INVERSION OF BLOCK MATRICES WITH BLOCK BANDED INVERSES: APPLICATION TO KALMAN-BUCY FILTERING

Amir Asif

Information Technology
Technical University of British Colombia
Surrey, BC V3R 7P8, Canada
Email: asif@techbc.ca

José M. F. Moura

Room 35-203, LIDS, EECS
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
Email: moura@mit.edu

ABSTRACT

We investigate the properties of block matrices with block banded inverses to derive efficient matrix inversion algorithms for such matrices. In particular, we derive the following: (1) a recursive algorithm to invert a full matrix whose inverse is structured as a block tridiagonal matrix; (2) a recursive algorithm to compute the inverse of a structured block tridiagonal matrix. These algorithms are exact. They reduce the computational complexity respectively by two and one orders of magnitude over the direct inversion of the associated matrices. We apply these algorithms to develop a computationally efficient approximate implementation of the Kalman-Bucy filter (KBf) that we refer to as the local KBf. The computational effort of the local KBf is reduced by a factor of I^2 over the exact KBf while exhibiting near-optimal performance.

1. INTRODUCTION

A major issue that precludes the application of sophisticated signal processing algorithms in computer vision or in the physical sciences in general is the large dimension of the state vector. In typical image-based applications, for example, the dimension of the state vector is on the order of the number of pixels in the image, typically 10^6 elements. The covariance matrices associated with an optimal filter applied to such problems have dimensions on the order of $10^6 \times 10^6$. The storage and the subsequent manipulation of such large matrices as required by the optimal filters is prohibitive, necessitating the use of a sub-optimal approach.

The paper investigates the use of structured banded matrices to derive near-optimal implementations of the Kalman-Bucy filter (KBf), [1]. We study block banded matrices and derive important properties that relate the constituent blocks of the block banded matrix to the block entries in its inverse, which is a full matrix. These inter-relationships between the block entries result in recursive algorithms for:
1. inverting a full matrix whose inverse is block banded;
2. computing the inverse of a block-banded matrix.
For tridiagonal block matrices, Algorithm 1 reduces the computational complexity by two orders of magnitude over the direct inversion of a full matrix whereas Algorithm 2 provides savings of one order of magnitude.

The block properties alluded to in our discussion above are used to derive a sub-optimal implementation of the KBf, referred to as the local KBf, that provides computational savings by a factor of I^2 over the direct KBf defined on an $I \times I$ image field. We approximate the inverse of the error covariance matrix, i.e., of the information matrix, by a block banded matrix. Banded approximations to information matrices correspond to modeling the error field as a reduced order Gauss Markov random field (GMrf). The estimate provided by the local KBf follows closely the estimate provided by the exact KBf.

The paper is organized as follows. In section 2, we derive three relevant results for matrices with block tri-diagonal banded inverses. In section 3, we present efficient algorithms to invert such matrices and band-limited block tridiagonal matrices. In section 4, we apply the band-limited approximation to develop the local KBf and illustrate the validity of the approximation through an experiment. In section 5 we conclude the paper.

2. BLOCK BANDED MATRICES

Consider a positive-definite, symmetric matrix P represented by its $(I \times I)$ constituent blocks $P = \{P_{ij}\}$, $1 \le i, j \le I$. The matrix P has dimensions of $(I^2 \times I^2)$. We assume the inverse of P, $A = P^{-1}$, is block banded. Such matrices occur in several applications like in Gauss Markov random fields (GMrf), where P is the covariance matrix and the inverse A is referred to as the information or potential matrix. For first order GMrf, the potential matrix A is tridiagonal and highly structured, [2]. In the paper, we restrict ourselves to covariance matrices P with block tridiagonal inverses.

