IDIAP RESEARCH REPORT

Efficient Kalman Smoothing for Harmonic State-Space Models

David Barber [a]

IDIAP–RR 05-87

[a] IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland

IDIAP Research Report 05-87

Efficient Kalman Smoothing for Harmonic State-Space Models

David Barber

Abstract. Harmonic probabilistic models are common in signal analysis. Framed as a linear-Gaussian state-space model, smoothed inference scales as $O(TH^2)$ where H is twice the number of frequencies in the model and T is the length of the time-series. Due to their central role in acoustic modelling, fast effective inference in this model is of some considerable interest. We present a form of 'rotation-corrected' low-rank approximation for the backward pass of the Rauch-Tung-Striebel smoother. This provides an effective approximation with computation complexity $O(TSH)$ where S is the rank of the approximation.

1 Introduction

Harmonic signal decompositions are one of the main tools in audio analysis and the harmonic plus noise model has recently been used in several applications [1, 2, 3]. In its simplest form we model a signal by a superposition of harmonic oscillators, this being essentially the Fourier Representation. The probabilistic interpretation of a Harmonic representation of a one-dimensional signal from time 1 to time T, $y_{1:T}$, is useful since a generative model enables one to build in known constraints about the signal generation process (see e.g. [2] for an application). Here we concentrate on the simplest form of these models, being essentially a bank of harmonic oscillators, and show how inference can be computed efficiently.

A useful state-space representation of a single harmonic oscillator is based on a two-dimensional latent linear dynamics $x_{t+1} = A(\theta)x_t$, where $A(\theta)$ is a Givens Rotation matrix:

$$A(\theta) = \rho \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

Then x_t describes a damped rotation in which at each timestep, the vector rotates anticlockwise by θ degrees, with a length reduction factor $0 < \rho < 1$. The projection of this two-dimensional vector onto the first dimension, $x_{1,t}$ then describes a one-dimensional harmonic oscillator, as shown in fig(1). The time-dependent energy related to this harmonic component is given by the length of the vector x_t. To describe a bank of such oscillators, rotating at different frequencies, we form the block-diagonal damped rotation matrix

$$A = \begin{pmatrix} A(\theta_1) & 0 & \cdots & 0 \\ 0 & A(\theta_2) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & A(\theta_n) \end{pmatrix}$$

To cope with the fact that a real signal will deviate from a perfect damped oscillator, we introduce additive Gaussian noise both in the state-space

$$x_{t+1} = Ax_t + w_t, \qquad w_t \sim \mathcal{N}(0, \Sigma^x) \tag{1}$$

and in the signal observation.

$$y_t = Bx_t + v_t, \qquad v_t \sim \mathcal{N}(0, \Sigma^y) \tag{2}$$

Here x_t is a $H \times H$ dimensional matrix, where H is equal to twice the number of frequencies in the harmonic representation. $B = [1, 0, 1, 0, \ldots, 1, 0]$ is a $1 \times H$ matrix (i.e. the transpose of a $H \times 1$ vector). This model accounts for both amplitude and phase noise – a sample from such a model is given in fig(2). An alternative probabilistic formulation of the above equations is

$$p(x_t|x_{t-1}) = \mathcal{N}(A_t x_{t-1}, \Sigma^x) \tag{3}$$

$$p(y_t|x_t) = \mathcal{N}(C_t x_t, \Sigma^y) \tag{4}$$

Figure 1: A damped oscillator in state space form. Left: At each time step, the state vector x rotates by θ and its length becomes shorter. Right: The actual waveform is a one dimensional projection from the two dimensional state vector. The stochastic model assumes that there are two independent additive noise components that corrupt the state vector x and the sample y, so the resulting waveform $y_{1:T}$ is a damped sinusoid with both phase and amplitude noise.

Figure 2: A sample from the Gaussian linear dynamical system defined by equation (1) and equation (2) from a bank of 100 oscillators evenly spread between 0 and 2000 Hz.

which define a joint Gaussian probability distribution

$$p(x_{1:T}, y_{1:T}) = \prod_{t=1}^{T} p(y_t|x_t)p(x_t|x_{t-1}) \tag{5}$$

where, by convention, $p(x_1|x_0)$ is a Gaussian distribution with mean 0 and covariance Σ_1^x. The above model is then a constrained form of a Kalman Filter[4]. Given a set of observations $y_{1:T}$, the two main interests are in calculating the *filtered* posterior inference $p(x_t|y_{1:t})$ and the *smoothed* posterior inference $p(x_t|y_{1:T})$. We shall see that filtering is computationally efficient, but smoothing is an order H more computationally demanding. The aim of this paper is to introduce an effective smoothing approximation for this important class of models.

