Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density

Simo Särkkä
Lab. of Computational Engineering
Helsinki University of Technology
Espoo, Finland
E-mail: simo.sarkka@hut.fi

Aki Vehtari
Lab. of Computational Engineering
Helsinki University of Technology
Espoo, Finland
E-mail: aki.vehtari@hut.fi

Jouko Lampinen
Lab. of Computational Engineering
Helsinki University of Technology
Espoo, Finland
E-mail: jouko.lampinen@hut.fi

***Abstract*— This article presents a classical type of solution to the time series prediction competition, the CATS benchmark, which is organized as a special session of the IJCNN 2004 conference. The solution is based on sequential application of the Kalman smoother, which is a classical statistical tool for estimation and prediction of time series. The Kalman smoother belongs to the class of linear methods, because the underlying filtering model is linear and the distributions are assumed Gaussian. Because the time series model of the Kalman smoother assumes that the densities of noise terms are known, these are determined by cross-validation.**

I. Introduction

A. Overview of the Approach

The purpose of this article is to present a solution to the CATS time series prediction competition, which is organized as a special session of the IJCNN 2004 conference. The approach is selected to be a simple one, such that all the models are linear Gaussian and methods are based on application of classical statistical linear filtering theory.

The proposed method uses linear models for long and short term behavior of the signal. The computation is based on the Kalman smoother, where the noise densities are estimated by cross-validation. In time series prediction the Kalman smoother is applied three times in different stages of the method.

B. Optimal Linear Filtering and Smoothing

The success of *optimal linear filtering* is mostly due to the journal paper of Kalman [1], which describes a recursive solution to the discrete linear filtering problem. Although the original derivation of *Kalman filter* was based on least squares approach, the same equations can be derived from pure probabilistic Bayesian analysis. The Bayesian analysis of Kalman filtering is well covered in the classic book of Jazwinski [2] and more recently in the book of Bar-Shalom et al. [3].

Kalman filtering, mostly because of its least squares interpretation, has been widely used in stochastic optimal control. A practical reason to this is that the inventor of Kalman filter, Rudolph E. Kalman, has also made several contributions to the theory of *linear quadratic Gaussian* (LQG) regulators, which are fundamental tools of stochastic optimal control.

As discussed in the book of West and Harrison [4], in the sixties, Kalman filter type recursive estimators were also used in Bayesian community and it is not clear if theory of Kalman filtering or theory of *dynamic linear models* (DLM) was the first. Although these theories were originally derived from slightly different starting points, they are equivalent. This article approaches the Bayesian filtering problem in Kalman filtering point of view, because of its useful connection to the theory and history of stochastic optimal control.

C. Kalman Filter

The Kalman filter (see, e.g. [2], [3]), which originally appeared in [1], considers a discrete filtering model, where the dynamic and measurements models are linear Gaussian

$$\begin{aligned} \mathbf{x}_k &= \mathbf{A}_{k-1}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \\ \mathbf{y}_k &= \mathbf{H}_k\mathbf{x}_k + \mathbf{r}_k, \end{aligned} \tag{1}$$

where $\mathbf{q}_{k-1} \sim N(0, \mathbf{Q}_{k-1})$ and $\mathbf{r}_k \sim N(0, \mathbf{R}_k)$. If the prior distribution is Gaussian, $\mathbf{x}_0 \sim N(\mathbf{m}_0, \mathbf{P}_0)$, then the optimal filtering equations can be evaluated in closed form and the resulting distributions are Gaussian

$$\begin{aligned} p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) &= N(\mathbf{x}_k \mid \mathbf{m}_k^-, \mathbf{P}_k^-) \\ p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) &= N(\mathbf{x}_k \mid \mathbf{m}_k, \mathbf{P}_k) \\ p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}) &= N(\mathbf{y}_k \mid \mathbf{H}_k\mathbf{m}_k^-, \mathbf{S}_k). \end{aligned}$$

