \hat{x}_C)}{\|\hat{x}_S - \hat{x}_A\|\,\|\hat{x}_S - \hat{x}_C\|}$$

and $\hat{x}_C$ is the estimator based on model-set difference $C = B - A$, that is, those models in B but not in A.

The geometric interpretation of this criterion is simple and interesting: Refer to Fig. 20. Model set B is better than set A if and only if the estimators based on those models in B but not in A falls inside the corresponding circle (a ball if dimension is higher than two) determined by $b = \frac{P\{s=m_i|s\in B\}}{P\{s=m_i|s\in A\}}$ for some $m_i \in A$; that is, b is the ratio of model probability in set B to the model probability in set A for any identical model.

Note that this result requires the knowledge of the optimal estimator $\hat{x}_S$ using the optimal model set. We demonstrate below how this seemingly unrealistic theoretical result can be used for model-set choice and comparison.

Suppose that the true turn rate s at a given time is a discrete random variable with sample (mode) space $\mathbf{S} = \{0, \pm 1^\circ/\text{s}, \pm 2^\circ/\text{s}, \pm 3^\circ/\text{s}, \pm 4^\circ/\text{s}, 5^\circ/\text{s}, \pm 6^\circ/\text{s}\}$ and its pmf is the following discrete version of (42): s is generated by (42) using the rectangular window and rounded to the nearest one of the above 13 possible turn rates. In other words,

$$\begin{aligned} P_i^{\text{true}} &= P\{s = \omega_i | s \in \mathbf{S}\} \\ &= \begin{cases} \int_{5.5}^{\infty} f(s)ds & \omega_i = \pm 6^\circ/\text{s} \\ \int_{\omega_i - 0.5}^{\omega_i + 0.5} f(s)ds & \omega_i \neq \pm 6^\circ/\text{s} \end{cases} \end{aligned} \qquad (48)$$

Note that at the given time the system is a coordinated-turn system governed by (39) with $F_{\text{CT}}(\omega_i)$ given by (41).

Consider two MM estimators $\hat{x}_A$ and $\hat{x}_B$ based on two model sets $A = \{0, \pm 3^\circ/\text{s}\}$ and $B = \{0, \pm 1^\circ/\text{s}, \pm 3^\circ/\text{s}, \pm 7^\circ/\text{s}\}$, respectively. Consider the following model-set choice problem: decide whether model set A is better than set B using criterion $\|\hat{x}_S - \hat{x}_B\| \leq \|\hat{x}_S - \hat{x}_A\|$ or equivalently (47), for the set of scenarios of interest given above in the form of a pmf (48).

To use criterion (47), we need the estimators $\hat{x}_C$ based on model set $C = B - A = \{\pm 1^\circ/\text{s}, \pm 7^\circ/\text{s}\}$ and the optimal MM estimator $\hat{x}_S$, where its models have the same probability mass as the true modes, given by (48); that is, $P_i^S \triangleq P\{m = \omega_i | m \in \mathbf{S}\} = P_i^{\text{true}}, \forall i$. The model probabilities for the other estimators are defined similarly:

$$\begin{aligned} P_i^B &\triangleq P\{m = \omega_i | m \in B\} \\ &= \begin{cases} 2\int_0^{0.5} f(s)ds, & \omega_i = 0 \\ \int_{0.5}^{2} f(s)ds, & \omega_i = \pm 1^\circ/\text{s} \\ \int_2^5 f(s)ds, & \omega_i = \pm 3^\circ/\text{s} \\ \int_5^\infty f(s)ds, & \omega_i = \pm 7^\circ/\text{s} \end{cases} \end{aligned}$$

P_i^A and P_i^C are induced by P_i^B in that they are obtained from P_i^B by deleting the probabilities of the models in B but not in A, C, respectively, and scaling up the remaining model probabilities such that they sum up to one. For example,

$$\begin{aligned} P_i^A &\triangleq P\{m = \omega_i | m \in A\} \\ &= \begin{cases} \frac{2}{c}\int_0^{0.5} f(s)ds & \omega_i = 0 \\ \frac{1}{c}\int_2^5 f(s)ds & \omega_i = \pm 3^\circ/\text{s} \end{cases} \end{aligned}$$

where $c = 2\left[\int_0^{0.5} f(s)ds + \int_2^5 f(s)ds\right]$.

Note that each MM estimator $\hat{x}_A$, $\hat{x}_B$, or $\hat{x}_C$ would be optimal should its model set match exactly the mode space in the sense that there is no approximation in the estimation algorithm used — suboptimality arises only from the fact that none of A, B, and C are equal to $\mathbf{S}$.