2 Kalman Filtering

The smoothed posterior $p(x_t|y_{1:t})$ is a Gaussian, whose mean and covariance we denote by f_t and F_t respectively. To compute f_t and F_t, we may use the well-known Kalman Filter recursions[4] given in Algorithm 1. Here P_t and F_t are $H \times H$ symmetric matrices and G_t is a $H \times 1$ vector. The length of signal T that we wish to perform filtering and smoothing may be of the order of 10^4, so that storing these matrices is prohibitive. In practice, the matrices P_t, G_t, F_t converge quickly to a stationary value and we therefore adopt the usual approach of replacing these quantities by their converged values, as given in Algorithm 2[4].

Algorithm 1 Kalman Filter

procedure KalmanFilter
  $F_0 \leftarrow 0,\ f_0 \leftarrow 0$
  for $t \leftarrow 1, T$ **do**
    $P_t \leftarrow A_t F_{t-1} A_t^T + \Sigma^x$
    $G_t \leftarrow P_t B^T \left(B P_t B^T + \Sigma^y\right)^{-1}$
    $F_t \leftarrow (I - G_t B)\, P_t$
    $f_t \leftarrow A f_{t-1} - G_t\, (B A f_{t-1} - y_t)$
  end for
end procedure

Algorithm 2 Approximate Kalman Filter

procedure ApproxKalmanFilter
  $F \leftarrow 0$, $f \leftarrow 0$
  repeat $P \leftarrow AFA^T + \Sigma^x$
  until P converges
  $G \leftarrow PB^T \left(BPB^T + \Sigma^y\right)^{-1}$
  $F \leftarrow (I - GB)\, P$
  for $t \leftarrow 1, T$ **do**
    $f_t \leftarrow Af_{t-1} - G\,(BAf_{t-1} - y_t)$
  end for
end procedure

Algorithm 3 Kalman Smoothing : Rauch Tung Striebel

procedure KalmanSmoother
  $R_T \leftarrow F_T$, $r_T \leftarrow f_T$
  for $t \leftarrow T-1, 1$ **do**
    $X \leftarrow F_t A^T \left(AF_t A^T + \Sigma^x\right)^{-1}$
    $U = I - XA$
    $R_t \leftarrow XR_{t+1}X^T + UF_t$
    $r_t \leftarrow Xr_{t+1} + Uf_t$
  end for
end procedure

Since P, F and G may be computed offline in a one-off computation, the complexity of Algorithm 2 for filtering a signal $y_{1:T}$ is determined by the recursion $f_t \leftarrow Af_{t-1} - G\,(BAf_{t-1} - y_t)$. Since BA is a (transposed) vector which may be precomputed, the scalar BAf_{t-1} takes order $O(H)$ computations. The term Af_{t-1} would ordinarily take $O(H^2)$ operations. However, since A is block diagonal (consisting of 2×2 rotation matrices on the diagonals), this also takes $O(H)$ operations. Hence, the complexity of computing $f_{1:T}$ takes only order $O(TH)$ operations. That is, the complexity of filtering (given the converged approximate values for P) is linear in the number of harmonics desired and the length of the time series – an agreeable complexity.

3 Kalman Smoothing

Here we want to compute $p(x_t|y_{1:T})$ which is a Gaussian with mean r_t and covariance R_t. The standard approach to smoothing is to use the Rauch-Tung-Striebel smoother[4], as presented in Algorithm 3, which makes use of the Kalman Filter results. As in the Filtering recursions, the posterior covariance rapidly converges to a constant value, and we may also replace the time-dependent forward covariances by their converged estimates F. Since R_t does not depend on the observations, this may also be pre-computed. Hence, our main concern is with the following equation

$$r_t \leftarrow Xr_{t+1} + f_t - XAf_t \tag{6}$$

Using the converged values, X is given by

$$X = FA^T \left(AFA^T + \Sigma^x\right)^{-1}$$

which is time-independent. However, unlike in the Filter recursions, we cannot simply write X exactly as the outer-product of two vectors and, unfortunately, this means that the computation of Xr_{t+1} is order $O(H^2)$. This is unacceptable since H will typically be of the order of several hundred to a thousand. Similarly, the term XAf_t is problematic and also has an exact complexity of order

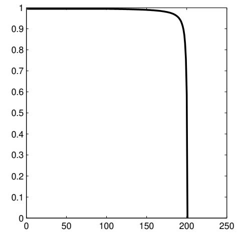

Figure 3: The singular values for a the matrix X formed from a bank of 100 oscillators evenly spread from 0 to 2000 Hz. The coefficient ρ is set to 0.999. The state covariances were set to $\Sigma^x = 10^{-3} I_H$ and the observation variance was set to $\Sigma^y = 10^{-6}$.