The parameters of the distributions above can be calculated by Kalman filter *prediction* and *update steps*:

- *Prediction step* is

$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{A}_{k-1}\mathbf{m}_{k-1} \\ \mathbf{P}_k^- &= \mathbf{A}_{k-1}\mathbf{P}_{k-1}\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1}. \end{aligned}$$

- *Update step* is

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\mathbf{m}_k^- \\ \mathbf{S}_k &= \mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k \\ \mathbf{K}_k &= \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{S}_k^{-1} \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\mathbf{v}_k \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T. \end{aligned}$$

D. Kalman Smoother

The Kalman smoother (see, e.g., [2], [3]) calculates recursively the state posterior distributions

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:T}) = N(\mathbf{x}_k \mid \mathbf{m}_k^s, \mathbf{P}_k^s),$$

for the linear filtering model (1). The difference to the posterior distributions calculated by the *Kalman filter* is that the smoothed distributions are conditioned to all the measurement data $\mathbf{y}_{1:T}$, while the filtered distributions are conditional only to the measurements obtained before and at the time step k, that is, to measurements $\mathbf{y}_{1:k}$.

The smoothed distributions can be calculated from the Kalman filter results by recursions

$$\begin{aligned}
\mathbf{P}_{k+1}^- &= \mathbf{A}_k \mathbf{P}_k \mathbf{A}_k^T + \mathbf{Q}_k \\
\mathbf{C}_k &= \mathbf{P}_k \mathbf{A}_k^T [\mathbf{P}_{k+1}^-]^{-1} \\
\mathbf{m}_k^s &= \mathbf{m}_k + \mathbf{C}_k [\mathbf{m}_{k+1}^s - \mathbf{A}_k \mathbf{m}_k] \\
\mathbf{P}_k^s &= \mathbf{P}_k + \mathbf{C}_k [\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-] \mathbf{C}_k^T,
\end{aligned}$$

starting from last time step T, with $\mathbf{m}_T^s = \mathbf{m}_T$ and $\mathbf{P}_T^s = \mathbf{P}_T$.

II. Description of the Model

A. Long Term Model

For long term prediction, a linear dynamic model is likely to be a good approximate model because if we ignore the short term periodicity of the data, the data could be well generated by a locally linear Gaussian process with Gaussian measurement noise. The data seems to consist of lines with suddenly changing derivatives. Thus, it would be reasonable to model the derivative as Brownian noise process, which leads to white noise model for second derivative.

There is no evidence of benefit for using higher derivatives, because the curve consists of set of lines rather than set of parabolas or other higher order curves. The dynamic model is formulated as a continuous time model, and then discretized to allow varying sampling rate, that is, the prediction over the missing measurements.

The selected dynamic linear model for long term prediction is

$$\begin{pmatrix} x_k \\ \dot{x}_k \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{pmatrix} + \begin{pmatrix} q_{1,k-1}^x \\ q_{2,k-1}^x \end{pmatrix},$$

where the process noise, $\mathbf{q}_k^x = (q_{1,k-1}^x \; q_{2,k-1}^x)^T$, has zero mean and covariance

$$\mathbf{Q}_{k-1} = \begin{pmatrix} \Delta t^3/3 & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t \end{pmatrix} q^x,$$

where Δt is the time period between samples and q^x defines the strength (spectral density) of the process noise. Suitable measurement model is

$$y_k = x_k + r_k^x, \quad r_k^x \sim N(0, \sigma_x^2).$$

A quick testing of the long term model produces a smooth curve as shown in Fig. 1. It can be seen that the locally linear dynamic model may be a bit too simple, because there still seems to be noticeable periodicity in the residual signal. This periodicity can be best seen from the residual autocorrelation in Fig. 2.

