IPN Progress Report 42-233 • May 15, 2023

Square-Root Formulas for Kalman Filter, Information Filter, and RTS Smoother: Links via Boomerang Prediction Residual

Toshio Michael Chin*

ABSTRACT. — A self-contained derivation of several square-root filter and smoother formulas is presented. The formulas include square-root versions of the Kalman filter, Rauch-Tung-Striebel (RTS) smoother, and Dyer-McReynolds Covariance Smoother (DMCS), along with the square-root information filter (SRIF). A stabilized version of the RTS smoother is also included. While most of the presented formulas have been independently derived and well documented, this presentation takes an unified approach through a random process named *boomerang prediction residual*, which simplifies both the derivations and formulas.

I. Introduction

Since its first application to aerospace navigation and guidance in 1960s, the Kalman filter has become a common numerical tool in a wide range of applications for tracking the state of a dynamic system based on incomplete and noisy observations. From the beginning, algebraic realizations of the filter algorithm played a key role in successful applications, as the legendary invention of the square-root filter enabled onboard deployment for the Apollo missions [1]. The square-root version significantly reduces round-off errors from digitization that can cause filter failure. While the principal formulations for the square-root Kalman filter had mostly been established by the mid-1970s [2, 3, 4], research on the square-root algorithms continued along with advances in computational technology, including dedicated hardware for Kalman filters [5] and method for parallel and distributed computing [6]. More recently, Kalman filter implementations can rely on highly optimized linear algebra software packages such as LINPACK/LAPACK [7]. Use of the QR-factorization, in particular, is seen often in square-root filter implementations [8, 9].

---

*Tracking Systems and Applications Section.

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

This note presents a set of square-root realizations for the Kalman filter as well as the Rauch-Tung-Striebel (RTS) smoother [10], which produces the fixed-interval smoothing estimates by retrospectively upgrading the Kalman filter results. Development here has been motivated by applications to space geodetic data for terrestrial reference frame estimation [11, 12] in which the variance estimates are susceptible to becoming negative due to digital round-offs. Generally, the standard Kalman filter and RTS smoother formulas share a weakness in preserving numerical positive semi-definiteness in their covariance matrix estimates. Addressing such an algorithmic weakness is a focal point of this note. The square-root Kalman filter algorithms presented here are not new and are well documented [3, 13, 14]. The accompanying square-root RTS smoother algorithm, on the other hand, appears to have a simpler algebraic form than the existing ones [15] due partly to a new derivation presented here. The smoother works well in tandem with either the square-root Kalman filter or the square-root information filter (SRIF) [16]. The filter and smoother algorithms here can be executed using only the standard LAPACK/BLAS linear algebra subroutines for the QR and Cholesky factorizations and for triangular matrix inversion and multiplication. The presentation is intended for general filtering and smoothing problems beyond the target applications.

II. Background

Filtering and smoothing are sequential estimation of a stochastic process x_t based on its observations y_t indexed by time $t = 1, \ldots, T$. The original Kalman filter algorithm caters to the linear stochastic dynamics and observation equations of the form

$$x_{t+1} = F_t x_t + w_t \tag{1}$$

$$y_t = H_t x_t + v_t \tag{2}$$

given F_t, H_t, $x_1 \sim \left(\bar{x}_1,\ \bar{P}_1 \right)$, $w_t \sim (\ \bar{q}_t,\ Q_t\ )$, and $v_t \sim (\ 0,\ R_t\ )$, where "$a \sim (\ b,\ C\ )$" denotes that "the random vector a has mean b and covariance matrix C". The random vectors x_1, w_t, and v_t are assumed to be mutually independent for all t.

The mean $\bar{q}_t$ can become non-zero if the stochastic process has a forcing or control input. The *zero-mean condition* of $\bar{q}_t = 0$ and $\bar{x}_1 = 0$, however, can be assumed without any loss of generality, since x_t and y_t from (1, 2) can be considered as the perturbation about the mean state trajectory $\bar{x}_t$ (computed as $\bar{x}_{t+1} = F_t \bar{x}_t + \bar{q}_t$ given $\bar{x}_1$) and simulated observation $\bar{y}_t \equiv H_t \bar{x}_t$ according to the common practice known as the *Extended Kalman filter* [13, 17]. The zero-mean condition also facilitates development of square-root algorithms and is assumed throughout this presentation.

