Structure selection of noise covariance matrices for linear Kalman filter design

Federico Bianchi, Simone Formentin, and Luigi Piroddi

***Abstract*— In state reconstruction problems, the statistics of the noise affecting the state equations is often supposed to be known. Since an incorrect description of the model stochastic properties may have detrimental effects on the final filtering performance, many algorithms have been proposed to estimate the noise covariance matrices together with the unknown state. Due to the high computational load, a typical practical assumption is that the process noise covariance can be parameterized as a diagonal matrix. In this paper, we show by counterexamples that this is not always the best compromise between computational complexity and tracking accuracy. Furthermore, a combinatorial optimization algorithm originally employed for model structure selection in nonlinear identification applications is here adapted to the task of selecting the structure of the process noise covariance matrices. The effectiveness of the proposed approach is illustrated by means of some numerical examples.**

I. Introduction

In modern engineering applications involving dynamical systems, the on-line measurement of the state vector is often needed for process monitoring, anomaly detection and/or feedback control. Since such a measurement may not be available for several reasons (*e.g.*, high cost or low reliability of sensors), model-based state reconstruction methods like Kalman filtering are usually employed to estimate the current state values from the corresponding outputs [1]. A major limitation of such an approach is that, even in the linear time-invariant framework, the Kalman estimation of the state is proven to be optimal (*i.e.*, it minimizes the variance of the estimation error) only if the model *coincides* with the mathematical description of the system [2]. Then, in practical applications where no prior knowledge on the physics of the system is available and a state space model needs to be identified from data, modeling errors can jeopardize the filtering performance. Although this is a well known issue in Kalman filtering applications, little effort has been devoted in the system identification community to the development of filtering-oriented stochastic models. To be more precise, many contributions have been dedicated to the problem of identifying the state space matrices, whereas less attention has been given to the modeling of noise statistics. In fact, in many papers the process noise covariance matrix Q and the measurement noise covariance matrix R are provided as *prior knowledge* or their estimation is "declassified" into an *empirical tuning* problem [3]. Recently, the joint estimation of the state and the covariance matrices (CMs) from data has been more thoroughly investigated and several algorithms have been proposed. Some of these techniques are formulated within an *adaptive control* framework, where a feedback mechanism is used to update the CMs based on the quality of state estimation. The available approaches can be classified into 4 groups: correlation methods [4], [5], maximum-likelihood (ML) methods [6], covariance matching methods [7], and Bayesian approaches [8] (see [9] for a nice and comprehensive overview). A common feature of all the existing algorithms is that the CMs are assumed to be *retrievable* from data, *i.e.*, that the data are informative enough to estimate them and that the matrices are correctly parameterized. As it is well known since the work in [4], tuning the CMs may require more degrees of freedom than tuning the optimal gain of a Kalman filter, which poses an identifiability issue on the noise CMs. Indeed, these matrices cannot be estimated univocally in such conditions. This problem can be dealt with in two ways. A first strategy consists in estimating directly the optimal gain through an iterative process, bypassing completely the estimation of Q and R. However, this approach is suitable if only the state estimation is of interest. More in general, the noise model itself is a crucial complement to the process model, which is typically the result of some approximations, and therefore its estimation deserves attention *per se*. A second strategy is that of simplifying the structure of Q and R to match the degrees of freedom of the Kalman gain matrix. For example, a diagonal structure is often adopted for these matrices (thereby assuming incorrelation of the individual noise components), which greatly reduces the number of free elements.

In this paper, we first illustrate the effects related to the mentioned identifiability issue, and then analyze in more detail the structural simplification approach, showing that *the diagonal parametrization of the CMs does not always provide the best compromise between computational complexity and tracking accuracy*. We then propose an algorithm to suitably select the structure of the noise CM. Some examples are illustrated to emphasize the potential benefits of this structural selection algorithm.

The remainder of the paper is as follows. In Section II, the problem of estimating the noise CMs is formally stated. Section III highlights the identifiability issues by means of some simple but significant numerical examples. The case of diagonal parametrization of the process CM is discussed in Section IV, while the algorithm for structure selection is provided in Section V. The paper is ended by some concluding remarks.

All authors are with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, via G. Ponzio 34/5, 20133 Milano, Italy name.surname@polimi.it

II. Problem Statement and Numerical Setup

Consider the state space model of a linear time-invariant discrete-time dynamic stochastic system S with additive white Gaussian noise (shortly WGN):

$$\begin{aligned} x(k+1) &= Fx(k)+v(k) \\ y(k) &= Hx(k)+w(k) \end{aligned} \tag{1}$$

where $x(k) \in \mathbb{R}^n$ denotes the state vector, $y(k) \in \mathbb{R}^p$ is the output vector, $v \sim WGN(0,Q)$ is the process noise with $Q = Q^T \in \mathbb{R}^{n\times n}$ and $Q \succeq 0$ and $w \sim WGN(0,R)$ indicates the measurement noise with $R = R^T \succ 0$ and $R \in \mathbb{R}^{p\times p}$. F and H are the dynamic and output matrices, respectively. We assume that v and w are uncorrelated. To ease the discussion, no deterministic exogenous inputs are here considered, without loss of generality. Let $\hat{x}(k|k)$ be the optimal filter, that is the estimate of $x(k)$ given the outputs y up to the current discrete time instant k. This estimation is carried out in two phases: the predictive phase and the update phase. In the *predictive phase*, the measurements up to the previous time instant $(k-1)$ are used to compute the value of the states (and the outputs) at the current instant k, as well as $P(k)$, which is the covariance matrix of the *state estimation error*:

$$\hat{x}(k|k-1) = F\hat{x}(k-1|k-1) \tag{2}$$

$$\hat{y}(k|k-1) = H\hat{x}(k|k-1) \tag{3}$$

$$P(k|k-1) = FP(k-1|k-1)F^T + Q \tag{4}$$

The *update phase* starts as soon as the measurement relative to the current time instant (that is being estimated) is available. Both the estimates of $x(k)$ and $P(k)$ are adjusted taking into account this new information:

$$\hat{x}(k|k) = \hat{x}(k|k-1) + K(k)\left[y(k) - H\hat{x}(k|k-1)\right] \tag{5}$$

$$P(k|k) = \left[I - K(k)H\right]P(k|k-1) \tag{6}$$

where the Kalman gain $K(k)$ is defined as:

$$K(k) = P(k|k-1)H^T\left[HP(k|k-1)H^T + R\right]^{-1} \tag{7}$$

Notice that the *a priori* estimate of $x(k)$ calculated in the predictive phase is adjusted by a correction term which equals the estimation error $y(k) - H\hat{x}(k|k-1)$, also called *innovation*, weighted by $K(k)$. When both phases are concluded, index k is incremented and the procedure iterated.

Property 1: Let $e(k) = x(k) - \hat{x}(k|k)$ and consider the quadratic loss function:

$$\mathscr{L}(e(k)) = e(k)^T e(k). \tag{8}$$

The filter $\hat{x}(k|k)$ defined in Eq. (5) is the minimum error variance filter, *i.e.,* it minimizes the average loss or risk:

$$\mathscr{R}(\hat{x}(k|k)) \equiv \mathbb{E}\left[\mathscr{L}(e(k))\right], \tag{9}$$

provided that the two CMs are known.

Proof: By Theorem 1 in [10], the minimum error variance filter $\hat{x}(k|k)$ is the conditional expectation

$$\hat{x}(k|k) = \mathbb{E}\left[x(k)|y(k_0),\ldots,y(k)\right],$$

which, under the constraint that the filter is linear, takes the form of Eq. (5) (see Theorem 2 in [10]). ■

Furthermore, it can be easily proved (see Section 3.1 in [11]) that the minimum error variance filter has the following property:

$$\mathscr{R}(\hat{x}(k|k)) = \text{trace}\{P(k|k)\}, \tag{10}$$

which motivates the choice of the trace of the state estimation covariance matrix as an optimality criterion in the sequel.

Property 1 relies on the availability of the true noise CMs, which are typically unknown. This makes the design of Kalman filters hard, as these matrices are required in the design procedure. In this respect, the problem of estimating Q is particularly critical (R can be deduced from the sensor characteristics). This has led to the development of several approaches for estimating the Q from data.

In the following, we denote as $\mathscr{R}(\hat{x}(k|k),Q)$ the risk value associated with the filter $\hat{x}(k|k)$ computed employing the given matrix Q in the design procedure (and using the right R).

III. Estimation of Q

The general idea behind all methods devoted to the *joint* estimation of the state and the CMs, is to exploit the existing relation between the CMs, the state vector $x(k)$ and $P(k|k)$. From Property 1 and relation (10), a possible way to solve the Q estimation problem is to address it as an optimization problem:

$$\underset{Q}{\text{minimize}} \quad \mathscr{R}(\hat{x}(k|k),Q), \tag{11}$$

with $\mathscr{R}(\hat{x}(k|k),Q)$ defined as in (9). Indeed, assuming that the observed data $\mathscr{D} = \{(x(k),y(k)\}_{k=1}^N$ have been generated from (1) using $v \sim WN(0,Q^\circ)$, it follows that

$$\mathscr{R}(\hat{x}(k|k),Q^\circ) \leq \mathscr{R}(\hat{x}(k|k),Q), \forall Q \neq Q^\circ.$$

However, it is not always possible to retrieve Q° from $\mathscr{D}$ due to structural issues. Indeed, as discussed in [4], for a system of order n with p outputs, Property 2 holds.

Property 2: Given a system in the form (1) with n states and p outputs, if

$$p < \lceil \frac{(n+1)}{2} \rceil, \tag{12}$$

then there exist infinite optimal solutions for problem (11).

Proof:

The lower bound in (12) follows by imposing that the DOFs associated to Q be at most equal the number of $K(k)$ components, *i.e.,* $\frac{n(n+1)}{2} \leq np$. ■

Property 2 implies that the q_{ij}, $i,j = 1,\ldots,n$ are not exactly retrievable from data when the estimation of the state *and* the CMs is jointly addressed, if over-parameterization occurs.