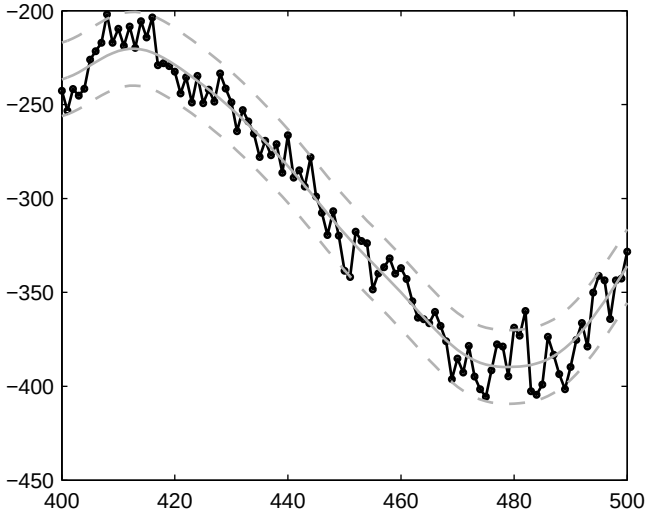

Fig. 1. Data 400–500 (black) and the result of prediction with the long term model (gray).

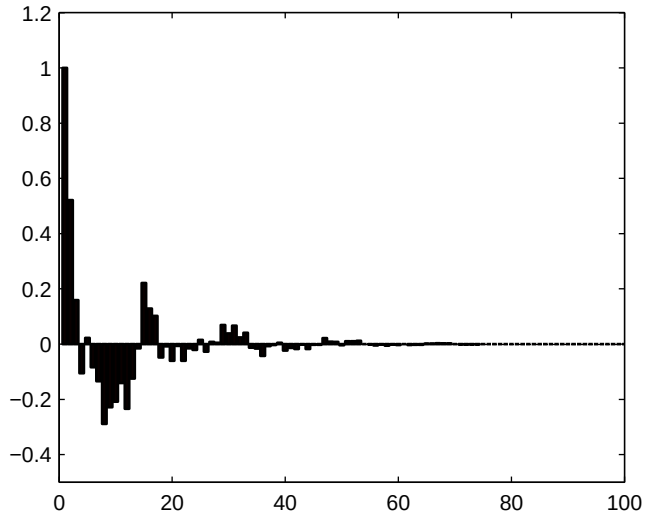

Fig. 2. Autocorrelation in residual of long term prediction model.

B. Short Term Model

The short term periodicity of the residual time series $\{e_k : k = 1, ..., N\}$ can be modeled by a linear prediction or AR-model (see, e.g., [5]), where as an extension to typical models, we allow the weight to vary according to a Gaussian random walk model

$$\begin{aligned}
\mathbf{w}_k &= \mathbf{w}_{k-1} + \mathbf{v}_k^{\mathsf{ar}} \\
e_k &= \sum_i w_{i,k} e_{k-i} + r_k^{\mathsf{ar}}.
\end{aligned} \tag{2}$$

The process noise $\mathbf{v}_k^{\mathsf{ar}}$ has zero mean and covariance $\mathbf{Q} = q^{\mathsf{ar}}\mathbf{I}$. The weight vector $\mathbf{w}_k$ is to be estimated from the known part of residual time series. The measurement noise has Gaussian distribution $r_k^{\mathsf{ar}} \sim N(0, \sigma_{\mathsf{ar}}^2)$.

After the AR-model has been estimated from the residual time series data, the final estimation solution is obtained from

$$\begin{aligned} d_k &= \sum_i w_{i,k} d_{k-i} + v_k^{\mathsf{p}} \\ e_k &= d_k + r_k^p, \quad r_k^p \sim N(0, \sigma_p^2), \end{aligned} \tag{3}$$

where the process noise v_k^{p} has variance q^{p}. The final signal estimate is then given as $\hat{y}_k = \hat{x}_k + \hat{d}_k$, where $\hat{x}_k$ is the estimate produced by applying Kalman smoother to the long term model, and $\hat{d}_k$ is produced by the short term model.