For an estimator using model set M, where M could be $\mathbf{S}$, A, B, or C, the average value of the estimate at time T is given by

$$\begin{aligned} \bar{x}_M &= E[\hat{x}_M | m \in M] \\ &= E[E(x|z, m \in M) | m \in M] = E[x | m \in M] \\ &= \sum_{m_i \in M} E[x | m = m_i] P\{m = m_i | m \in M\} \\ &= \sum_{\omega_i \in M} \left[F_{\text{CT}}(\omega_i)\bar{x}_0 + \bar{w}_i\right] P_i^M \end{aligned}$$

where z is the measurement, $x = Fx_0 + w$ follows from (39), and $\bar{x}_0$ is the prior state of the system, assumed be identical for all models. Consider a specific example with (43) and $\bar{x} = [1000, 100, 200, 120]'$, $\bar{w}_i = 0, T = 5$. Then we have

$$\begin{aligned} P_i^S &= \{0.1316, 0.1009, 0.1009, 0.1008, 0.1008, 0.1296, \\ &\quad 0.1296, 0.0807, 0.0807, 0.0202, 0.0202, 0.0021, 0.0021\} \\ P_i^A &= \{0.1901, 0.4050, 0.4050\} \\ P_i^B &= \{0.1316, 0.1462, 0.1462, 0.2804, 0.2804, \\ &\quad 0.0076, 0.0076\} \\ P_i^C &= \{0.4753, 0.4753, 0.0247, 0.0247\} \end{aligned}$$

and thus

$$\begin{aligned} r^2 &= \frac{(\bar{x}_S - \bar{x}_C)'(\bar{x}_S - \bar{x}_C)}{(\bar{x}_S - \bar{x}_A)'(\bar{x}_S - \bar{x}_A)} = 16.7880^2 \\ b &= \frac{P\{m = 0 | m \in B\}}{P\{m = 0 | m \in A\}} \\ &= \frac{P\{m = 3^\circ/\text{s} | m \in B\}}{P\{m = 3^\circ/\text{s} | m \in A\}} = 0.6925 \\ \cos\theta &= \frac{(\bar{x}_S - \bar{x}_A)'(\bar{x}_S - \bar{x}_C)}{\|\bar{x}_S - \bar{x}_A\| \, \|\bar{x}_S - \bar{x}_C\|} = -0.9999 \end{aligned}$$

It follows that

$$16.7880 = r > r_t = 5.5041$$

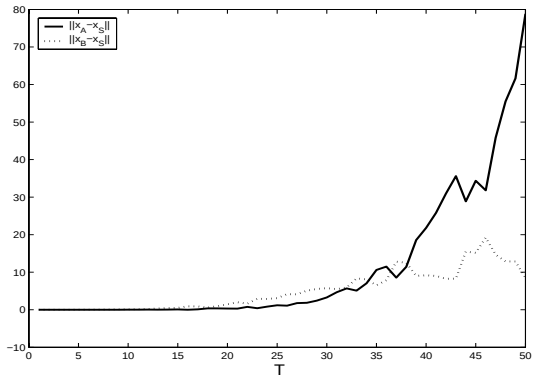
and thus $\|\hat{x}_S - \bar{x}_B\| > \|\hat{x}_S - \bar{x}_A\|$. Consequently, we conclude that model set A is better than set B for this problem. Note that it is hard to say based on our intuition or experience which model set is better. Fig. 20 illustrates the corresponding circular criterion, except that $\bar{x}_C$ should be added at approximately $(16.788, 180^\circ)$.

Fig. 21 shows the RMS position errors of $\hat{x}_S$, $\hat{x}_A$, $\hat{x}_B$ averaged over 500 runs for the time steps $k = 1, 2, \ldots, 10$, which correspond to sampling time $T = 5, 10, \ldots, 50$, respectively. The above numbers correspond to the points of the curves at $T = 5$s. The corresponding curves for the relative merit factor $r - r_t$ of model sets A and B are plotted in Fig. 22 for $T = 5, 10, \ldots, 50$. It is clear that the two figures agree almost perfectly.

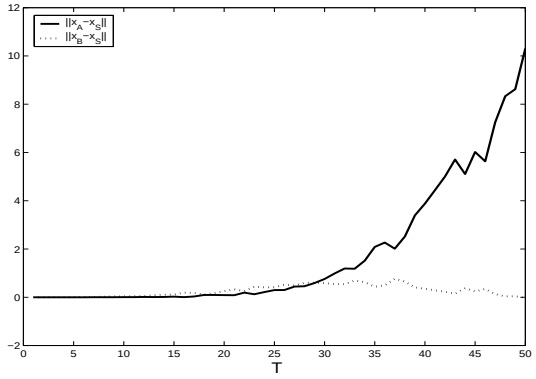
In the simulation, the true modes were generated randomly with a distribution given by (48). They were not known to the estimators, not allowed to jump (for simplicity), and correspondingly AMM estimators were used based on the model sets S, A, B, respectively, assuming that the true mode has a distribution given by their corresponding pmfs P_i^M, $\forall i \in M$, where $M = \mathbf{S}, A$, or B. The initial state of the system was $x_0 = [1000, 100, 200, 120]'$ and each estimator used x_0 as the initial state estimate.