Consider *e.g.* the following second order system:

$$S_1 : \begin{cases} x(k+1) = \begin{bmatrix} 0.9 & -0.4 \\ 0.2 & 0.9 \end{bmatrix} x(k) + v(k) \\ y(k) = \begin{bmatrix} 0.5 & 0 \end{bmatrix} x(k) + w(k) \end{cases} \tag{13}$$

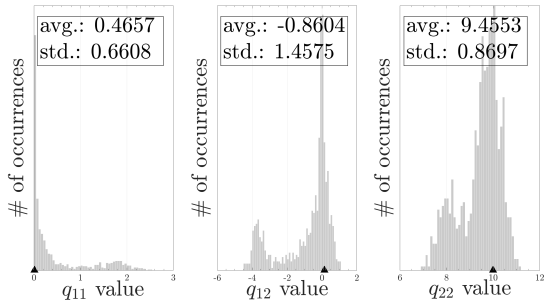

Fig. 1. S_1: Distribution of the optimal values for q_{11}, q_{12}, and q_{22} over 1000 MC runs with different noise realizations. The black triangular markers indicate the true values.

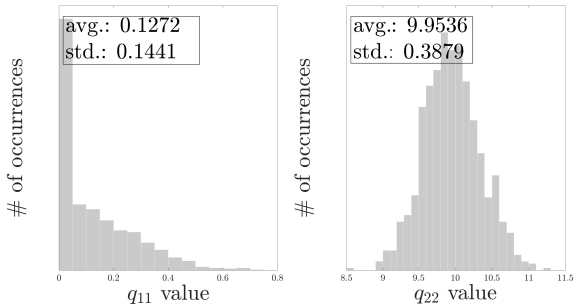

Fig. 2. S_1: Distribution of the optimal values for q_{11} and q_{22} over 1000 MC runs with different noise realizations. In this case, over-parameterization does not occur since a diagonal structure is adopted for Q.

with $Q^\circ = \begin{bmatrix} 0.0125 & 0.15 \\ 0.15 & 10 \end{bmatrix}$, $R = \frac{\text{var}[y(k)]}{10}$, $x(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

A Monte Carlo (MC) study has been carried out analyzing 1000 different noise realizations of length $N = 10000$. On each run, the optimization problem (11) has been solved on an unbounded domain using the `fminsearch` Matlab routine [12]. Figure 1 shows the distribution of the optimal values of the coefficients of Q w.r.t. the noise realization. In general, there is a bias in the estimation of Q° induced by the specific noise realization. Furthermore, the distributions show other peaks different from the true values. This experimental evidence demonstrates that it is not always possible to deduce directly from $\mathscr{D}$ the Q°, not even asymptotically, when the estimation of the state *and* the CMs is jointly addressed, if over-parameterization occurs. Indeed, for the considered system S_1, the Kalman gain $K^\circ(k)$ is a 2×1 vector, whereas the Q matrix has 3 degrees of freedom (DOFs). Therefore, the same $K^\circ(k)$ may be obtained for several values of Q (which are all optimal).

Note that, despite the result stated in Property 2, answering to how many and under which conditions the noise CMs elements can be estimated is still missing in literature. This motivated the interest in estimating directly the optimal gain of a Kalman filter independently on the knowledge of Q, as proposed in [4], or in the usage of a Q with a simplified structure. Indeed, notice that condition (12) is automatically satisfied if *e.g.,* a diagonal structure is chosen for Q, since in that case Q has only n DOFs which is at most equal to the number np of elements of the Kalman gain (equal when $p = 1$). Figure 2 exemplifies this concept, where the experiment on S_1 is repeated with a diagonal Q, *i.e.* with only 2 DOFs.

Discussed the existence of an identifiability issue, from now on we focus on the more general problem of choosing a suitable structure for Q when it is unknown. This motivates the selection of the numerical optimization in (11), w.r.t correlation methods which provide an analytical solution to the Q estimation problem but cannot handle structural constraints.

IV. Which structure for Q?

A common practice adopted by several methods is to simplify the structure of matrix Q to be diagonal, with obvious computational advantages. While this often works out satisfactorily because it reduces the degrees of freedom of Q to n, it is not always the most appropriate choice leading to an unnecessary loss of filtering accuracy. To investigate this issue in more detail, we here address the following constrained optimization problem:

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & \mathscr{R}(\hat{x}(k|k), Q) \\ \text{subject to} \quad & Q \in \mathscr{Q}_\kappa, \end{aligned} \tag{14}$$

where $\mathscr{Q}_\kappa$ is the set of all positive semidefinite matrices $Q \in \mathbb{R}^{n\times n}$ with $\kappa \leq \frac{n(n+1)}{2}$ free parameters. To ensure that $Q = Q^T \succeq 0$, Q is parameterized using a Cholesky decomposition

$$Q = ZZ^T, \tag{15}$$

where Z is a lower triangular matrix, and the optimization process is carried out over the elements z_{ij} of Z.