In practice, only the distributions of weight vectors $\mathbf{w}_k$ are known, not their actual values, and in order to use the model (3) we would have to integrate over these distribution on every step. In this article we have used a common approach, where this integration is approximated by using the most likely estimate of weights vectors and this value is regarded as being known in advance. In classical statistical signal processing this estimate is calculated by linear least squares (see, e.g., [5]). Because our weight vector is allowed to vary in time, the corresponding estimate in this case is produced by applying Kalman smoother to model (2).

In this article we chose to use a second order AR-model such that the weight vector was two dimensional,

$$\mathbf{w}_k = \begin{pmatrix} \mathbf{w}_{1,k} \\ \mathbf{w}_{2,k} \end{pmatrix}. \tag{4}$$

C. The Prediction Method

The long term prediction is then done in two steps:

1) Run *Kalman filter* over the data sequence and store the estimated means and covariances. Predict the missing measurement such that the filtering result contains estimates also for the missing steps.
2) Run *Kalman smoother* over the Kalman filter estimation result, which results in smoothed (MAP) estimate of the time series including the missing parts.

The short term prediction consists of four steps:

1) Run *Kalman filter* over the residual sequence with model (2) in order to produce the filtering estimate of AR weight vectors. Predict the weights over the missing parts.
2) Run *Kalman smoother* over the Kalman filter estimation result above, which results in smoothed (MAP) estimate of the weight time series including the missing parts.
3) Run *Kalman filter* over the residual sequence with model (3) in order to produce filtering estimate of the short term periodicity. The periodicity is also predicted over the missing parts.
4) Run *Kalman smoother* over the Kalman filter estimation result above, which results in smoothed (MAP) estimate of the periodicity time series including the missing parts.

Due to Gaussian random walk model of weights the short term model has potentially a large effective number of parameters. Simple error minimization procedure with respect to to noise parameters (e.g., Maximum Likelihood) would lead to a badly over-fitted estimation solution. By application of cross-validation we can maximize the predictive performance and avoid over-fitting.

III. Results

A. Selection of Measurement Noises

Long term measurement noise strength can be approximated by looking at a short time period of the curve. If we assume that we would approximate it by a dynamic linear model, we can approximate the standard deviation of the model's measurement noise by looking at the strengths of residuals. The selected variance of noise was $\sigma_x^2 = 10^2$, which quite well fits to the observed residual as can be seen in the Fig. 1.

The choice of measurement noises both in long term and short term models can be done, for example, by visual inspection, because the exact choice of noise strength is not crucial. In fact, the choice does not matter at all when the cost function of the CATS competition is considered, because in this case the selection is dependent on the selection of process noise strength in all the models. The process noise strength is selected based on cross-validation, which implicitly corrects also the choice of measurement noise strength. By visual inspection a suitable measurement noise for the AR-estimation model (2) was $\sigma_{\mathsf{ar}}^2 = 1^2$.

Because we are only interested in the missing parts of data in prediction with model (3), the best way to do this is to follow the measurements exactly whenever there are measurements and use AR-model for prediction only when there are no measurements. This happens when the measurement noise level is set to as low as possible and the process noise is set to a moderate value. Our choice for noise level in model (3) was $\sigma_{\mathsf{p}}^2 = 10^{-9}$.

B. Cross-Validation of Process Noises

Process noise parameters q^x and q^{ar} were selected using a decision theoretic approach by minimizing the expected cost where cost function is the target error criterion. Expected cost can easily be computed by cross-validation, which approximates the formal Bayes procedure of computing the expected costs (see, e.g., [6]). Based on cross-validation, the best process noises were

$$\begin{aligned} q^x &= 0.14 \\ q^{\mathsf{ar}} &= 0.0005. \end{aligned} \tag{5}$$

As discussed in the previous Section, the only requirement for selection of process noise q^{p} is that it should be high enough. Because the measurement noise was selected to be very low, our choice was $q^{\mathsf{p}} = 1$.

