A NOTE ON STATE ESTIMATION AS A CONVEX OPTIMIZATION PROBLEM

Thomas Schön, Fredrik Gustafsson, Anders Hansson

Department of Electrical Engineering
Linköping University, SE-581 83 Linköping, Sweden
Email: (schon, fredrik, hansson)@isy.liu.se

ABSTRACT

The Kalman filter computes the *maximum a posteriori* (MAP) estimate of the states for linear state space models with Gaussian noise. We interpret the Kalman filter as the solution to a convex optimization problem, and show that we can generalize the MAP state estimator to any noise with log-concave density function and any combination of linear equality and convex inequality constraints on the states.

We illustrate the principle on a hidden Markov model, where the state vector contains probabilities that are positive and sum to one.

1. INTRODUCTION

State estimation in stochastic linear models is an important problem in many model-based approaches in signal processing and automatic control applications, where the Kalman filter is the standard method. However, if we have prior information of some kind it is often impossible to incorporate this in the Kalman filter framework. We will in this paper show how we can use prior information by considering the optimization problem that the Kalman filter solves. A similar treatment can be found in [1], however they only consider quadratic problems, whereas we will consider a larger class of convex problems.

2. CONVEX OPTIMIZATION

In this section we will give a very brief introduction to convex optimization (see also [2]).

The main message in convex optimization is that one should *not* differ between linear and non-linear optimization problems, but instead between convex and non-convex problems. This is due to the fact that the class of convex problems is much larger than that covered by linear problems, and we know that for a convex problem any local optimum is also a global optimum. Moreover, there exist efficient algorithms for solving convex optimization problems. A convex optimization problem is defined as

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \qquad i = 0, \ldots, m \\ & a_i^T x = b_i, \qquad i = 0, \ldots, n \end{aligned} \tag{1}$$

where the functions $f_0, \ldots, f_m$ are convex and the equality constraints are linear. We will in the following sections try to identify some estimation problems that can be cast as convex optimization problems.

3. NOTATION AND BACKGROUND

Maximum *a posteriori* (MAP) estimation [3] is about finding an estimator of a stochastic variable z that maximizes the conditional density $p(z|y)$, given the observation y ($y \in \mathbb{R}^{n_y}$ and $z \in \mathbb{R}^{n_z}$). Thus, the MAP problem is

$$\max_z \quad \log p(z|y) \tag{2}$$

In the sequel, the measurements vectors y_i from time 1 to time k will be denoted $y_{1:k}$, and similarly $z_{0:k}$ denotes all unknowns including the initial values. The operator $z_i^{(j)}$ extracts the j*th* element from the vector z_i.

The assumptions commonly used in the literature are that the elements in the z vectors are spatially and temporally independent ('white noise') and Gaussian distributed. We will insist on the independence assumption, but not on the assumption of Gaussian densities, giving us the following form of $\log p(z)$ (supressing the dependence on y)

$$\log p(z_{0:k}) = \log \prod_{i=0}^{k} p_{z_i}(z_i) = \sum_{i=0}^{k} \log p_{z_i}(z_i). \tag{3}$$

Depending on the distribution, the objective function in (1) can be explicitely written as in Table 1, see also [2].

4. CONVEX OPTIMIZATION ESTIMATION

In this section we will discuss the estimation problem in the presence of constraints. In Table 1 the objective functions are given for several log-concave [1] densities. Constraints

This work was supported by the Swedish Research Council

[1] A function function $f : \mathbb{R}^n \to \mathbb{R}$ is *log-concave* if $f(x) > 0$ for all x in the domain of f, and $\log f$ is a concave function [2].

PDF	Objective function	Extra constraints
Gaussian	$\sum_{i=0}^{k} \|z_i\|^2$	
Exponential	$\sum_{i=0}^{k} \sum_{j=1}^{n_z} z_i^{(j)} - 1$	$z \geq 0$
Laplacian	$\sum_{i=0}^{k} \sum_{j=1}^{n_z} \|z_i^{(j)}\|$	
Uniform	constant	$-\sqrt{3} \leq z \leq \sqrt{3}$

Table 1. Objective functions in (1) for different normalized (zero mean and unit variance) probability density functions.

arise in the derivation of some of these probability density functions (PDF), but constraints also arise from prior information of some kind, e.g., a model assumption. This will be discussed in the Section 6.

Assume we want to estimate (x, z), where z has a certain known distribution, and that x and z are related through the constraints

$$A \begin{bmatrix} x \\ z \end{bmatrix} = b, \tag{4}$$

If we now want to use (2) we are faced with the problem of finding the joint distribution of x and z, which can be quite tedious.