The filter and smoother seek the optimal estimate of the dynamic state x_t as a *linear* function of the observations y_t. Optimality is defined by the least mean-squares error condition (Appendix II.A). For brevity, we adopt the notations found in [16, 18] for the state estimates:

$$
\begin{array}{rclcll}
x_t^{\sim} & \equiv & \mathcal{L}[\,x_t|y_1,\dots,y_{t-1}\,] & = & \hat{x}_{t|t-1} & : \text{pre-update prediction} \\
x_t^{\circ} & \equiv & \mathcal{L}[\,x_t|y_1,\dots,y_t\,] & = & \hat{x}_{t|t} & : \text{filtered estimate} \\
x_t^{\star} & \equiv & \mathcal{L}[\,x_t|y_1,\dots,y_T\,] & = & \hat{x}_{t|T} & : \text{smoothed estimate}
\end{array} \tag{3}
$$

where $\mathcal{L}[a|b]$ denotes "the linear function of b that optimally estimates a" and the associated estimation error covariance matrices:

$$
\begin{array}{rclcll}
P_t^{\sim} & \equiv & \mathcal{E}[(x_t - x_t^{\sim})(x_t - x_t^{\sim})^{\top}] & = & P_{t|t-1} & : \text{pre-update error covariance} \\
P_t^{\circ} & \equiv & \mathcal{E}[(x_t - x_t^{\circ})(x_t - x_t^{\circ})^{\top}] & = & P_{t|t} & : \text{filter error covariance} \\
P_t^{\star} & \equiv & \mathcal{E}[(x_t - x_t^{\star})(x_t - x_t^{\star})^{\top}] & = & P_{t|T} & : \text{smoother error covariance}
\end{array} \tag{4}
$$

where $\mathcal{E}$ denotes the expectation operator.[1] If x_1, w_t, and v_t all have Gaussian distributions, the optimal linear estimator is also the conditional mean (Appendix II.A) so that each $\mathcal{L}$ in (3) could be replaced by the expectation operator $\mathcal{E}$. For reference only, the third columns of (3, 4) show notations common in the literature under such a Gaussian assumption [13].

For further brevity, we omit the time index unless the time is other than t. From here on, any variable without a subscript can be assumed to be implicitly indexed by t (except for the identity matrix I). Also, the superscripts $^{\top}$, $^{-1}$, and $^{-\top}$ respectively denote the transpose, inverse, and transposed-inverse of a matrix. The standard Kalman filter equations can then be written as

$$Y = HP^{\sim}H^{\top} + R \tag{5}$$
$$K = P^{\sim}H^{\top}Y^{-1} \tag{6}$$
$$x^{\circ} = x^{\sim} + K(y - Hx^{\sim}) \tag{7}$$
$$P^{\circ} = P^{\sim} - KYK^{\top} \tag{8}$$
$$x_{t+1}^{\sim} = Fx^{\circ} \tag{9}$$
$$P_{t+1}^{\sim} = FP^{\circ}F^{\top} + Q \tag{10}$$

where K is often called the *Kalman gain* and Y is the covariance matrix associated with the *innovation process* defined as $\eta_t \equiv y_t - H_t x_t^{\sim}$. The recursions are initialized as $x_1^{\sim} = 0$ and $P_1^{\sim} = \bar{P}_1$. The data update steps (5–8) revise the estimate-covariance pair based on the new observation (2), while the time update steps (9, 10) do so based on the dynamics (1).

The RTS smoother equations are initialized by the final filter estimates $(\,x_T^{\star}, P_T^{\star}\,) = (\,x_T^{\circ}, P_T^{\circ}\,)$ and iterate backward in time $t = T-1, T-2, \dots, 1$ as

$$S = P^{\circ}F^{\top}(P_{t+1}^{\sim})^{-1} \tag{11}$$
$$x^{\star} = x^{\circ} + S(x_{t+1}^{\star} - x_{t+1}^{\sim}) \tag{12}$$
$$P^{\star} = P^{\circ} - S(P_{t+1}^{\sim} - P_{t+1}^{\star})S^{\top} \tag{13}$$

---

[1] Given the probability density function $p(x)$ and a generic function $q(x)$ for a random vector x, $\mathcal{E}[q(x)] = \int q(x)\,p(x)\,dx$ where the integral is taken over the entire domain of x.

where S_t will be referred to as the *smoother gain*. Tables 1–3 summarize the standard filter and smoother formulas.