As a result, the structural constraints on Q translate to conditions on the parameters of the Cholesky factors. For example, for $n = 3$ one has that:

$$Q = ZZ^T = \begin{bmatrix} z_{11}^2 & z_{11}z_{21} & z_{11}z_{31} \\ z_{21}z_{11} & z_{21}^2 + z_{22}^2 & z_{21}z_{31} + z_{22}z_{32} \\ z_{31}z_{11} & z_{31}z_{21} + z_{32}z_{22} & z_{31}^2 + z_{32}^2 + z_{33}^2 \end{bmatrix}. \tag{16}$$

Therefore, the constraint $z_{21}z_{31} + z_{22}z_{32} = 0$ must be applied in the optimization problem to ensure that $q_{23} = q_{32} = 0$, and so on. More simply, a diagonal Q is obtained by setting Z to be diagonal ($z_{21} = z_{31} = z_{32} = 0$).

Consider now the third order system described below:

$$S_2 : \begin{cases} x(k+1) = \begin{bmatrix} 0.0218 & 0.9243 & -0.2750 \\ 0.4645 & -0.2466 & -0.8076 \\ 0.8451 & 0.1167 & 0.4530 \end{bmatrix} x(k) + v(k) \\ y(k) = \begin{bmatrix} 0.9936 & 0 & 0.6539 \end{bmatrix} x(t) + w(k) \end{cases} \tag{17}$$

with

$$Q^\circ = \begin{bmatrix} 6.3557 & 6.2921 & -0.7910 \\ 6.2921 & 6.5128 & -0.0420 \\ -0.7910 & -0.0420 & 6.7963 \end{bmatrix},$$

$R = \text{var}[y(k)]/10$, $x(0) = [0\ 0\ 0]^T$. As done previously, an observation window of $N = 10000$ samples has been considered.

An exhaustive analysis over all possible matrix structures has been carried out and the aggregated results are reported in Figure 3. In particular, for each structure, 100 MC runs have been carried out on the same data realization for different random initializations of the optimization solver, and the identified optimal Qs have been validated over a test set. In this plot, the black marker refers to the best diagonal solution, while the red marker refers to $\mathscr{R}(\hat{x}(k|k), Q^{\circ})$. Observing the parameterizations with 3 DOFs, one can note that the full diagonal case does not represent the best solution and more efficient *non-diagonal* structures can be pursued with the same number of parameters. Indeed, the minimum $\mathscr{R}(\hat{x}(k|k), Q)$ associated to the diagonal case (*d*), computed on a test set, is 42.6770 w.r.t. 40.3954 (i.e., a relative error of 5.6482%), which corresponds to the best *non-diagonal* (*nd*) structure with 3 DOFs. The two corresponding identified matrices are listed below:

$$Q_d = \begin{bmatrix} 2.3232 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 11.7474 \end{bmatrix}, \; Q_{nd} = \begin{bmatrix} 6.2337 & 8.2759 & 0 \\ 8.2759 & 10.9879 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It is worth noting that both these matrices make the pair $\left(F, \sqrt{Q}\right)$ uncontrollable, which results in the Kalman filter neglecting the effect of new measurements on the state corresponding to the row of zeros. Unusual as it may seem, this is perfectly legitimate and does not limit *per se* the applicability of the Kalman filter (see Section 3.1 in [13]).

As for S_1, an identifiability issue arises also for S_2 due to over-parameterization, when the optimization is carried out on a Q matrix with more than 3 DOFs. This prevents one from obtaining the "true" Q in the unconstrained case (*i.e.* with 6 DOFs).

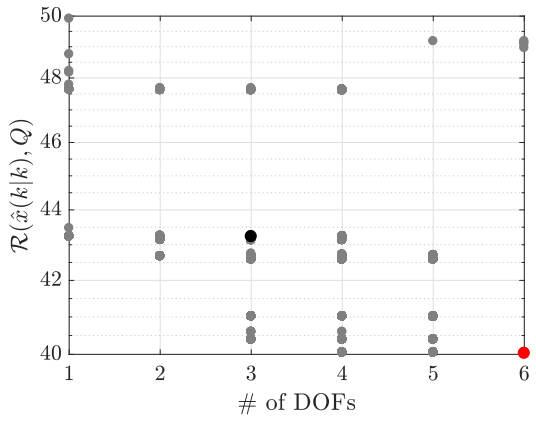

Fig. 3. S_2: estimation error corresponding to different parameterizations. The black marker represents the diagonal case, while the red one the case $\mathscr{R}(\hat{x}(k|k), Q^{\circ})$. The reported values have been computed on a test data-set.

V. Matrix Structure Selection

The analysis carried out in the previous sections shows that while the Q matrix should be endowed with a limited number of degrees of freedom, the diagonal structure does not necessarily always provide the best possible low-complexity approximation of the real Q. Building on this observation, we next propose an approach to address the Matrix Structure Selection (MSS) of Q. In particular, we take inspiration from a recent work presented in [14] addressing the model structure selection problem for Nonlinear Auto-Regressive with eXogenous (NARX) models in a probabilistic framework. To facilitate the explanation, the probabilistic framework will be presented directly in the context of this paper. The interested reader can find the proof of the local convergence properties of this approach in the original paper [14].