The Cholesky factor U of the inverse, $P^{-1} = U^T U$, has the following upper bidiagonal block structure

$$U = \begin{bmatrix} U_1 & O_1 & \underline{0} & \underline{0} & . & \underline{0} \\ \underline{0} & U_2 & O_2 & \underline{0} & . & \underline{0} \\ \underline{0} & \underline{0} & U_3 & O_3 & . & \underline{0} \\ . & . & . & \ddots & \ddots & . \\ \underline{0} & . & . & \underline{0} & U_{I-1} & O_{I-1} \\ \underline{0} & . & . & \underline{0} & \underline{0} & U_I \end{bmatrix}. \quad (1)$$

This work is supported by the ONR grant N00014-97-1-0800. The second author is on sabbatical from CMU (moura@ece.cmu.edu).

Matrix U is upper block bidiagonal. Its inverse U^{-1} has the following structure

$$U^{-1} = \begin{bmatrix} U_1^{-1} & \underline{*} & \underline{*} & \underline{*} & \cdot & \underline{*} \\ \underline{0} & U_2^{-1} & \underline{*} & \underline{*} & \cdot & \underline{*} \\ \underline{0} & \underline{0} & U_3^{-1} & \underline{*} & \cdot & \underline{*} \\ & & \ddots & \ddots & & \\ \cdot & \cdot & & & \cdot & \cdot \\ \underline{0} & \cdot & \cdot & \underline{0} & U_{I-1}^{-1} & \underline{*} \\ \underline{0} & \cdot & \cdot & \underline{0} & \underline{0} & U_I^{-1} \end{bmatrix} \quad (2)$$

where the notation $\underline{*}$ stands for unknown entries. The diagonal block entries of U^{-1} are the inverse of the corresponding diagonal blocks of U. The lower triangular block entries in U^{-1} are all zero blocks. These features are used to derive the following results where we show how to obtain, [3]:

(a) the block entries $\{U_i, O_i\}$ of the Cholesky factor U from the blocks $\{P_{ii}\}$ and $\{P_{ii+1}\}$ of the diagonal and first upper diagonal of the covariance matrix P;

(b) the block entries $\{P_{ij}\}$ of P from the block entries $\{U_i, O_i\}$ of the Cholesky factors of P^{-1}; and

(c) the off-diagonal entries P_{ij}, $1 \le i \le I$, $i+2 \le j \le I$ from the main diagonal blocks P_{ii} and the first upper main diagonal blocks P_{ii+1} of P.

Result 1: Given the blocks $\{P_{ii}\}$ and $\{P_{ii+1}\}$ of a covariance matrix P with a block tridiagonal inverse, the blocks $\{U_i, O_i\}$ of the Cholesky factor $U = \mathrm{chol}(P)$ are given by

$$U_I = \mathrm{chol}\left((P_{II})^{-1}\right) \quad (3)$$

$$\left.\begin{aligned} U_i &= \mathrm{chol}\left((P_{ii} - P_{ii+1}P_{i+1i+1}^{-1}P_{ii+1}^T)^{-1}\right) \\ O_i &= -(P_{i+1i+1}^{-1}P_{ii+1}^T U_i^T)^T \end{aligned}\right\} \quad (4)$$

$$\text{for } 1 \le i \le (I-1)$$

where $\mathrm{chol}(\cdot)$ means Cholesky's factorization. ∎

Proof: From the equality

$$P = (U^T U)^{-1}, \text{ we get } PU^T = U^{-1}. \quad (5)$$

Now, replace U by its value in (1), substitute U^{-1} from (2), and express P as $\{P_{ij}\}$, $1 \le i,j \le I$. Then we multiply out the left hand side of (5) and equate the diagonal and lower diagonal block entries on the left and the right hand sides,

$$\begin{aligned} P_{II}U_I^T &= U_I^{-1} && (6) \\ P_{ij}U_j^T + P_{ij+1}O_j^T &= (U_j)^{-1}\delta_{ij} && (7) \\ &\text{for} \quad 1 \le i < (I-1), \quad 1 \le j \le i. \end{aligned}$$

Equation (3) is obtained directly by rearranging terms in (6). To derive (4), we substitute $i = \ell + 1$ and $j = \ell$ in (7),

$$P_{\ell+1\ell}U_\ell^T + P_{\ell+1\ell+1}O_\ell^T = 0. \quad (8)$$

On rearranging (8), the bottom equation in (4) is verified. To prove the expression for U_i substitute $i = j = \ell$ in (7),

$$P_{\ell\ell}U_\ell^T + P_{\ell\ell+1}O_\ell^T = (U_\ell)^{-1} \quad (9)$$

from which the equality for U_ℓ in (4) can be derived by substituting for O_ℓ and rearranging.