A well-known weakness in these formulas is that subtractions of one covariance matrix from another, which appear in (8) and (13), are vulnerable to digital round-offs that may make the difference matrix non-definite, violating the fundamental characteristic of a covariance matrix to be positive semi-definite and leading often to swift divergence from optimality.

III. Stabilized Forms

The covariance formulas (8, 13) can be altered into "stabilized forms" to replace the numerically vulnerable covariance differences with quadratic terms. A stabilized form of RTS covariance recursion (13), rare or absent in the literature, is presented here. Its derivation depends on the introduction of a process labeled here as "the boomerang prediction residual," which reappears prominently in the square-root forms of the Kalman filter and RTS smoother presented later.

A. Joseph Stabilized Form

The Kalman filter data-update formula (7) can be written for the estimation error $x_t - x_t^\circ$ as

$$\begin{aligned} x - x^\circ &= x - x^\sim - K(y - Hx^\sim) \\ &= x - x^\sim - KH(x - x^\sim) - Kv \\ &= (I - KH)(x - x^\sim) - Kv. \end{aligned}$$

Since the prior estimation error $x_t - x_t^\sim$ is setup to be independent from the observation error v_t,

$$\begin{aligned} P^\circ &= \mathcal{E}[(x - x^\circ)(x - x^\circ)^\top] \\ &= (I - KH)P^\sim(I - KH)^\top + KRK^\top \end{aligned} \quad (14)$$

which is the *Joseph stabilized form* [18] that can impose positive semi-definiteness on P_t° through its algebraic form that, unlike (8), contains no difference between covariance matrices. The computational cost can be significantly higher with (14) than (8) due to an increase in matrix multiplications.

B. Boomerang Prediction Residual

Define $\delta_t \equiv x_t - S_t x_{t+1}$ to be the *boomerang prediction residual* process.[2] Its estimate based on the filter data set $\{y_1, \ldots, y_t\}$ is denoted as δ_t° and given by

$$\delta^\circ = x^\circ - S x_{t+1}^\sim \quad (15)$$

[2] $S_t x_{t+1}$ can be considered as a hindcast of x_t based on x_{t+1}, which in turn is a forecast based on x_t itself.

Table 1. Filter parameters from the system equations

Filter parameter		Source
F_t	State transition matrix	$x_{t+1} = F_t x_t + w_t$
H_t	Observation matrix	$y_t = H_t x_t + v_t$
Q_t	Process noise covariance	$w_t \sim (0, Q_t)$
R_t	Observation noise covariance	$v_t \sim (0, R_t)$
$\bar{P}_1$	Initial state covariance	$x_1 \sim (0, \bar{P}_1)$
T	Time (index) interval	$t = 1, 2, \ldots T$

Table 2. Variables in the filter/smoother recursions

State estimate	Type	Estimation error covariance
$x_t^{\sim} \equiv \mathcal{L}[\,x_t \vert y_1, \ldots, y_{t-1}\,]$	Pre-update prediction	$P_t^{\sim} \equiv \mathcal{E}(x_t - x_t^{\sim})(x_t - x_t^{\sim})^\top$
$x_t^{\circ} \equiv \mathcal{L}[\,x_t \vert y_1, \ldots, y_t\,]$	Filtered estimate	$P_t^{\circ} \equiv \mathcal{E}(x_t - x_t^{\circ})(x_t - x_t^{\circ})^\top$
$x_t^{\star} \equiv \mathcal{L}[\,x_t \vert y_1, \ldots, y_T\,]$	Smoothed estimate	$P_t^{\star} \equiv \mathcal{E}(x_t - x_t^{\star})(x_t - x_t^{\star})^\top$