A. A probabilistic approach to MSS

Let $\vartheta = [z_{11}\ z_{21}\ z_{22}\ ...]^T$ be a vector containing all the $N_T = \frac{n(n+1)}{2}$ parameters z_{ij}, $i \geq j$, in lexicographical order. Let also s be a binary vector taking values in the set $\Sigma = \{0,1\}^{N_T}$, where the k-th element s_k encodes the presence (or absence) of ϑ_k in the model structure. For example, a third order Q matrix with the following structure:

$$Q = \begin{bmatrix} q_{11} & 0 & 0 \\ 0 & q_{22} & q_{23} \\ 0 & q_{23} & q_{33} \end{bmatrix}$$

can be obtained by selecting $\vartheta_1 = z_{11}$, $\vartheta_3 = z_{22}$, $\vartheta_5 = z_{32}$ (see (16)), *i.e.* by employing a Z factor with the structure $s = [1\ 0\ 1\ 0\ 1\ 0]$. Notice that the same structure for Q could have also been achieved using a larger structure $s = [1\ 0\ 1\ 0\ 1\ 1]$, encompassing also the term z_{33}, but the inclusion of this additional parameter does not lead to any improvements in terms of accuracy (see again the Cholesky factor in (16)).

We aim to find the matrix Q with the *smallest number of non-zero parameters* that provides an "acceptable" performance. Accordingly, we can formulate the structure selection problem as follows:

$$\underset{s}{\text{minimize}} \quad \mathscr{R}(\hat{x}(k|k), Q(s)) + \lambda \mathscr{P}(s), \tag{18}$$

where $\mathscr{R}(\hat{x}(k|k), Q(s))$ is defined in (9), $Q(s)$ being a symmetric semidefinite positive matrix with structure induced by s, $\mathscr{P}(s) : \Sigma \to \mathbb{R}^+$ is a penalization term, and parameter λ determines the relative importance of the two sub-objectives. Since s is a binary vector, the most intuitive choice for the penalization term is the zero-norm $\|s\|_0 = \text{card}\{s_i : s_i \neq 0\}$, which actually coincides in this specific case with the one-norm $\|s\|_1 = \sum_{i=1}^{N_T} |s_i|$. In view of the convexity of the one-norm, this is preferred to the zero-norm [15].

Without loss of generality, we will reformulate problem (18) as follows

$$\underset{s}{\text{maximize}} \quad J(s) = e^{-\beta \cdot [\mathscr{R}(\hat{x}(k|k), Q(s)) + \lambda \mathscr{P}(s)]}, \tag{19}$$

where $\beta > 0$ is a design parameter. The exponential form of the cost function provides a natural normalization in the interval $(0,1]$, and allows a sharper discrimination between structures with similar performance. In the following, we will denote as $s^\star$ the optimal solution of problem (19). We will further assume that $J(s) < J(s^\star)$, $\forall s \in \Sigma \setminus \{s^\star\}$. An exhaustive solution of the optimization problem (19) for all possible

values of s is not convenient for high order systems, due to the complexity of the combinatorial part of the problem. Indeed, there are 2^{N_T} different values for vector s, and in turn N_T grows as the square of the system order n. For this purpose we here introduce an heuristic approach to the solution of the optimization problem.

Following [14], problem (19) is reformulated in a probabilistic form, by assigning a probability to each possible value of the structure vector to be the target solution $s^\star$. Starting from an initial tentative distribution (*e.g.,* such that all possible structures have equal probability), we then employ a *sample-and-evaluate* approach to progressively update the probability distribution, enhancing the probability that high-performance structures are extracted. This iterative process terminates when the distribution converges to a limit distribution, where one structure has probability 1 (and all the others have probability 0).

To this end, let σ be a discrete variable taking values in Σ according to a probability distribution $\mathbb{P}_\sigma$. The average performance of σ can be evaluated as follows:

$$\mathbb{E}[J(\sigma)] = \sum_{s\in\Sigma} \mathbb{P}_\sigma(s) J(s), \tag{20}$$

which is a convex combination of the performance of all structures in Σ. If we let $\mathbb{P}_\sigma$ vary over all possible distributions on Σ, the maximum value of (20) is obtained by making all probability mass concentrate on the target structure $s^\star$ (denoting such distribution as $\mathbb{P}^\star_\sigma = \arg\max_{\mathbb{P}_\sigma} \mathbb{E}[J(\sigma)]$, it holds that $\mathbb{P}^\star_\sigma(s) = 1$ for $s = s^\star$, and 0 otherwise). Thus, the optimization problem

$$\underset{s\in\Sigma}{\text{maximize}} \quad \mathbb{P}^\star_\sigma(s), \tag{21}$$

provides the same solution of problem (19). A key feature of the proposed algorithm is the adopted parametrization for $\mathbb{P}_\sigma$. The idea is to associate each term ϑ_i with a Bernoulli random variable $\rho_i \sim Be(\mu_i)$, where μ_i is the success probability of ρ_i. Accordingly, ϑ_i is present in the extracted Cholesky factor Z (*i.e.* $s_i = 1$) if $\rho_i = 1$ and is absent if $\rho_i = 0$. Parameter μ_i represents the confidence level that ϑ_i belongs to the target structure $s^\star$. The vector $\mu = [\mu_1, \ldots, \mu_{N_T}]$ induces the following probability distribution over the solution space Σ:

$$\mathbb{P}_\sigma(s) = \prod_{i:s_i=1} \mu_i \prod_{i:s_i=0} (1-\mu_i). \tag{22}$$

Therefore, one can optimize (21) by refining μ so as to concentrate $\mathbb{P}_\sigma$ onto $s^\star$. To do so, we employ a sampled version of the index:

$$\ell_i = \mathbb{E}[J(\sigma)|\sigma_i = 1] - \mathbb{E}[J(\sigma)|\sigma_i = 0], \tag{23}$$

which evaluates for each term ϑ_i the average performance of the structures that contain it, weighted with the respective probabilities as given by $\mathbb{P}_\sigma$, and compares it with the corresponding average performance of the remaining models weighted in probability. Theorem 1 proves that if the probability distribution induced by μ is not far from $\mathbb{P}^\star_\sigma$, then $\ell_i > 0$ iff $s^\star_i = 1$.