Clearly, the above procedure still works even if the optimal model set $\mathbf{S}$ is large. In many practical problems, the optimal model set for a given set of scenarios of interest is (approximately) known but may be too large to be used in an MM estimator. This example demonstrates how to choose between two model sets given this optimal model set, without using actual measurements or simulation. Clearly, the introduction of an proper probabilistic

model of the scenarios, such as the Gaussian mixture model (42), is a key here. We point out that the circular criterion is also applicable for the cases where measurements are involved.



(a) RMS position errors



(b) RMS velocity errors

Fig. 21: Difference in prediction of MM estimators: $\|\bar{x}_S - \bar{x}_A\|$ and $\|\bar{x}_S - \bar{x}_B\|$.

14 Selection of Estimatee

Consider for simplicity a parameter estimation problem by minimizing $\|p - \hat{p}\|^2$, where p is an unknown parameter.

As an example, suppose we would like to use an MM estimator to determine sampling interval T [28] using the following empirical relation [30, 7]

$$T \approx 0.4 P_D \left[\frac{\sigma_0\sqrt{\tau}}{\sigma}\right]^{0.4} \frac{v_0^{2.4}}{1 + 0.5 v_0^2} \tag{49}$$

where everything except σ is known. Is it good, as proposed in [28], to estimate

$$\widehat{\sigma^2} = \sum_{i=1}^{N} \sigma_i^2 P\{\sigma = \sigma_i | z\} \tag{50}$$

first and then put $\sqrt{\widehat{\sigma^2}}$ into the above formula? The answer is **no**! As pointed out in [13] this estimator is not even unbiased.

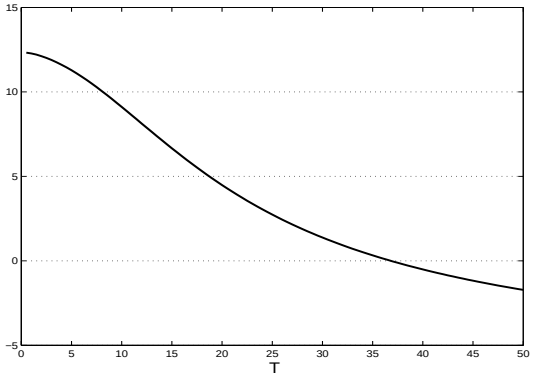


Fig. 22: Relative merit factor $r - r_t$ of model sets A and B versus sampling interval T (A is inferior to B if and only if $r - r_t < 0$).

In fact, if the ultimate goal is to estimate the sampling interval, then it is *best* to use

$$T \approx 0.4 P_D \hat{p} \left[\sigma_0\sqrt{\tau}\right]^{0.4} \frac{v_0^{2.4}}{1 + 0.5 v_0^2}$$

where

$$\hat{p} = \sum_{i=1}^{N} \sigma_i^{-0.4} P\{\sigma^{-0.4} = \sigma_i^{-0.4} | z\}$$

This is because if

$$p = f(\eta), \qquad \hat{p} = E[p|z], \qquad \hat{\eta} = E[\eta|z]$$

then

$$\begin{aligned} \hat{p} &= \sum_{i=1}^{N} p_i P\{p = p_i | z\} \\ &= f\left(\sum_{i=1}^{N} \eta_i P\{\eta = \eta_i | z\}\right) = f(\hat{\eta}) \end{aligned}$$

if and only if f is a linear function.

There are many other advantages of using $\hat{p}$ than using $f(\hat{\eta})$. In general, we should design the model set in the space of p rather than of η. For more details, the reader is referred to [13].

15 Conclusions

Model-set design, choice, and comparison have been considered in a general setting. We have not only argued for the need for and the benefit of probabilistic modeling of the models to be designed as well as the true mode, but also proposed that they be modeled as random variables for offline model-set design, choice, and comparison. Based on such probabilistic models, we have presented the following general results: (a) three general, systematic approaches to

model-set design based on minimizing mismatch in distribution, minimizing modal distance, and moment matching; (b) a variety of optimality criteria and measures for base-state estimation, mode estimation, mode identification, and hybrid-state estimation; (c) several (optimal or data efficient) computationally efficient hypothesis tests for solving representative model-set choice problems; and (d) the concept and two quantification functions of relative efficacy of each model in a set that describes how effectively it covers regions of the mode space. Several simple examples that illustrate how these theoretical results can be used for model-set design, choice and comparison have been given.

Many of the general results presented in this paper are also useful for performance evaluation of MM algorithms.