Problem 4.1 (Convex optimization estimation) *Assume that $p(z)$ is a known log-concave PDF. The MAP-estimate for x, z, where x and z are related via* (4) *is given by*

$$\begin{aligned} \max_{x,z} \quad & \log p_z(z) \\ \text{s.t.} \quad & A \begin{bmatrix} x \\ z \end{bmatrix} = b \end{aligned} \tag{5}$$

Remark: Any linear equalities and convex inequalities may be added to this formulation, and standard software applies.

This approach to estimation is presented in [2]. The standard estimation problem is to interpret x as the parameters conditioned on the measurements $x|y$, and then z is just a nuisance parameter. The standard approach, not often written explicitely, is to marginalize the nuisance parameters to get $p(x|y) = \int p(x|y,z)p(z|y)dz$ where the constraints are used explicitely. This works fine in a range of applications, and the solution most often has a quite simple form. In the general case, we can formulate the problem below.

5. LINEAR REGRESSION EXAMPLE

As an example of estimation, consider a linear regression problem in matrix form

$$Y = \Phi^T \theta + E. \tag{6}$$

Interpret $E \leftrightarrow z$ as a Gaussian nuisance parameter with variance σ^2, the regression parameter $\theta \leftrightarrow x$ as the parameter and $Y, \Phi \leftrightarrow y$ as the observations. The well-known result from marginalization is that

$$\theta \in \mathrm{N}((\Phi\Phi^T)^{-1}\Phi Y, \sigma^2(\Phi\Phi^T)^{-1}). \tag{7}$$

Alternatively, we can pose the problem as

$$\begin{aligned} \max_{x,\,z} \quad & \log p_E(E) \\ \text{s.t.} \quad & [\Phi^T,\ \mathbf{1}] \begin{bmatrix} \theta \\ E \end{bmatrix} = Y \end{aligned} \tag{8}$$

If this regression model happens to be an ARX model of a transfer function

$$G(e^{i\omega}) = \frac{\sum_l b^{(l)} e^{-i\omega l}}{1 + \sum_l a^{(l)} e^{-i\omega l}}, \tag{9}$$

in system identification, we use $\theta = (a^T, b^T)^T$. Now, we can simply add constraints such as bounded DC gain $L \leq G(0) \leq U$, or more generally, any lower and upper bound on the transfer function

$$L(\omega) \leq \frac{\sum_l b^{(l)} e^{-i\omega l}}{1 + \sum_l a^{(l)} e^{-i\omega l}} \leq U(\omega), \tag{10}$$

which is easily rewritten in the standard form. Similarly, any other interval for any other frequency of the transfer function can be bounded.

6. CONVEX OPTIMIZATION FILTERING

In the previous section we talked about constraints in general. We will in this section discuss a special type of constraints, namely the ones that appear in describing the dynamic behaviour of a model. In order to obtain convex problems we will use linear models of the dynamics. The following model

$$Ex_{k+1} = Ax_k + Bw_k + Ke_k, \tag{11a}$$

$$y_k = Cx_k + De_k, \tag{11b}$$

together with a density for the initial state, p_{x_0}, and p_e, p_w will constitute our model. With $E = I$, $K = 0$ we have the standard state space model, and with $E = I$, $B = 0$, $D = I$ we have the so called innovation form. If the E-matrix in (11a) is invertible we can rewrite the equation in a state space model. Otherwise we have what is commonly refered to as a descriptor model. [4].

To put state filtering in the general estimation form as in Problem 4.1, let

$$z = \left[x_0^T,\ w_{0:k-1}^T,\ e_{0:k}^T\right]^T, \tag{12}$$

and interpret x as $x_{1:k}|y_{1:k}$. To rewrite the conditional density more explicitly, use the independence assumption and (3), which gives

$$\begin{aligned} \log p(x_0, w_{0:k-1}, e_{0:k}) &= \log p_{x_0}(x_0) \\ &+ \sum_{i=0}^{k-1} \log p_{w_i}(w_i) + \sum_{i=0}^{k} \log p_{e_i}(e_i). \end{aligned} \tag{13}$$

Using Bayes' rule, $p(z|y) = p(y|z)p(z)/p(y)$ and the fact that

$$p(x_k) = p_{x_0}(x_0) \prod_{i=0}^{k-1} p_{w_i}(w_i), \tag{14}$$

$$p(y_k|x_k) = \prod_{i=0}^{k} p_{e_i}(e_i), \tag{15}$$

we obtain the following goal function

$$p(x_0, w_{0:k-1}, e_{0:k}) = \prod_{i=0}^{k} p_{e_i}(e_i) p_{x_0}(x_0) \prod_{i=0}^{k-1} p_{w_i}(w_i).$$

Conditioned on z in (12), the states in (11a) are uniquely defined by a deterministic mapping $x = f(z)$, which implies that $p(x|z) = f(z)$ contains nothing stochastic. That is, the MAP estimate of x and z are simply related by $\hat{x}^{MAP} = f(\hat{z}^{MAP})$. Similarly, the joint MAP estimate x, z in the convex optimization formulation is given by maximizing $p(z)$, since $p(z, x) = p(x|z)p(z) = f(z)p(z)$ by Bayes' rule. Hence we have now justified the following general convex estimation problem.