Theorem 1 (Local convergence - Part I, [14]): Let $\mathbb{P}_\sigma$ be the probability distribution induced by μ according to (22). Then, there exists $\delta \in (0,1)$ s.t. if $\mathbb{P}_\sigma(s^\star) \geq \delta$ it holds that $\ell_i > 0$ if $s^\star_i = 1$ and $\ell_i < 0$ otherwise.

Index (23) can thus be employed to progressively refine μ_i, $i = 1, \ldots, N_T$, by increasing the values of the parameters μ_i for which $\ell_i > 0$. In practice, since the expected values in (23) cannot be calculated exactly one has to resort to a sampled approximation. More precisely, at each iteration, several sample structures are extracted from $\mathbb{P}_\sigma$ (each structure is obtained by extracting one sample for each Bernoulli distribution) and evaluated. The latter operation amounts to solving problem (14) with the proper constraints on the structure of Q, and yields the parameters of the matrix Q with the assigned structure that maximizes the filter performance on the collected data-set of N state-measurement pairs $\mathscr{D}$. Then, one estimates the ℓ_i indices according to (23) approximating the expected values with averages over the extracted samples. Finally, one updates μ_i, for each term ϑ_i, according to the sign of the index ℓ_i (μ_i is increased to reinforce the probability of selecting that term in the subsequent steps if $\ell_i > 0$). The algorithm terminates when a limit distribution $\mathbb{P}^\star_\sigma$ is obtained, with all the μ_i in $\{0,1\}$, thus identifying a unique structure $s^\star$. The update equation for μ_i is as follows:

$$\mu_i(r+1) = \mu_i(r) + \gamma \cdot \ell_i, \tag{24}$$

where γ is a step size and r denotes the current iteration.

Theorem 2 (Local convergence - Part II, [14]): Let $\mu(r)$ be s.t. $\mathbb{P}^{(r)}_\sigma(s^\star) \geq \delta$, where δ is a value for which Theorem 1 holds. Then, the local convergence to the target limit distribution is guaranteed by the iterative application of (24) starting from $\mu(r)$.

To guarantee that parameters μ_i remain in the $[0,1]$ interval, the result of (24) is always saturated *a posteriori*. The overall algorithm is reported in Algorithm 1 in the next page.

B. An application of the MSS algorithm

In this example we consider a system S_3 in the form of (1) with 5 states and outputs, described by the following matrices:

$$F = \begin{bmatrix} 0.0964 & 0.1577 & 0.2783 & -0.0983 & 0.1116 \\ 0.1354 & 0.0687 & 0.1230 & -0.1257 & -0.2379 \\ 0.2974 & 0.0913 & -0.1032 & -0.0937 & 0.0525 \\ -0.1010 & -0.1266 & -0.0891 & 0.4644 & -0.0420 \\ 0.0863 & -0.2380 & 0.0884 & -0.0410 & -0.0409 \end{bmatrix},$$

$$H = \begin{bmatrix} 0.2380 & -0.5470 & 2.0726 & 0 & 1.4756 \\ -0.0458 & 0 & -0.7593 & 0.9201 & -1.5044 \\ 0.0523 & 1.7044 & -1.1369 & -0.0254 & 0.8159 \\ 0 & -0.1391 & 0 & -1.4746 & -0.3703 \\ 0.1182 & 0.0666 & 0.3504 & 2.1646 & 0.1038 \end{bmatrix}.$$

The data were generated using a sparse Q matrix:

$$Q = \begin{bmatrix} 0.3854 & 0.1160 & 0 & -0.4516 & 0 \\ 0.1160 & 0.3225 & 0 & -0.0832 & -0.0643 \\ 0 & 0 & 0.1000 & 0 & 0 \\ -0.4516 & -0.0832 & 0 & 0.6249 & -0.0165 \\ 0 & -0.0643 & 0 & -0.0165 & 0.0251 \end{bmatrix},$$

Algorithm 1 MSS algorithm

Require: $\{(x(k), y(k)), k = 1, \cdots, N\}$, $N_T, N_p, \beta, \lambda, \gamma, \mu_{min}, \mu_{max}$
Ensure: μ