A Appendix

A.1 Metric Space of Distributions

Let $\mathcal{F}$ be the set of cdfs. By the Lebesgue decomposition, it consists of absolute continuous, piecewise constant, and continuous but not absolute continuous cdfs — which correspond to continuous, discrete, and singular random variables, respectively — and their convex linear combinations. Define the distance between any two cdfs $F_1(x)$ and $F_2(x)$ (i.e., elements of $\mathcal{F}$) by

$$d(F_1, F_2) = \max_{x \in R} |F_1(x) - F_2(x)|$$

This distance definition is legitimate, as seen below. Clearly, it satisfies positive definiteness and symmetry. To see it satisfies triangle inequality, let x_* be such that

$$|F_1(x_*) - F_2(x_*)| = \max_{x \in R} |F_1(x) - F_2(x)|$$

Thus the triangle inequality follows: For any $F_1, F_2, F_3 \in \mathcal{F}$, we have

$$\begin{aligned} d(F_1, F_2) &= |F_1(x_*) - F_2(x_*)| \\ &\le |F_1(x_*) - F_3(x_*)| + |F_3(x_*) - F_2(x_*)| \\ &\le \max_{x \in R} |F_1(x) - F_3(x)| + \max_{x \in R} |F_3(x) - F_2(x)| \\ &= d(F_1, F_3) + d(F_3, F_2) \end{aligned}$$

As a result, $(\mathcal{F}, d)$ is a metric space. Furthermore, we have the following proposition.

Proposition. Let $\mathcal{D}$ be the set of cdfs of discrete random variables. Then, it is dense in $(\mathcal{F}, d)$.

For model-set design, we are interested mainly in the following constructive proof of this proposition, rather than the proposition per se.

Proof. Given any cdf $F \in \mathcal{F}$ and scalar $\epsilon > 0$, there exists a positive integer $N = \lceil 1/2\epsilon \rceil$ = smallest integer not smaller than $1/2\epsilon$, such that $d(F, D) \le \epsilon$, that is,

$$\max_{x \in R} |F(x) - D(x)| \le \epsilon$$

where $D \in \mathcal{D}$ is the cdf of a discrete random variable, defined by the pmf $p(x_i) = \frac{1}{N}, i = 1, \ldots, N$, with

$$\begin{aligned} &x_i = \arg\underset{x \in R}{} \left[F(x) = \frac{i - 1/2}{N} \right], \\ &\text{if } F(x) = \frac{i - 1/2}{N} \text{ has a solution} \\ &x_i = \arg\underset{x \in R}{} \left[F(x^-) < \frac{i - 1/2}{N} < F(x^+) \right], \\ &\text{if } F(x) = \frac{i - 1/2}{N} \text{ has no solution} \end{aligned}$$

Note that in the case where $F(x) = \frac{i-1/2}{N}$ has no solution, x_i may be a repeated point in the sense that

$$\begin{aligned} \frac{i - 1 - 1/2}{N} < F(x^-) < \frac{i - 1/2}{N} < \cdots \\ < \frac{i + j - 1/2}{N} < F(x^+) < \frac{i + j + 1 - 1/2}{N} \end{aligned}$$

(e.g., when $F(x_i)$ has a jump of magnitude greater than ϵ). In this case, we use $x_i = \cdots = x_{i+j}$ and thus $p(x_i) = \cdots = p(x_{i+j}) = \frac{j+1}{N}$. This completes the proof.

Fig. 1 also serves to illustrate this proof.

An important by-product of this proof is that it provides a systematic procedure of finding the cdf of a discrete random variable that has a minimum number N of jumps among all such cdfs that are within a given distance ϵ to any given cdf. This optimality is clear: Corresponding to any other cdf $D' \in \mathcal{D}$ with a number $N' < N$ of jumps, there is always a cdf $F(x)$ that would make the requirement $\max_{x \in R} |F(x) - D'(x)| \le \epsilon$ violated.

Although $\mathcal{D}$ is dense in $(\mathcal{F}, d)$, the latter is not separable with respect to $\mathcal{D}$ because $\mathcal{D}$ is not a denumerable set. It seems that $(\mathcal{F}, d)$ is not separable.

A.2 Mathematical Details of Minimum–Distance Design

Proof of Theorem 6.1. (a) Given any partition $\mathcal{S} = \{S_1, \ldots, S_N\}$ of mode space $\mathbf{S}$, we have, for any model set M with N elements,

$$\begin{aligned} E[d(s, m)] &= \sum_i \int_{S_i} d(s, m_i) dF(s) \\ &= \sum_i P\{s \in S_i\} E[d(s, m_i) | s \in S_i] \\ &\ge \sum_i P\{s \in S_i\} \min_m E[d(s, m) | s \in S_i] \end{aligned}$$

By definition, the (generalized) centroid $m_i = s_i^*$ minimizes $E[d(s, m)|s \in S_i]$. The optimality condition A thus follows.