$O(H^2)$. An obvious strategy would be to replace X by a low-rank approximation. However, we may empirically observe that no-such low rank approximation of X exists, see for example fig(3), where typically nearly all the singular values from a Singular Value Decomposition (SVD) are close to unity. Hence, a naive strategy of projecting X to a low-rank subspace will fail since nearly all the singular values will be required for an accurate representation of X.

3.1 A low rank 'rotation' compensated Approximation

First we write $\hat{r}_{t+1} = r_{t+1} - Af_t$, so that equation (6) can be written as

$$r_t = f_t + X\hat{r}_{t+1}$$

Computing $\hat{r}_{t+1}$ is $O(H)$ again thanks to the fact that A is block diagonal. We now concentrate on X. This is given by

$$X = FA^T \left(AFA^T + \Sigma^x\right)^{-1}$$

Using the converged Kalman Filter equations

$$P = AFA^T + \Sigma^x, \qquad F = P - GBP$$

we may write

$$X = PA^T P^{-1} - GBPA^T P^{-1} \tag{7}$$

Bearing in mind that if X was of low rank, then the computational complexity would be modest, our aim is to find an approximate suitable decomposition of X. The term $GBPA^TP^{-1}$ in equation (7) is unproblematic since this is indeed trivially of the form of the outerproduct of the vector G with the vector $P^{-T}AP^TB^T$. Hence, this term causes no difficulty. Unfortunately, the term PA^TP^{-1} does not

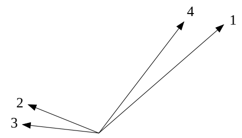

Figure 4: The effect of the operations PA^TP^{-1}. First the vector depicted 1 is transformed by P^{-1} into a new representation, vector 2. Then this vector is rotated by A^T to the vector 3, and then transformed back to the original basis, depicted by vector 4. If the rotation A^T is not too strong, then this will be roughly equivalent to rotating the original vector 1 by A^T.

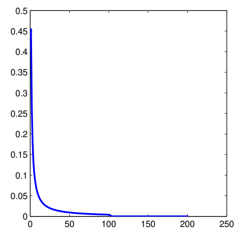

Figure 5: The singular values for a the matrix $K = PA^TP^{-1} - A^T$ formed from a bank of 100 oscillators even spread from 0 to 2000 Hz. The coefficient ρ is set to 0.999. The state covariances were set to $\Sigma^x = 10^{-3}I_H$ and the observation variance was set to $\Sigma^y = 10^{-6}$. Contrast this with fig(3).

possess a low rank approximation. The fundamental reason for this is that A is (proportional to) a rotation matrix – even if P were the identity, then A itself cannot have a low-rank approximation since it rotates *all* components. However, we may gain some insight into forming a useful approximation by the following reasoning. One may view the matrix PA^TP^{-1} as follows : P^{-1} first transforms into a new basis, we then perform a rotation in the basis (performed by A^T which corresponds to inverse rotation of A), and then transform back to the original basis. Hence, if the rotation A^T is relatively weak, then we may expect that the transformation PA^TP^{-1} has roughly the same effect as a rotation A^T in the original basis. This is depicted in fig(4). The idea, therefore, is that $K = PA^TP^{-1} - A^T$ may have a low rank approximation. The singular values of this rotation-corrected matrix are depicted in fig(5), where we see that indeed, a low rank approximation would be reasonable. A more sophisticated approximation would be to assume that $PA^TP^{-1} \approx \hat{P}A^T\hat{P}^{-1}$ where $\hat{P}$ is formed from a block diagonal approximation of P, although we have found that, in practice, the simpler approximation produces reasonable results.