Result 2: Given blocks $\{U_i, O_i\}$ of the Cholesky factor $U = \mathrm{chol}(P)$ where P has a block tridiagonal inverse:

(a) The main diagonal blocks P_{ii}, $1 \le i \le I$ of P can be obtained recursively from the following expressions

$$\begin{aligned} P_{II} &= \left(U_I^T U_I\right)^{-1} && (10) \\ P_{ii} &= \left(U_i^T U_i\right)^{-1} + \left(U_i^{-1}O_i\right)P_{i+1i+1}\left(U_i^{-1}O_i\right)^T && (11) \\ &\text{for } (I-1) \ge i \ge 1. \end{aligned}$$

(b) The remaining upper triangular blocks P_{ii+k}, $1 \le i \le (I-1)$, $1 \le k \le (I-i)$, are given by

$$P_{ii+k} = \left(\prod_{\tau=i}^{i+k-1}\left(-U_\tau^{-1}O_\tau\right)\right)P_{i+ki+k} \quad (12)$$

or alternatively, for $1 \le i \le (I-1)$, $(i+1) \le j \le I$, by

$$P_{ij} = \left(\prod_{\tau=i}^{j-2}\left(-U_\tau^{-1}O_\tau\right)\right)P_{j-1j}. \quad ∎ \quad (13)$$

Proof: Equation (10) follows directly by rearranging terms in (6). We prove (12) by induction. Due to lack of space, we do not include here the proof of (13). It follows the same lines as that of (12). Finally, we prove expression (11).

Case 1: Equation (12).

In (8), right multiply by U_ℓ^{-T}, take the transpose on both sides, and rearrange, to get

$$P_{\ell\ell+1} = \left(-U_\ell^{-1}O_\ell\right)P_{\ell+1\ell+1} \quad (14)$$

which is equation (12) for $k = 1$.

By the induction step, (12) is valid for $k = \wp$, i.e.,

$$P_{ii+\wp} = \left(\prod_{\tau=i}^{i+\wp-1}\left(-U_\tau^{-1}O_\tau\right)\right)P_{i+\wp i+\wp}. \quad (15)$$

We now prove equation (12) for $k = \wp + 1$.

With $i = \ell + \wp + 1$ and $j = \ell$ in equation (7) and following the steps that led to (14), we get

$$P_{\ell\ell+\wp+1} = \left(-U_\ell^{-1}O_\ell\right)P_{\ell+1\ell+\wp+1}. \quad (16)$$

Substituting for $P_{\ell+1\ell+\wp+1}$ from equation (15) with $i = \ell+1$ proves equation (12) for $k = \wp + 1$.

Case 2: Equation (11) follows by right multiplying (9) by U_ℓ^{-T}, and substituting for $P_{\ell\ell+1}$ from Result (12).

Result 3: Given the main diagonal and the upper diagonal blocks $\{P_{ii}\}$ and $\{P_{ii+1}\}$ of the covariance matrix P with block tridiagonal inverse, the upper triangular blocks P_{ij}, $1 \le i \le I$, $i+2 \le j \le I$, are given by

$$P_{ij} = \left(\prod_{\tau=i}^{j-2}\left(P_{\tau+1\tau+1}^{-1}P_{\tau+1\tau}\right)^T\right)P_{j-1j}. \quad ∎ \quad (17)$$

Proof: Result 3 is derived from (13) by expressing the product $\left(-U_\tau^{-1}O_\tau\right)$ in terms of the main and the upper diagonal blocks of P using (8).