$\mu \leftarrow \frac{1}{N_T} \cdot \mathbf{1}_{N_T \times 1}$;
repeat
  for $p = 1$ **to** N_p **do** ▷ Generate structures
    $s^{(p)} = [\,]$;
    for $i = 1$ **to** N_T **do**
      Extract t_i from Be(μ_i); ▷ Generate terms
      $s^{(p)} \leftarrow [s^{(p)},\, t_i]$;
    end for
    Define $Q^{(p)}$ according to structure $s^{(p)}$;
    $\{Q^{(p)}, \mathscr{R}(\hat{x}(k|k), Q^{(p)})\}$ CM estimation;
    $J^{(p)} \leftarrow e^{-\beta \cdot \left[\mathscr{R}(\hat{x}(k|k), Q^{(p)}) + \lambda \mathscr{P}(s^{(p)})\right]}$; ▷ Structure evaluation
  end for
  for $i = 1$ **to** N_T **do** ▷ Update μ_i
    $J^{\oplus} \leftarrow 0$; $n^{\oplus} \leftarrow 0$; $J^{\ominus} \leftarrow 0$; $n^{\ominus} \leftarrow 0$;
    for $p = 1$ **to** N_p **do**
      if $s_i^{(p)} = 1$ **then**
        $J^{\oplus} \leftarrow J^{\oplus} + J^{(p)}$; $n^{\oplus} \leftarrow n^{\oplus} + 1$;
      else
        $J^{\ominus} \leftarrow J^{\ominus} + J^{(p)}$; $n^{\ominus} \leftarrow n^{\ominus} + 1$;
      end if
    end for
    $\mu_i \leftarrow \mu_i + \gamma\left(\frac{J^{\oplus}}{\max(n^{\oplus},1)} - \frac{J^{\ominus}}{\max(n^{\ominus},1)}\right)$;
    $\mu_i \leftarrow \max\left(\min\left(\mu_i, \mu_{max}\right), \mu_{min}\right)$; ▷ Saturation
  end for
until $\max_{i=1,\ldots,N_T} \left(\max\{\mu_{max} - \mu_i,\, \mu_i - \mu_{min}\}\right) \le \varepsilon$ ▷ Stopping criterion

and

$$R = \mathrm{diag}\left(0.1055,\, 0.1402,\, 0.0978,\, 0.1916,\, 0.3802\right).$$

A sparse Q can be representative, *e.g.*, of distributed systems with various interconnected devices, whereby only state variables associated to a device or to neighboring devices are correlated.

To account for the randomization inherent in the MSS algorithm 1, an MC study has been carried out by running the algorithm 100 times on the same data realization. The algorithm has been set up with $N_p = 40$, $\beta = 1$, $\mu_{min} = 0.001$, and $\mu_{max} = 0.999$. A common issue in solving optimization problems which include regularization terms is the selection of the optimal regularization weight λ. A popular method for the selection of this parameter is the *L-curve criterion* [16] [17]. It is based on the study of the *L-curve* which is a log-log plot of the norm of a regularized solution versus the norm of the corresponding residual. It is a convenient graphical tool for displaying the trade-off between the size of a regularized solution and its fit to the given data, as the regularization parameter varies. According to the performed *L-curve* analysis (see Figure 4), a λ value equal to 0.02 is used as it represents a good trade-off. Finally, as suggested in [14], we adopt a time varying step size γ with the following law:

$$\gamma = \frac{1}{10\left(\mathtt{max}\left(J_p\right) - \mathtt{mean}\left(J_p\right)\right) + 0.1}.$$

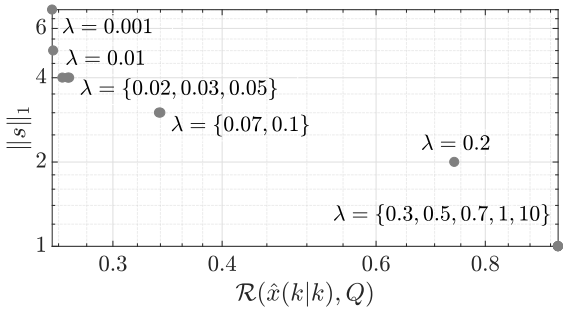

Fig. 4. MSS - L-*curve*. The values have been computed on the training data-set.

The aggregated results are reported in Table I. One can note that the most frequently selected structure corresponds to Q matrices with 5 DOFs in the form:

$$Q = \begin{bmatrix} q_{11} & 0 & 0 & q_{14} & 0 \\ 0 & q_{22} & 0 & 0 & 0 \\ 0 & 0 & q_{33} & 0 & 0 \\ q_{14} & 0 & 0 & q_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

In the remaining 3% of cases, the algorithm selected the following structures (defined in terms of the z_{ij} terms):
$(z_{11},\, z_{41},\, z_{22},\, z_{42},\, z_{43})$,
$(z_{11},\, z_{41},\, z_{22},\, z_{42},\, z_{33})$,
$(z_{11},\, z_{21},\, z_{22},\, z_{42},\, z_{43})$.
It is worth mentioning that the algorithm converges to the mentioned solution by exploring a tiny fraction of the Σ space, *i.e.,* 611.29 explored solutions (on average) w.r.t. 32768 possible structures.