C. Prediction Results

Fig. 3 shows the estimated AR-coefficients for each time instance. It can be seen that the weights vary a bit over time, but the periodic short term process seems to be quite stationary.

Figs. 4, 5, 6, 7 and 8 show the results of predicting over the missing intervals. It can be seen that on the missing intervals the short term model differs from long term model only near

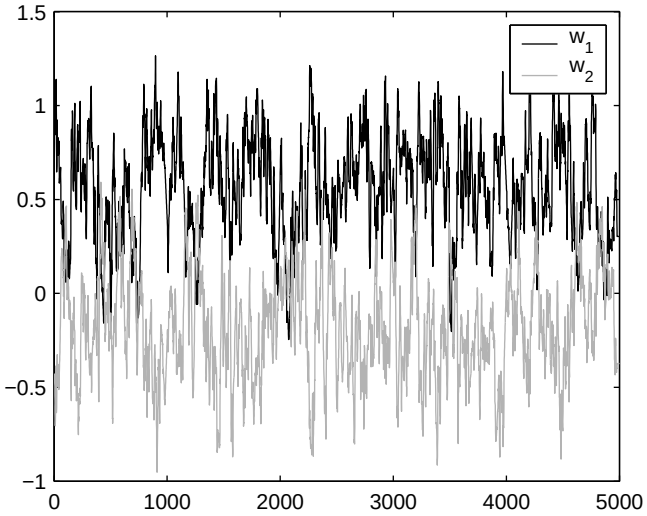

Fig. 3. Estimated filter coefficients for the AR-model.

the measurements and the combined estimate is closest to the long term prediction in the middle of prediction period. The result is intuitively sensible, because when going away from the measurement we have less information on the phase of local periodicity and it is best to just guess the mean given by the long term model.

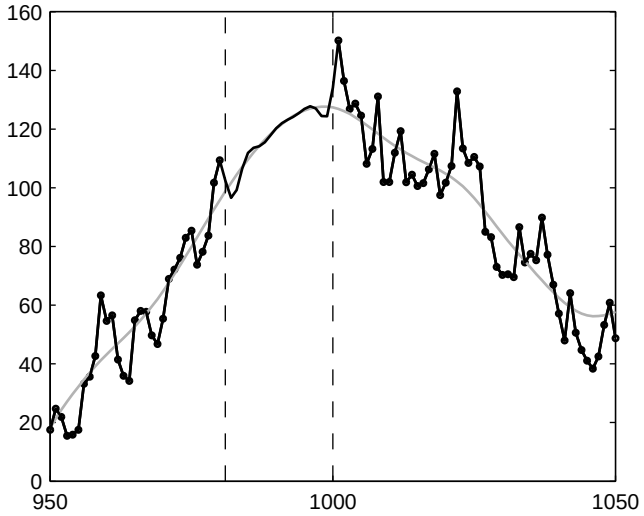

Fig. 4. Prediction over missing data at 981 – 1000. The gray line is the long term prediction result and black line is the combined long and short term prediction result.

IV. Conclusions

A. Summary of the Results

In this article we applied the classical Kalman smoother method for estimating the long term and short term statistical models for the CATS benchmark time series. The results indicate that the long term prediction gives a very good overall

Fig. 5. Prediction over missing data at 1981 – 2000. The gray line is the long term prediction result and black line is the combined long and short term prediction result.