Table 3. A summary of standard Kalman filter and RTS smoother equations

	Kalman Filter	RTS Smoother
Initializations	$x_1^{\sim} = 0, \quad P_1^{\sim} = \bar{P}_1$	$x_T^{\star} = x_T^{\circ}, \quad P_T^{\star} = P_T^{\circ}$
Recursions	(for $t = 1, 2, \ldots, T$)	(for $t = T-1, T-2, \ldots, 1$)
	$Y_t = H_t P_t^{\sim} H_t^\top + R_t$	$S_t = P_t^{\circ} F_t^\top (P_{t+1}^{\sim})^{-1}$
	$K_t = P_t^{\sim} H_t^\top Y_t^{-1}$	$x_t^{\star} = x_t^{\circ} + S_t(x_{t+1}^{\star} - x_{t+1}^{\sim})$
	$x_t^{\circ} = x_t^{\sim} + K_t(y_t - H_t x_t^{\sim})$	$P_t^{\star} = P_t^{\circ} - S_t(P_{t+1}^{\sim} - P_{t+1}^{\star}) S_t^\top$
	$P_t^{\circ} = P_t^{\sim} - K_t Y_t K_t^\top$	
	$x_{t+1}^{\sim} = F_t x_t^{\circ}$	
	$P_{t+1}^{\sim} = F_t P_t^{\circ} F_t^\top + Q_t$	

and the associated error covariance matrix is denoted as Δ_t° and given by

$$\begin{aligned} \Delta^\circ &= \mathcal{E}[(\delta - \delta^\circ)(\delta - \delta^\circ)^\top] \\ &= (I - SF)P^\circ(I - SF)^\top + SQS^\top \end{aligned} \tag{16}$$

whose derivation is through the estimation error

$$\begin{aligned} \delta - \delta^\circ &= (x - x^\circ) - S(x_{t+1} - x^\sim_{t+1}) \\ &= (x - x^\circ) - S(Fx + w - Fx^\circ) \\ &= (I - SF)(x - x^\circ) - Sw \end{aligned}$$

expressed in terms of mutually independent x_t and w_t.

Lemma III.1. *The filter error for the boomerang prediction residual is uncorrelated to all future smoother error, or* $\mathcal{E}[(\delta_t - \delta_t^\circ)(x_{t+j} - x^\star_{t+j})^\top] = 0$ *for* $j = 1, \ldots, T-t$.

`Proof:` For each j in $j = 1, \ldots, T-t$.

$$\begin{aligned} \mathcal{E}[(\delta - \delta^\circ)(x_{t+j} - x^\star_{t+j})^\top] &= \mathcal{E}[\{(x - x^\circ) - S(x_{t+1} - x^\sim_{t+1})\}(x_{t+j} - x^\star_{t+j})^\top] \\ &= \mathcal{E}[\{(x - x^\circ) - S(x_{t+1} - x^\sim_{t+1})\}(x_{t+j} - x^\sim_{t+j})^\top (W^f_{t+j})^\top] \\ &= P^\circ F^\top \cdot F_1^j (W^f_{t+j})^\top - SP^\sim_{t+1} \cdot F_1^j (W^f_{t+j})^\top \\ &= \mathbf{0} \end{aligned} \tag{17}$$

where $F_1^j \equiv I$ for $j = 1$ and is the product $F_{t+1}^\top \cdots F_{t+j-1}^\top$ for $j > 1$, the second equality is due to (49–51) from Appendix II.C, each W^f_{t+j} is a constant matrix, and the last equality is due to (11).

By the orthogonality principle [19], Lemma III.1 implies that the boomerang prediction residual process provides independent (additive) data to the smoother estimates during backward recursion.

C. Stabilized RTS Smoother

The RTS equation for the smoother estimate (12) can be rewritten as a recursion driven by the filtered estimate δ_t° of the boomerang prediction residual as

$$x^\star = Sx^\star_{t+1} + \delta^\circ. \tag{18}$$

The smoother error can then be expressed in terms of the two mutually independent processes from Lemma III.1 as

$$x - x^\star = S(x_{t+1} - x^\star_{t+1}) + (\delta - \delta^\circ)$$

which, helped by (17), leads to a recursion for the smoother error covariance matrix

$$P^\star = SP^\star_{t+1}S^\top + \Delta^\circ. \tag{19}$$

Substitution of (16) into (19) leads to a stabilized RTS smoother covariance recursion

$$P^\star = S(P^\star_{t+1} + Q)S^\top + (I - SF)P^\circ(I - SF)^\top \tag{20}$$

which, unlike (13), is free from differencing two positive semi-definite matrices.