Figure 5 shows a comparison between a portion of the real x_1 trajectory and those estimated by employing in the filter design the true Q (denoted Q_{true}), the CM estimated based on the structure suggested by the MSS algorithm (Q_{opt}), as well as the optimal diagonal one (Q_{diag}). The curves displayed in Figure 5 have been computed on a test data-set. Apparently, the lack of information about q_{14} in the diagonal structure, prevents the filter from estimating accurately the x_1 trajectory. On the other hand, the filter based on the structure-optimized Q_{opt} yields state estimates very close to those obtained with Q_{true}, thus motivating such a selection. This also reflects on the corresponding values of $\mathscr{R}(\hat{x}(k|k), Q)$, respectively $\mathscr{R}(\hat{x}(k|k), Q_{\text{true}}) = 0.2280$, $\mathscr{R}(\hat{x}(k|k), Q_{\text{opt}}) = 0.2557$, and $\mathscr{R}(\hat{x}(k|k), Q_{\text{diag}}) = 0.5123$.

Average elapsed time [s]	381.34
Average number of iterations	50.19
Average # of explored matrix structures	611.29
Total # of possible matrix structures	$2^{15} = 32768$
Selected structure (z_{ij} terms)	$(z_{11},\, z_{41},\, z_{22},\, z_{33})$
Extraction of the selected structure [%]	97

TABLE I
MSS - MONTE CARLO STUDY RESULTS.

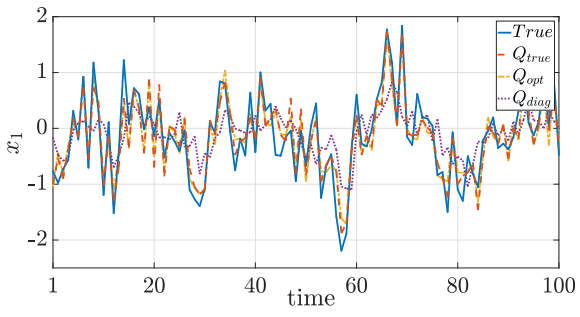

Fig. 5. MSS - State estimation: real and estimated trajectories for state variable x_1. The state estimates are obtained by employing Q_{true}, Q_{opt}, and Q_{diag}. All the trajectories are computed on a test data-set.

VI. Conclusions

In this paper, we investigated three main aspects regarding the identification of the noise process CM in Kalman filtering problems. First, we discussed the identifiability issues associated to the estimation of Q from data, related to the over-parametrization in the multidimensional case. Indeed, the tuning of the CMs is based on the estimation of the Kalman gain, which usually provides less degrees of freedom than those associated with the CM. This key observation opened a window onto the problem of choosing a suitable parametrization for Q. Accordingly, as a second point of discussion, we questioned the common practice of assuming a diagonal structure for the CM, showing that this is not always the best choice. Finally, building upon the last observation, we developed a method for solving explicitly the structure selection problem, by decoupling it from the estimation problem.

References

[1] M. S. Grewal, "Kalman filtering," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 705–708.

[2] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of basic engineering*, vol. 83, no. 1, pp. 95–108, 1961.

[3] S. Formentin and S. Bittanti, "An insight into noise covariance estimation for kalman filter design," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 2358–2363, 2014.

[4] R. Mehra, "On the identification of variances and adaptive kalman filtering," *IEEE Transactions on automatic control*, vol. 15, no. 2, pp. 175–184, 1970.

[5] J. Duník, O. Straka, and M. Šimandl, "On autocovariance least-squares method for noise covariance matrices estimation," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 967–972, 2017.

[6] V. A. Bavdekar, A. P. Deshpande, and S. C. Patwardhan, "Identification of process and measurement noise covariance for state and parameter estimation using extended kalman filter," *Journal of Process control*, vol. 21, no. 4, pp. 585–601, 2011.

[7] X. Wang, M.-q. Liu, Z. Fan, and S.-l. Zhang, "A novel approach of noise statistics estimate using h infinity filter in target tracking," *Frontiers of Information Technology & Electronic Engineering*, vol. 17, no. 5, pp. 449–457, 2016.

[8] P. Matisko and V. Havlena, "Noise covariance estimation for kalman filter tuning using bayesian approach and monte carlo," *International Journal of Adaptive Control and Signal Processing*, vol. 27, no. 11, pp. 957–973, 2013.

[9] J. Duník, O. Straka, O. Kost, and J. Havlík, "Noise covariance matrices in state-space models: A survey and comparison of estimation methods - Part i," *International Journal of Adaptive Control and Signal Processing*, vol. 31, no. 11, pp. 1505–1543, 2017.

[10] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[11] B. D. Anderson and J. B. Moore, "Optimal filtering," *Englewood Cliffs*, vol. 21, pp. 22–95, 1979.

[12] "Matlab optimization toolbox," Version 7.6, the MathWorks, Natick, MA, USA.

[13] B. Southall, B. F. Buxton, and J. A. Marchant, "Controllability and observability: Tools for kalman filter design." in *BMVC*, 1998, pp. 1–10.

[14] A. Falsone, L. Piroddi, and M. Prandini, "A randomized algorithm for nonlinear model structure selection," *Automatica*, vol. 60, pp. 227–238, 2015.

[15] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[16] C. Lawson and R. J. R, Hanson, *Solving least squares problems*. Siam, 1995, vol. 15.

[17] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the l-curve," *SIAM review*, vol. 34, no. 4, pp. 561–580, 1992.