3. INVERTING BANDED MATRICES

In this section, we use Result 1 to derive an algorithm for inverting a full matrix, P, whose inverse is block tridiagonal. Result 2 is then used to solve the converse problem, i.e., inverting a block tridiagonal matrix. Since in both cases, we are dealing with symmetric matrices, we compute only the diagonal and the upper triangular blocks.

Algorithm 1: Inverting matrices that have tridiagonal block inverses. Using the notation introduced in section 2, Algorithm 1 computes $A = P^{-1}$ from P as follows.
Step 1: Compute P_{ii}^{-1}, $1 \leq i \leq I$.
Step 2: Calculate the product terms $U_i^T U_i$, $U_i^T O_i$ and $O_i^T O_i$, $1 \leq i \leq (I-1)$, using the following recursive expressions

$$U_i^T U_i = (P_{ii} - P_{ii+1} P_{i+1i+1}^{-1} P_{ii+1}^T)^{-1} \quad (18)$$
$$U_i^T O_i = -(U_i^T U_i)\, P_{ii+1} P_{i+1i+1}^{-1} \quad (19)$$
$$O_i^T O_i = -P_{i+1i+1}^{-1} P_{ii+1}^T\, (U_i^T O_i). \quad (20)$$

The only other product term needed to compute A is

$$U_I^T U_I = P_{II}^{-1}. \quad (21)$$

Equations (18)-(21) follow directly from Result 1.
Step 3: The diagonal block entries A_{ii} and the upper diagonal entries, A_{ii+1}, $1 \leq i \leq (I-1)$, are given by

$$A_{i+1i+1} = U_{i+1}^T U_{i+1} + O_{i+1}^T O_{i+1} \quad (22)$$
$$A_{ii+1} = U_i^T O_i. \quad (23)$$

The top left block A_{11} equals $U_1^T U_1$. The lower diagonal blocks $A_{i+1i} = A_{ii+1}^T$. Since A is a tridiagonal block matrix, the remaining blocks in A are all zero blocks, $\underline{0}$.

Equations (22)-(23) are derived by expanding the expression $A = U^T U$ in terms of the constituent blocks.

By counting the number of operations, it can be verified that the above algorithm is $O(I^4)$, a reduction of I^2 over the direct inversion of the covariance matrix P.

Algorithm 2: Inverting tridiagonal block matrices. Following the notation of section 2, Algorithm 2 computes P from A or its Cholesky factors $\{U_i, O_i\}$ as described below.
Step 1: Compute U_i^{-1}, $1 \leq i \leq I$.
Step 2: Compute the product terms

$$(U_i^T U_i)^{-1} = U_i^{-1} U_i^{-T} \quad \text{and } U_i^{-1} O_i, \quad 1 \leq i \leq I \quad (24)$$

Step 3: We use Result 2 to derive alternative expressions for the main diagonal block entries, P_{ii} and for the entries on the upper diagonals P_{ii+k}, $k \geq 1$. The lower right main diagonal block entry is given by

$$P_{II} = \left(U_I^T U_I\right)^{-1}. \quad (25)$$

The remaining diagonal blocks, P_{ii}, are computed recursively by starting from $i = I-1$ and decrementing i till $i = 1$ in the following expression

$$P_{ii} = \left(U_i^T U_i\right)^{-1} + \left(U_i^{-1} O_i\right) P_{i+1i+1} \left(U_i^{-1} O_i\right)^T. \quad (26)$$

The upper diagonal entries P_{ii+k}, $1 \leq i \leq (I-1)$ and $1 \leq k \leq (I-i)$, are given by

$$P_{ii+k} = \left(-U_i^{-1} O_i\right) P_{i+1i+k}^T. \quad (27)$$

Algorithm 2 highlights the following important points:

1. The diagonal blocks can be found directly from $\{U_i, O_i\}$ without computing any off diagonal blocks, P_{ij}, $i \neq j$.
2. In general, the block entries on the n'th diagonal, $k = n$, can be computed from $\{U_i, O_i\}$ and the blocks on the diagonal just before it, $k = (n-1)$.
3. Computing each entry on the main diagonal is of $O(I^3)$. Since there are I blocks on the main diagonal, the operation count for computing all entries on the main diagonal is of $O(I^4)$.
4. The total number of blocks on the n'th diagonal, $k = n$, is $(I-n)$. If the block entries on the preceeding diagonal, $k = (n-1)$, are known, the total computations for the n'th diagonal are $(I-n)O(I^3)$.
5. Inverting A using Algorithm 2 is of $O(I^5)$, an improvement of I over the direct computation of P.