IV. Square-root Covariance Filter and Smoother

The Kalman filter covariance recursion (5, 6, 8, 10) is independent from the state estimate recursion (which in contrast is dependent on the covariance recursion through the gain matrix). The same is true for the RTS smoother covariance recursion (11, 13). The classical approach to square-root filter and smoother is to perform these covariance recursions with a square-root $\sqrt{P}$ of the covariance matrix P, defined here as a *square* matrix such that $\sqrt{P}\sqrt{P}^\top = P$. Using existing software packages, Cholesky factorization of P usually yields $\sqrt{P}$ in a lower-triangular form. If P is singular, Cholesky factorization fails, and finding $\sqrt{P}$ requires a more sophisticated routine such as Matlab's `sqrtm` [20].

A. Data Update

Consider updating a given estimate-covariance pair $(x_t^\sim, P_t^\sim)$ with the observation given as (2). The joint distribution of the observation and state vectors then becomes

$$\begin{bmatrix} y \\ x \end{bmatrix} \sim \left(\begin{bmatrix} Hx^\sim \\ x^\sim \end{bmatrix}, \begin{bmatrix} Y & HP^\sim \\ P^\sim H^\top & P^\sim \end{bmatrix} \right) \tag{21}$$

whose (joint) mean vector and covariance matrix provide the ingredients for the textbook formulas for the optimal linear estimator of x_t given y_t, which are the Kalman filter data-update formulas (7, 8). In particular, given the joint covariance matrix, the formula (8) for the data-updated covariance P_t° is the *Schur complement* of Y_t (Appendix II.A).

Keys to the square-root filter algorithms are that two different square-roots of the same (symmetric) matrix are related by a unitary transform (Appendix I.A) and that square-root factors with (block) triangular matrix structures can produce a Schur complement (Appendix I.B). In particular, the joint covariance matrix in (21) can be factored individually as $\mathbf{A}_1^\top \mathbf{A}_1$ and $\mathbf{B}_1^\top \mathbf{B}_1$ by the square-root matrices

$$\mathbf{A}_1^\top \equiv \begin{bmatrix} \sqrt{R} & H\sqrt{P^\sim} \\ \mathbf{0} & \sqrt{P^\sim} \end{bmatrix}, \quad \mathbf{B}_1^\top \equiv \begin{bmatrix} \sqrt{Y} & \mathbf{0} \\ K\sqrt{Y} & \sqrt{P^\circ} \end{bmatrix}$$

where $\mathbf{A}_1^\top$ contains only the given parameters and the lower-triangular $\mathbf{B}_1^\top$ exposes the desired parameters including the square-root Schur complement $\sqrt{P_t^\circ}$ for the data-updated square-root covariance matrix. Since $\mathbf{B}_1$ has an upper triangular structure, applying the QR factorization to $\mathbf{A}_1$ would yield $\mathbf{B}_1$. Some linear algebra packages including LAPACK offer "LQ factorization", an alternate of QR factorization, that transforms $\mathbf{A}_1^\top$ directly to the lower-triangular $\mathbf{B}_1^\top$.

The Kalman gain can also be derived from the left column of $\mathbf{B}_1^\top$. Explicit computation for K_t is not necessary for the state estimate update

$$x^\circ \quad = \quad x^\sim - (K\sqrt{Y}) \cdot (\sqrt{Y})^{-1}(y - Hx^\sim) \tag{22}$$

since the vector $(\sqrt{Y})^{-1}(y - Hx^\sim)$ can be computed first via back-substitution with the upper-triangular $\sqrt{Y}$ and then be multiplied by $K\sqrt{Y}$ extracted from $\mathbf{B}_1^\top$.

B. Time Update and Smoother Parameters

The covariance time-update formula (10) is a sum of two symmetric matrices, which can be expressed as $\mathbf{B}_{2f}^\top \mathbf{B}_{2f} = \mathbf{A}_{2f}^\top \mathbf{A}_{2f}$ by the square-root factors

$$\mathbf{A}_{2f}^\top \equiv \begin{bmatrix} \sqrt{Q} & F\sqrt{P^\circ} \end{bmatrix}, \quad \mathbf{B}_{2f}^\top \equiv \begin{bmatrix} \sqrt{P^\sim_{t+1}} & \mathbf{0} \end{bmatrix}.$$

An LQ or QR factorization can transform $\mathbf{A}_{2f}^\top$ to the lower-triangular $\mathbf{B}_{2f}^\top$, from which the desired $\sqrt{P^\sim_{t+1}}$ can be extracted.