Problem 6.1 (Convex optimization filtering) *Assume that the densities p_{x_0},p_{w_i}, and p_{e_i} are log-concave. In the presence of constraints in terms of a dynamic model* (11a) – (11b) *the MAP-estimate is the solution $\hat{x}_k = x_k$ to the following problem*

$$\begin{aligned} \max_{x_{1:k}, z} \quad & \log p_{x_0}(x_0) + \sum_{i=0}^{k-1} \log p_{w_i}(w_i) + \sum_{i=0}^{k} \log p_{e_i}(e_i) \\ \text{s.t.} \quad & Ex_{k+1} = Ax_k + Bw_k + Ke_k \\ & y_k = Cx_k + De_k \end{aligned}$$

Remark: Any linear equalities and convex inequalities may be added to this formulation, and standard software applies.

As is evident from Problem 6.1 we see that we are free to use different densities for the different disturbances p_{x_0}, p_{w_i}, and p_{v_i}. It is here also worth noting that the recursive solution to Problem 6.1 under the assumptions of Gaussian densities and a non-singular E-matrix is the celebrated Kalman filter. This has been known for a long time, see e.g., [5], and [6] for nice historical accounts of this fact, and for a proof see e.g., [7]. It is also worthwhile noting that Problem 6.1 under the assumption of Gaussian disturbances is a weighted least-squares problem. To see this combine 6.1 and the Gaussian case in Table 1, where the weights are the inverse of the covariance matrices. This provides a deterministic interpretation of the problem that the Kalman filter solves. For more on the similarities and differences between deterministic and stochastic filtering see e.g., [8]. We also see that if we solve Problem 6.1 we will not only obtain the filtered estimate $\hat{x}_{k|k}$, but also all the smoothed estimates, $\hat{x}_{i|k}, i = 0, \ldots, k-1$.

So why should we solve the estimation problem via 6.1, which demands more computations, instead of via the Kalman filter? There are two reasons. The first reason is that we can handle all log-concave density functions, not just the Gaussian. The second reason is that we can add any prior information, in convex form, to problem 6.1. That is we can add linear equality constraints and convex inequality constraints, and still find the optimal estimate. We will see an illustration of this in the example in the subsequent section.

7. HMM EXAMPLE

There are mainly two filtering problems, where there exist finite-dimensional recursive optimal filters, and in particular a finite-dimensional MAP estimator. One is, as already mentioned, linear state space models with Gaussian noise. Here the Kalman filter is optimal in ML, MAP and minimum variance senses. For non-Gaussian noises, the Kalman filter computes the linear state estimate with minimum variance, but it is no longer the MAP or ML estimator.

The other case is hidden Markov models (HMM). Interestingly, it has been pointed out [9] that the HMM can be written in a state space model. That is, the Kalman filter computes the best possible linear estimate of the Markov state. This fact makes it possible to compare conceptually different approaches on the same example!

A hidden Markov model is defined by a discrete variable $\xi \in (1, 2, \ldots, n)$ with a known transition probability matrix A, where $A^{(i,j)} = P(\xi_k = i|\xi_{k-1} = j)$, that is, given that $\xi = j$ at time $k-1$, the probability that $\xi = i$ at time k is $A^{(i,j)}$. We will assume an observation process $\nu \in (1, 2, \ldots, m)$, where $P(\nu = i|\xi = j) = C^{(i,j)}$. The filter for computing the *a posteriori* probabilities can be expressed as the recursion

$$\pi_k^{(i)} = p(\xi_k = i) \tag{16a}$$

$$= \frac{\sum_{j=1}^{n} \sum \pi_{k-1}^{(j)} A^{(i,j)} C^{(\nu_k, j)}}{\sum_{j=1}^{n} \sum \pi_{k-1}^{(j)} C^{(\nu_k, j)}}. \tag{16b}$$