3.1. Local KBf: Banded Matrix Approximation

In this section, we present an approach for approximating the covariance matrix computation in the Kalman-Bucy filter (KBf). We describe our approximate filter in the context of filtering random fields that evolve with time. In other words, we have image fields defined on a grid whose temporal dynamics follow a certain discretization of a partial differential equation. The image fields are stacked in a long vector – the state vector. The key to our approach is approximating the inverse of the error covariance matrix (the information matrix) by a block banded matrix. This corresponds to modeling the error field in the spatial dimensions at each point in time as a reduced-order noncausal Gauss Markov random field (GMrf). For a first order GMrf approximation, the information matrix is modeled by a tridiagonal block matrix. In [4], it is proved that these are GMrf approximations that optimize the Kullback-Leibler mean information distance criterion.

Our approximation algorithm uses Result 3. Instead of updating the entire covariance matrix at each time iteration we update only a few diagonals of the error covariance matrix. For a first order GMrf approximation, only the main diagonal block entries P_{ii} and the upper main diagonal block entries P_{ii+1} are updated. Any other blocks, P_{ij}'s if required may then be obtained directly from the P_{ii}'s and the P_{ii+1}'s using Result 3. Since fewer blocks are updated, the local KBf requires less computations. In fact for the problem defined below, the computations are reduced by two orders of magnitude on the linear dimension of the grid.

We illustrate our approach by applying the banded matrix approximation to computing the error covariance matrix in a KBf problem. Our experiment shows how the estimates of the approximate KBf – the local KBf – closely tracks the optimal estimates. To be able to compute the optimal estimates, we need to run the exact KBf. Due to this, we run our experiment intentionally on a grid of dimensions 11×11 for which the optimal KBf is of $O(10^8)$ and is still computationally feasible.

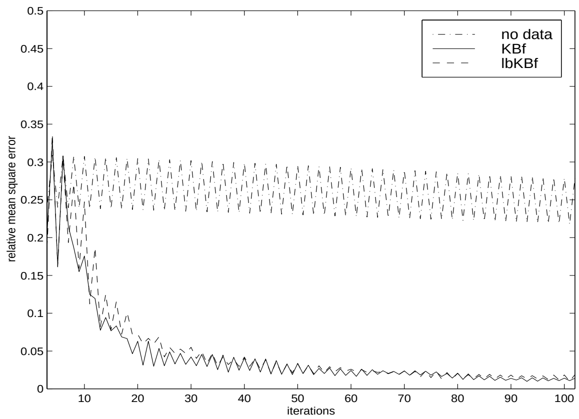

Figure 1: Mean Square Error (MSE) for the local KBf and for the optimal KBf.

4. LOCAL KBF

To examine the effect of our approximation, consider applying the KBf to the following dynamic problem

$$\frac{\partial u}{\partial t} - k_1 y v + k_2 \frac{\partial \eta}{\partial x} = X, \quad \frac{\partial v}{\partial t} + k_1 y u + k_2 \frac{\partial \eta}{\partial y} = 0 \quad (28)$$

$$\frac{\partial \eta}{\partial t} + k_3 \frac{\partial u}{\partial x} + k_3 \frac{\partial v}{\partial y} = 0 \quad (29)$$

where (u, v, η) are the coupled fields to be estimated, (x, y, t) are the independent spatial and time variables, (k_1, k_2, k_3) are constants, and X is the input forcing term assumed Gaussian. The dynamical model considered above is used frequently in thermodynamics and other flow problems.