(b) Note that the nearest-neighbor partition $\{S_1^*, \ldots, S_N^*\}$ of mode space $\mathbf{S}$ has the smallest expected metric given any model set $M = \{m_1, \ldots, m_N\}$:

$$\begin{aligned} E[d(s,m)] &= \sum_i \int_{S_i} d(s,m_i) dF(s) \\ &= \sum_i \int_{S_i^*} d(s,m_i) dF(s) \\ &= \int_{\mathbf{S}} \min_{m \in M} d(s,m) dF(s) \end{aligned}$$

because

$$d(s,m_i) = \min_{m \in M} d(s,m), \quad \forall s \in S_i^*$$

On the other hand, any partition with nonzero probability of the set of points such that $d(s,m) > \min_{m \in M} d(s,m)$ would clearly have

$$\begin{aligned} E[d(s,m)] &= \sum_i \int_{S_i} d(s,m_i) dF(s) \\ &> \int_{\mathbf{S}} \min_{m \in M} d(s,m) dF(s) \end{aligned}$$

where $\{S_1, \ldots, S_N\}$ is an arbitrary partition of the mode space $\mathbf{S}$. For a given pair $\{m_i, m_j\}$ of distinct models, since

$$\begin{aligned} P\{s &\in S_i^*\} \min_{m \in M} E[d(s,m) | s \in S_i^*] \\ +P\{s &\in S_j^*\} \min_{m \in M} E[d(s,m) | s \in S_j^*] \\ +P\{s &\in S_{ij}\} \min_{m \in M} E[d(s,m) | s \in S_{ij}] \\ &= P\{s \in S_i^*\} \min_{m \in M} E[d(s,m) | s \in S_i^*] \\ +P\{s &\in (S_j^* \cup S_{ij})\} \min_{m \in M} E[d(s,m) | s \in (S_j^* \cup S_{ij})] \\ = P\{s &\in (S_i^* \cup S_{ij})\} \min_{m \in M} E[d(s,m) | s \in (S_i^* \cup S_{ij})] \\ +P\{s &\in S_j^*\} \min_{m \in M} E[d(s,m) | s \in S_j^*] \end{aligned}$$

grouping points in S_{ij} with S_i^* or with S_j^* leads to the same $E[d(s,m)]$. Thus, points in S_{ij} may be assigned to either S_i or S_j. Note, however, that this results in different (generalized) centroid pairs $\{s_i^*, s_j^*\}$, either or both of which may differ from the given model pair $\{m_i, m_j\}$. The optimality condition B thus follows.

Proof of Theorem 6.2. (a) By the total expectation theorem, we have

$$\begin{aligned} E[s] &= \sum_i E[s | s \in S_i] P\{s \in S_i\} \\ &= \sum_i m_i P\{s \in S_i\} = \sum_i m_i P\{m = m_i\} = E[m] \end{aligned}$$

where $P\{s \in S_i\} = P\{m = m_i\}$ because S_i is covered by m_i exclusively.

(b) Note first that the random model (with S_i covered by m_i exclusively) can clearly be written as a linear combination of the observables $1(s; S_i)$ (i.e., $m = \hat{s}$, which is random before s is known)

$$m = \sum_i m_i \delta_{m - m_i} = \sum_i m_i 1(s; S_i)$$

As such, model set design (with S_i covered by m_i exclusively) can be viewed a problem of estimating s using observables $1(s; S_i)$; this estimator is always linear whether it is optimal or not. On the other hand, it is well known that a conditional-mean (i.e., centroid) estimator is optimal in the sense of minimizing the MSE $E[(s-m)'(s-m)]$. Such an (optimal linear) estimator satisfies the orthogonality principle: $E[1(s; S_i)(s-m)'] = 0, \forall i$. Therefore,

$$\begin{aligned} E[m(s-m)'] &= E\left[\left(\sum_i m_i 1(s; S_i)\right)(s-m)'\right] \\ &= \sum_i m_i E\left[1(s; S_i)(s-m)'\right] = 0 \end{aligned}$$

(c) It follows directly from (b) that

$$\begin{aligned} E[mm'] &= E[m[s-(s-m)]'] = E[ms'] \\ &= E[[s-(s-m)]m'] = E[sm'] \end{aligned}$$

Taking trace yields $E[m's] = E[s'm] = E[m'm]$.

(d) By (b), we have

$$\begin{aligned} &E[(s-m)(s-m)'] = E[s(s-m)'] \\ = \; &E[ss'] - E[sm'] = E[ss'] - E[mm'] \end{aligned}$$

Taking trace yields $E[(s-m)'(s-m)] = E[s's] - E[m'm]$.

(e) It follows from (b) directly.