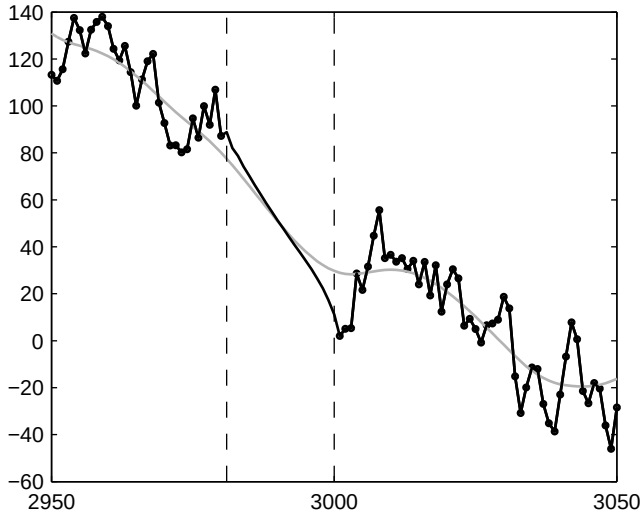

Fig. 6. Prediction over missing data at 2981 – 3000. The gray line is the long term prediction result and black line is the combined long and short term prediction result.

approximation of the signal and the short term prediction catches the local periodicity ignored by the long term model.

Although all the used models were linear (and dynamic) in nature they seem to model well this non-linear time series. It is likely that by using non-linear state space models (filtering models) the prediction results would be better, but it is very hard to judge what kind of model is really the best. This judgment naturally also depends on the criterion used.

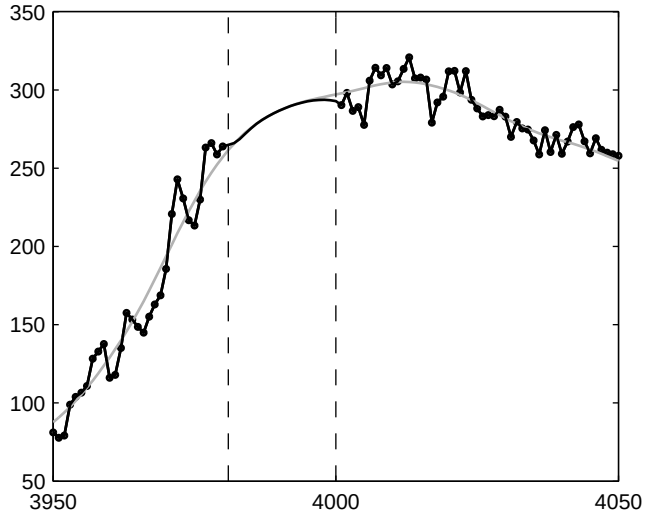

Fig. 7. Prediction over missing data at 3981 – 4000. The gray line is the long term prediction result and black line is the combined long and short term prediction result.

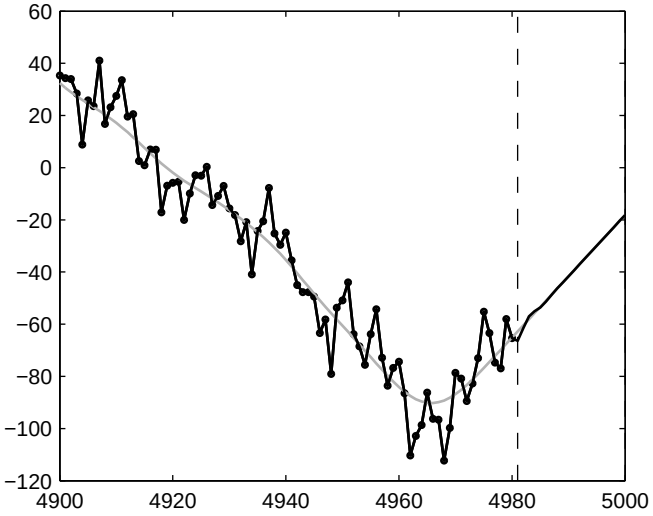

Fig. 8. Prediction over missing data at 4981 – 5000. The gray line is the long term prediction result and black line is the combined long and short term prediction result.

ACKNOWLEDGMENT

The authors would like to thank Ilkka Kalliomäki for his help on proofreading the article.

REFERENCES

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, pp. 34–45, March 1960.

[2] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[3] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. Wiley Interscience, 2001.

[4] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, 1997.

[5] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., 1996.

[6] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, 1994.