A low rank approximation for K is then obtained by computing the Singular Value Decomposition $K = UDV^T$. Then by taking only the first S singular values, we obtain an approximation $K \approx \hat{U}\hat{V}$ where $\hat{U}$ is obtained from the first S columns of U, and $\hat{V}$ is obtained from the first S rows of DV^T. With this we then may write

$$X \approx \hat{U}\hat{V} + A^T - GBPA^TP^{-1}$$

$$r_t = f_t + \hat{U}\left(\hat{V}\hat{r}_{t+1}\right) + A^T\hat{r}_{t+1} - G\left(BPA^TP^{-1}\right)\hat{r}_{t+1}$$

The complexity of the final term $G\left(BPA^TP^{-1}\right)\hat{r}_{t+1}$ is $O(H)$, as is $A^T\hat{r}_{t+1}$ and f_t. The complexity of $\hat{U}\left(\hat{V}\hat{r}_{t+1}\right)$ is $O(SH)$ where S is the rank of the SVD approximation. Note that the SVD approximation can be computed offline, and is a one-time only computation.

3.2 Demonstration

In fig(6) we show a sample waveform for which we wish to find a harmonic representation using 200 frequencies evenly distributed between 0 and 2000 Hz. In fig(7) we plot the filtered spectrogram (1.3 seconds of computation using a Pentium III processor with 1 Gbyte of RAM), the exact smoothed posterior (90 seconds) and the rank 30 approximation of the smoothed posterior (7 seconds). The errors made by the approximation are given in fig(8), where we see that, crucially, in the regions where the posterior components are large, then the approximation is very accurate, with a mean absolute deviation of 0.002. Plotting the relative deviation is less meaningful since the spectrogram has mainly small values, although it is the relatively few larger values for which the approximation needs to be accurate.

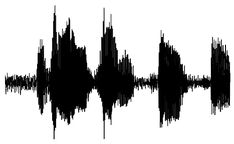

Figure 6: The waveform, corresponding to 2.5 seconds of speech, with 8000 samples per second.

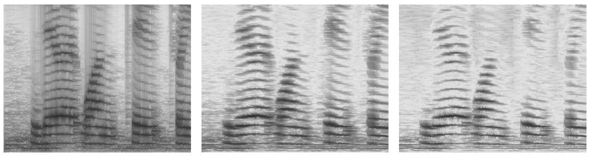

Figure 7: Spectrograms (log energy) of the waveform fig(6). Left: Filtered estimate $p(x_t|v_{1:t})$. Middle: exact smoothed posterior $p(x_t|v_{1:T})$. Right: approximation of the smoothed posterior using a rank 30 approximation.

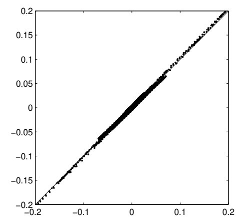

Figure 8: Each x, y point in this graph corresponds to a value $(r_{t,i}, \hat{r}_{t,i})$ where $r_{t,i}$ is the exact value of the i^{th} frequency component of the posterior vector $p(r_t|v_{1:T})$, and $\hat{r}_{t,i}$ is the corresponding rank $S = 30$ approximation. The mean absolute deviation of the approximation is 0.002.

4 Conclusion

Formulated as a linear dynamical system, filtered inference in Harmonic models can be carried out computationally efficiently using the Kalman Filter recursions, scaling as $O(TH)$ where H are the number of frequencies of the model and T is the length of the time series. However, the smoothed posterior cannot be exactly computed in a reasonable time, with the exact computation scaling as $O(TH^2)$. A naive low-rank approximation of the recursion also does not yield an effective approximation. However, our rotation-corrected low-rank approximation does provide an effective approximation to smoothing, with complexity $O(TSH)$ where S is the rank of the approximation. Typically, we have found that a rank of less than 20 is often sufficient for a reasonable approximation. This approximation technique can also be applied to more complex harmonic models (e.g. work deriving from [1]) and related probabilistic models in acoustics.

References

[1] R.J. McAulay and Quateri. T.F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on acoustics, speech and signal processing*, 34(4):744–754, 1986.

[2] A.T. Cemgil, H. J. Kappen, and D. Barber. A generative model for music transcription. *IEEE trans. Speech and Audio Processing, 2004*, 2004. submitted.

[3] A. T. Cemgil, B. Kappen, and D. Barber. Generative Model based Polyphonic Music Transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* , 2003.

[4] Y. Bar-Shalom and Xiao-Rong Li. *Estimation and Tracking : Principles, Techniques and Software.* Artech House, Norwood, MA, 1998.