A.3 Mathematical Details of Moment-Matching Design

Proof of Theorem 7.1. Since C_m is a positively weighted sum of dyads of $\tilde{m}_i, i \in J$, where $\tilde{m}_i = m_i - \bar{m}$, its rank is equal to the dimension of the linear space spanned by $\tilde{m}_i$, that is, the number of linearly independent vectors in the set $\{\tilde{m}_i, i \in J\}$. It thus follows from the fact $\sum_{i \in J} \tilde{m}_i p_i = 0$ that $\text{rank}(C_m) \leq \min\{|M| - 1, \dim(m)\}$; that is,

$$\min |M| \geq \text{rank}(C_m) + 1$$

For m to match $\bar{s}$ and C_s, we set $\bar{m} = \bar{s}, C_m = C_s$ and thus $\min |M| \geq \text{rank}(C_s) + 1$. The equality holds when we choose a minimal set $\{m_i, i \in \{1, \ldots, \text{rank}(C_s) + 1\}\}$ with as many as $\text{rank}(C_s)$ linearly independent vectors, which can be done, as the corollary of Theorem 7.2 states. This completes the proof.

Proof of Theorem 7.2. Use induction. For $n = 1$, it clearly satisfies (9). Assume that it satisfies (9) for $n = j - 1$; that is,

$$\sum_{i=0}^{j} p_i^{j-1} = 1, \quad \sum_{i=1}^{j} m_i^{j-1} p_i^{j-1} = \mathbf{0},$$
$$\sum_{i=1}^{j} m_i^{j-1} (m_i^{j-1})' p_i^{j-1} = I_{(j-1)\times(j-1)}$$

Then, for $n = j$, we have

$$\begin{aligned}
\sum_{i=0}^{j+1} p_i^j &= p_0 + \sum_{i=1}^{j} p_i^j + p_{j+1}^j \\
&= p_0 + \frac{1}{2}\sum_{i=1}^{j} p_i^{j-1} + (1-p_0)/2 \\
&= 1 \\
\sum_{i=0}^{j+1} m_i^j p_i^j &= \sum_{i=1}^{j} \left[(m_i^{j-1})', (1-p_0)^{-1/2}\right]' \frac{p_i^{j-1}}{2} \\
&\quad + \left[\mathbf{0}, -(1-p_0)^{-1/2}\right]' \frac{1-p_0}{2} \\
&= \left[\mathbf{0}, (1-p_0)^{-1/2}\right]' \sum_{i=1}^{j} \frac{p_i^{j-1}}{2} \\
&\quad + \left[\mathbf{0}, -(1-p_0)^{-1/2}\right]' \frac{1-p_0}{2} \\
&= \mathbf{0}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=0}^{j+1} m_i^j (m_i^j)' p_i^j \\
&= \sum_{i=1}^{j} \left[(m_i^{j-1})', (1-p_0)^{-1/2}\right]' \left[(m_i^{j-1})', (1-p_0)^{-1/2}\right] p_i^{j-1} \\
&\quad + \left[\mathbf{0}, -(1-p_0)^{-1/2}\right]' \left[\mathbf{0}, -(1-p_0)^{-1/2}\right] p_{j+1}^j \\
&= \begin{bmatrix} a & b \\ c & d \end{bmatrix} = I_{j\times j}
\end{aligned}$$

where

$$\begin{aligned}
a &= \sum_{i=1}^{j} m_i^{j-1} (m_i^{j-1})' p_i^{j-1} \\
b &= (1-p_0)^{-1/2} \sum_{i=1}^{j} m_i^{j-1} p_i^{j-1} \\
c &= (1-p_0)^{-1/2} \sum_{i=1}^{j} (m_i^{j-1})' p_i^{j-1} \\
d &= (1-p_0)^{-1} \sum_{i=1}^{j} p_i^{j-1} + (1-p_0)^{-1} p_{j+1}^j
\end{aligned}$$

That is, it satisfies (9) for $n = j$. This completes the proof. Note also that all self skewnesses are clearly equal to zero: $E[m(j)^3] = 0, \forall j \le n$, where $m(j)$ is the jth element of m because

$$\begin{aligned}
E[m(j)^3] &= \sum_{i=0}^{j+1} m_i^j (j)^3 p_i^j \\
&= \sum_{i=1}^{j} (1-p_0)^{-3/2} \frac{p_i^{j-1}}{2} - (1-p_0)^{-3/2} p_{j+1}^j \\
&= 0, \forall j \le n
\end{aligned}$$

Proof of Theorem 7.3. Clearly, the probability masses sum up to unity and the mean is zero. The covariance is given by

$$\begin{aligned}
&\sum_{i=0}^{2n} m_i m_i' p_i \\
&= \sum_{i=1}^{n} \text{diag}\left(\mathbf{0}_{(i-1)\times(i-1)}, \sum_{j=1}^{2} \frac{n}{1-p_0}\frac{1-p_0}{2n}, \mathbf{0}_{(n-i)\times(n-i)}\right) \\
&= I
\end{aligned}$$

This completes the proof.