Starting with two different initial conditions, the dynamical model, (28)-(29), is discretized using a leap-frog finite difference scheme and propagated forward in time. The resulting fields provide the starting point for the KBf experiment. One set of fields simulates the real world fields that we are trying to estimate from which the data is observed. The other set is used as the initial conditions in the optimal and the local implementations of the KBf.

The measurement model is given by

$$y(t) = H(t)\eta(t) + w(t) \quad (30)$$

where $H(t)$ corresponds to the time-varying observation matrix and $w(t)$ is the observation noise, assumed Gaussian. We assume that data is available only for the field η along a few known rows and that at each time iteration, different sets of rows are observed. This property defines the structure of $H(t)$ and makes the measurements extremely sparse. Such sparsity arises in remote sensing applications with satellite observations.

Figure 1 shows the evolution over time of the mean square error (MSE) for the image field η for the two KBf's. The solid line near the bottom represents the estimation error of the optimal KBf and the dotted line represents the estimation error of the local KBf. We include in Fig. 1 the MSE when no data is assimilated. This is represented by the semisolid oscillating line near the top. As it can be observed, the optimal and the local KBf's reduce significantly the error. More importantly, the local implementation follows closely the optimal KBf showing that the GMrf approximation is actually a very reasonable approximation.

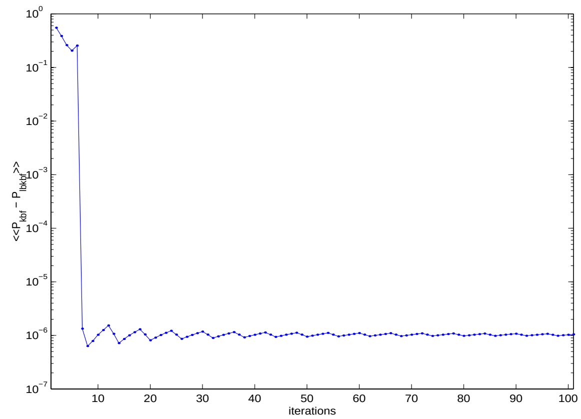

Figure 2: Comparison of the error covariance matrices for the local and for the optimal KBfs as a function of time.

To measure the closeness of the error covariance matrix of the local KBf to the optimal error covariance matrix, we plot in Figure 2 the approximation errors, defined as $\| P_{\text{optimal}} - P_{\text{suboptimal}} \| / I^2$ where we use the Frobenius norm and $I = 726$. Figure 2 shows that after a short transient the difference between the error covariance matrices of the optimal KBf and its local approximation is small.

5. SUMMARY

We have presented several important properties for a block banded matrix that relate its constituent blocks to the block entries of its inverse. These properties are used to derive efficient algorithms to invert band-limited matrices and to invert matrices whose inverse are known to be block banded. The algorithms provide computational savings of up to two orders of magnitude. We applied these algorithms to develop a sub-optimal implementation of the Kalman-Bucy filter (KBf) that approximates its error covariance matrix at each time step by a matrix whose inverse is block-banded. We call this approximate filter the local KBf. The resulting local KBf reduces the computational complexity by two orders of magnitude of the linear dimension of the field. As illustrated by an experiment, the local KBf follows the standard KBf closely and exhibits near-optimal performance.

6. REFERENCES

[1] R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," *ASME Journal of Basic Engineering*, vol. 83, pp. 95–108, 1961.

[2] J. M. F. Moura and N. Balram, "Recursive Structure of Noncausal Gauss Markov Random Fields," *IEEE Trans. on Inf. Theory*, vol. IT-38(2), pp. 334–354, 1992.

[3] A. Asif and J. Moura, "Data assimilation in large time varying multidimensional fields," accepted for publication in *IEEE Trans. of Image Processing*.

[4] A. Kavcic and J. M. F. Moura, "Information Loss in Markov Approximations," tech. rep., Department of Electrical Engineering, Carnegie Mellon University, 1998. manuscript of 30 pages, submitted to publication.