The MAP estimate is $\hat{\xi}_k = \arg\max_i \pi_k^{(i)}$. Now, the HMM can be written as the state space model

$$x_{k+1} = Ax_k + w_k, \tag{17a}$$

$$y_k = Cx_k + e_k, \tag{17b}$$

where $x_k^{(i)} = p(\xi_k = i)$ and $y_k^{(i)} = p(\nu_k = i)$. This is the state-space form (11a)–(11b) ($B = D = E = I, K = 0$) where the disturbances are zero-mean white noises, and the stationary covariance matrices can be shown to be

$$Q = \operatorname{Cov} w_k = \operatorname{diag}(\pi) - A \operatorname{diag}(\pi) A^T, \tag{18a}$$

$$R = \operatorname{Cov} e_k = \operatorname{diag}(C\pi) - C \operatorname{diag}(\pi) C^T, \tag{18b}$$

where π is the stationary solution to (in vector form)

$$\pi = \lim_{k \to \infty} A^k \pi_0, \text{where } \pi_0 > 0. \tag{19}$$

Since the states x we are estimating in a HMM are probabilities we have the following prior information on the states

$$\sum_{i=1}^{2} x^{(i)} = 1, \quad \text{and} \quad x^{(i)} \geq 0, \ i = 1, 2. \tag{20}$$

In the standard Kalman filter it is impossible to incorporate this prior information about the states, however in Problem 6.1 it is straightforward. We will now examine four different filters using an increasing amount of prior information (In 1-3 we have approximated w_t and e_t in (17) as Gaussian with zero mean and covariances (18)):

1. The Kalman filter.
2. The convex optimization filter with constraint $\sum_i x_k^{(i)} = 1$. This case can alternatively be computed by the Kalman filter using $P_0 = p_0 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and any $\sum_i x_0^{(i)} = 1$, or by using the fictitious measurement $y_0 = (1, 1, \ldots, 1)x_0 = 1$ with zero measurement noise. Note, however, that the Ricatti equation will be singular here, which may imply certain numerical difficulties. A more theoretically sound alternative is given in [9].
3. The convex optimization filter with constraint (20).
4. The optimal filter (16).

The numerical example is taken from [9], where

$$A = C = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \tag{21}$$

In Table 2, the root mean square error (RMSE) is given for these four cases and in Fig. 1 the states are shown. From

1. Kalman filter	0.585
2. 6.1 with $x_1 + x_2 = 1$	0.573
3. 6.1 with $x_1 + x_2 = 1$ and $x \geq 0$	0.566
4. Optimal filter	0.403

Table 2. RMSE values for the different filters.

this table it is obvious that we can obtain better estimates by using more information in this case. Of course, the convex optimization filters cannot compare to the performance of the optimal filter. However, the point is to show the flexibility of the approach, and the problem of consideration can be generalized with more contraints or a more complicated measurement relation, such that the optimal filter does no longer exist.

8. CONCLUSIONS

We have formulated the state estimation problem in a convex optimization framework. In this way, well-known numerical efficient algorithms can be used to compute the MAP

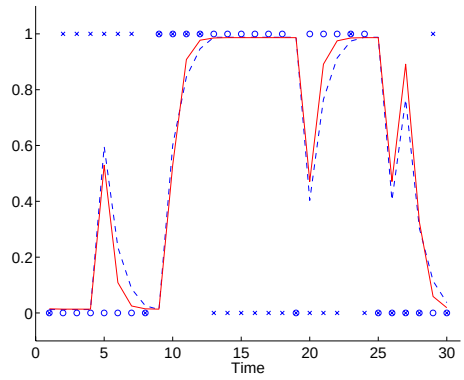

Fig. 1. The true state is marked by o, and the measured states by x. The dashed/solid line is the estimate from filter 3, respective 4.

estimate of the state vector, without any problems with local minima. Compared to the Kalman filter, the advantage is that any log-concave noise distribution can be used and any linear equality or convex inequality state contraint may be included, while the main drawback is that no recursive convex optimization algorithm is yet available, which makes the approach computer intensive.

9. REFERENCES

[1] D.G. Robertson and J.H. Lee, "On the use of constraints in least squares estimation and control," *Automatica*, vol. 38, pp. 1113–1123, 2002.

[2] L. Vandenberghe and S. Boyd, "Convex optimization," To be published, December 2001.

[3] A.H. Jazwinski, *Stochastic processes and filtering theory*, Mathematics in science and engineering. Academic press, New York, 1970.

[4] D.G. Luenberger, "Dynamic equations in descriptor form," *IEEE Transactions on Automatic Control*, vol. AC-22, no. 3, jun 1977.

[5] H.W. Sorenson, "Least-squares estimation: from gauss to kalman," *IEEE Spectrum*, vol. 7, pp. 63–68, July 1970.

[6] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Transactions on Information Theory*, vol. IT-20, no. 2, pp. 146–181, March 1974.

[7] C.V. Rao, *Moving Horizon Strategies for the Constrained Monitoring and Control of Nonlinear Discrete-Time Systems*, Ph.D. thesis, University of Wisconsin Madison, 2000.

[8] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear Estimation*, Information and System Sciences Series. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[9] S. Andersson, *Hidden Markov Models - Traffic Modeling and Subspace Methods*, Ph.D. thesis, Centrum for Mathematical Sciences, Mathematical Statistics, Lund University, Lund, Sweden, 2002.