Proof of Theorem 7.4. Clearly, the probability masses sum up to unity and the mean is zero. The covariance is given by

$$\begin{aligned}
&\sum_{i=0}^{2n} m_i m_i' p_i \\
&= \sum_{i=1}^{n} \text{diag}\left(\mathbf{0}_{(i-1)\times(i-1)}, \sum_{j=1}^{k} \frac{\alpha_j n}{1-p_0}\frac{1-p_0}{\alpha_j\beta_j n}, \mathbf{0}_{(n-i)\times(n-i)}\right) \\
&= I
\end{aligned}$$

This completes the proof.

Proof of Theorem 7.5. Clearly, the probability masses sum up to unity and the mean is zero by symmetry. Without loss of generality, assume the distance between any two adjacent models is 1. Assume a general diamond of k layers. Let each model on the circle of radius r_{ij} have probability p_{ij}. By symmetry, C_l is diagonal. Thus it suffices to show that C_l is proportional to identity matrix I (i.e., $C_l = \alpha I$), that is $C_l(1,1) = C_l(2,2)$ for the 2D case, where $C_l(q,q)$ is the qth diagonal entry of C_l. Now consider models on the circle of radius r_{ij}. Denote their total contribution to covariance by C_{ij}. For odd i with $1 \le j \le (i+1)/2$, the jth model on the ith hexagonal layer is the qth model counting back from the hexagonal vertex, where $q = (i+1)/2 - j$,

and thus we have

$$\begin{aligned} & C_{ij}(1,1)/p_{ij} \\ = \;& 4[((2j-1)/2)^2 + (i-q/2)^2 + (i/2+q/2)^2] \\ = \;& (2j-1)^2 + (2i-q)^2 + (i+q)^2 \\ & C_{ij}(2,2)/p_{ij} \\ = \;& 4[(i\sqrt{3}/2)^2 + (q\sqrt{3}/2)^2 + ((i-q)\sqrt{3}/2)^2] \\ = \;& 3[i^2+q^2+(i-q)^2] \end{aligned}$$

For even i with $2 \le j \le i/2+1$, the jth model on the ith hexagonal layer is the qth model counting back from the hexagonal vertex, where $q = i/2+1-j$, and thus we have

$$\begin{aligned} C_{ij}(1,1)/p_{ij} &= 4[(j-1)^2 + (i-q/2)^2 + (i/2+q/2)^2] \\ &= 4(j-1)^2 + (2i-q)^2 + (i+q)^2 \\ C_{ij}(2,2)/p_{ij} &= 4[(i\sqrt{3}/2)^2 + (q\sqrt{3}/2)^2 + ((i-q)\sqrt{3}/2)^2] \\ &= 3[i^2+q^2+(i-q)^2] \end{aligned}$$

For even i with $j = 1$, we have

$$\begin{aligned} C_{ij}(1,1)/p_{ij} &= 4[(3i/4)^2] \\ C_{ij}(2,2)/p_{ij} &= 2[(i\sqrt{3}/2)^2 + 2(i\sqrt{3}/4)^2] \end{aligned}$$

It can be easily verified that in all cases,

$$C_{ij}(1,1) = C_{ij}(2,2)$$

The case with a higher dimension follows from symmetry. This completes the proof.

References

[1] G. A. Ackerson and K. S. Fu, "On State Estimation in Switching Environments," *IEEE Trans. Automatic Control*, AC-15(1):10–17, Jan. 1970.

[2] E. Balas and M. W. Padberg, "Set Partitioning: A Survey," *SIAM Review*, 18(4):710–761, Oct. 1976.

[3] Y. Bar-Shalom and X. R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT: YBS Publishing, 1995.

[4] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*. New York: Wiley, 2001.

[5] Y. Baram, "Nonstationary Model Validation from Finite Data Records," *IEEE Trans. Automatic Control*, AC-25(1):10–19, Feb. 1980.

[6] Y. Baram and N. R. Sandell, Jr., "An Information Theoretic Approach to Dynamical Systems Modeling and Identification," *IEEE Trans. Automatic Control*, AC-23(1):61–66, Feb. 1978.

[7] S. S. Blackman, *Multiple Target Tracking with Radar Applications*. Norwood, MA: Artech House, 1986.

[8] H. A. P. Blom and Y. Bar-Shalom, "The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients," *IEEE Trans. Automatic Control*, AC-33(8):780–783, Aug. 1988.

[9] M. J. Caputi, "A Necessary Condition for Effective Performance of the Multiple Model Adaptive Estimator," *IEEE Trans. Aerospace and Electronic Systems*, AES-31(3):1132–1139, July 1995.

[10] C. B. Chang and M. Athans, "State Estimation for Discrete Systems with Switching Parameters," *IEEE Trans. Aerospace and Electronic Systems*, AES-14(5):418–425, May 1978.

[11] A. I. El-Fallah, R. P. Mahler, T. Zajic, E. Sorensen, M. G. Alford, and R. K. Mehra, "Scientific Performance Evaluation for Sensor Menagement," in *Signal Processing, Sensor Fusion, and Target Recognition IX*, vol. SPIE 4052, pp. 183–194, 2000.

[12] X. R. Li, "Hybrid Estimation Techniques," in *Control and Dynamic Systems: Advances in Theory and Applications* (C. T. Leondes, ed.), vol. 76, pp. 213–287, New York: Academic Press, 1996.

[13] X. R. Li, "Optimal Selection of Estimatee for Multiple-Model Estimation with Uncertain Parameters," *IEEE Trans. Aerospace and Electronic Systems*, AES-34(2):653–657, Apr. 1998.

[14] X. R. Li, "Engineer's Guide to Variable-Structure Multiple-Model Estimation for Tracking," in *Multitarget-Multisensor Tracking: Applications and Advances* (Y. Bar-Shalom and D. W. Blair, eds.), vol. III, ch. 10, pp. 499–567, Boston, MA: Artech House, 2000.

[15] X. R. Li, "Multiple-Model Estimation with Variable Structure—Part II: Model-Set Adaptation," *IEEE Trans. Automatic Control*, AC-45(11):2047–2060, Nov. 2000.

[16] X. R. Li and Y. Bar-Shalom, "Design of an Interacting Multiple Model Algorithm for Air Traffic Control Tracking," *IEEE Trans. Control Systems Technology*, 1(3):186–194, Sept. 1993. Special issue on Air Traffic Control.

[17] X. R. Li and Y. Bar-Shalom, "Performance Prediction of Hybrid Algorithms," in *Control and Dynamic Systems: Advances in Theory and Applications*, vol. 72 (C. T. Leondes, ed.), vol. 72, pp. 99–151, New York: Academic Press, 1995.

[18] X. R. Li and Y. Bar-Shalom, "Multiple-Model Estimation with Variable Structure," *IEEE Trans. Automatic Control*, AC-41(4):478–493, Apr. 1996.

[19] X. R. Li and V. P. Jilkov, "Expected-Mode Augmentation for Multiple-Model Estimation," in *Proc. 2001 International Conf. on Information Fusion*, (Montreal, QC, Canada), pp. WeB1–3–WeB1–10, Aug. 7-10 2001.

[20] X. R. Li, V. P. Jilkov, J.-F. Ru, and A. Bash, "Expected-Mode Augmentation Algorithms for Variable-Structure Multiple-Model Estimation," in *Proc. IFAC 15th World Congress*, (Barcelona, Spain), July 2002.

[21] X. R. Li and Y. M. Zhang, "Multiple-Model Estimation with Variable Structure—Part V: Likely-Model Set Algorithm," *IEEE Trans. Aerospace and Electronic Systems*, AES-36(2):448–466, Apr. 2000.

[22] X. R. Li, Y. M. Zhang, and X. R. Zhi, "Multiple-Model Estimation with Variable Structure—Part IV: Design and Evaluation of Model-Group Switching Algorithm," *IEEE Trans. Aerospace and Electronic Systems*, AES-35(1):242–254, Jan. 1999.

[23] G. Lorden, "2-SPRT's and the Modified Kiefer-Weiss Problem of Minimizing an Expected Sample Size," *Ann. Statist.*, 4:281–291, 1976.

[24] D. T. Magill, "Optimal Adaptive Estimation of Sampled Stochastic Processes," *IEEE Trans. Automatic Control*, AC-10:434–439, 1965.

[25] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting Multiple Model Methods in Target Tracking: A Survey," *IEEE Trans. Aerospace and Electronic Systems*, AES-34(1):103–123, 1996.

[26] W. H. Ross, "The Expectation of the Likelihood Ratio Criterion," *International Statistical Review*, 55:315–329, 1987.

[27] S. N. Sheldon and P. S. Maybeck, "An Optimizing Design Strategy for Multiple Model Adaptive Estimation and Control," *IEEE Trans. Automatic Control*, AC-38(4):651–654, Apr. 1993.

[28] H.-J. Shin, S.-M. Hong, and D.-H. Hong, "Adaptive-Update-Rate Target Tracking for Phased-Array Radar," *IEE Proc., G*, 142(2), 1995.

[29] D. D. Sworder and J. Boyd, *Estimation Problems in Hybrid Systems*. Cambridge University Press, 1999.

[30] G. VanKeuk, "Software Structure and Sampling Strategy for Automatic Target Tracking with a Phased Array Radar," in *Proc. of AGARD Conference, No. 252*, (Monterey, CA), pp. 11–1–11–13, 1978.

[31] T. Zajic, J. L. Hoffman, and R. P. Mahler, "Scientific Performance Metrics for Data Fusion: New Results," in *Signal Processing, Sensor Fusion, and Target Recognition IX*, vol. SPIE 4052, pp. 172–182, 2000.

[32] Y. M. Zhang and X. R. Li, "Detection and Diagnosis of Sensor and Actuator Failures Using IMM Estimator," *IEEE Trans. Aerospace and Electronic Systems*, AES-34(4):1293–1312, Oct. 1998.