If the smoother estimates are desired in addition, the equation of dynamics (1) can be considered as an "observation" of x_t by x_{t+1}. This perspective is relevant to smoothing because the retrospective (backward time) estimate of x_t given x_{t+1} is statistically independent from the causal (forward time) prediction of x_t based on x°_{t-1} due to the Markov property of the state-space model (Appendix II.C). Because (1) and (2) have similar algebraic forms, the square-root procedure developed for the data-update using (2) should be applicable to the time-update using (1) as well.[3] In particular, an analogue of (21) would be the joint distribution

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} \sim \left(\begin{bmatrix} Fx^\circ \\ x^\circ \end{bmatrix}, \begin{bmatrix} P^\sim_{t+1} & FP^\circ \\ P^\circ F^\top & P^\circ \end{bmatrix} \right) \tag{23}$$

whose joint covariance matrix has two square-root factors:

$$\mathbf{A}_2^\top \equiv \begin{bmatrix} \sqrt{Q} & F\sqrt{P^\circ} \\ \mathbf{0} & \sqrt{P^\circ} \end{bmatrix}, \quad \mathbf{B}_2^\top \equiv \begin{bmatrix} \sqrt{P^\sim_{t+1}} & \mathbf{0} \\ S\sqrt{P^\sim_{t+1}} & \sqrt{\Delta^\circ} \end{bmatrix}$$

where the top rows of $\mathbf{A}_2^\top$ and $\mathbf{B}_2^\top$ are $\mathbf{A}_{2f}^\top$ and $\mathbf{B}_{2f}^\top$, respectively, and Δ°_t is the error covariance matrix (16) for the boomerang prediction residual process δ_t. Expansion of (16) leads to a simpler form

$$\begin{aligned} \Delta^\circ &= P^\circ - SFP^\circ - P^\circ F^\top S^\top + SP^\sim_{t+1}S^\top \\ &= P^\circ - SP^\sim_{t+1}S^\top - SP^\sim_{t+1}S^\top + SP^\sim_{t+1}S^\top \\ &= P^\circ - SP^\sim_{t+1}S^\top \end{aligned} \tag{24}$$

which equals to $P^\circ_t - P^\circ_t F_t^\top P^\sim_{t+1} F_t P^\circ_t$ or the Schur complement of $P^\sim_{t+1}$ in the joint covariance matrix of (23). An LQ or QR factorization could transform $\mathbf{A}_2^\top$, containing only the prior parameters, to the lower triangular $\mathbf{B}_2^\top$. The left block column of $\mathbf{B}_2^\top$ yields the RTS smoother gain as

$$S = \left(S\sqrt{P^\sim_{t+1}} \right) \cdot \sqrt{P^\sim_{t+1}}^{\,-1} \tag{25}$$

which is to be stored along with $\sqrt{\Delta^\circ_t}$ for the smoother recursion. The boomerang prediction residual estimate δ°_t is also computed (15) and stored.

---

[3] Note also the analogous algebraic forms for K_t and S_t in (6, 11), as well as for Y_t and $P^\sim_{t+1}$ in (5, 10). However, actual observation of x_{t+1} is not available during the filter time update. A value $x^\star_{t+1}$ becomes available only during the backward smoother recursion.

C. Smoothing

The stabilized smoother recursions (18, 19) can be identified in algebraic form as the Dyer-McReynolds Covariance Smoother (DMCS) [16]. DMCS, in turn, is identical to the RTS smoother. In particular, substitutions of (15) and (24) into (18) and (19), respectively, lead to the RTS recursions (12) and (13). The DMCS recursions (18, 19) are thus considered equivalent to the RTS recursions.

The smoother covariance recursion (19) can be expressed as $\mathbf{B}_3^\top \mathbf{B}_3 = \mathbf{A}_3^\top \mathbf{A}_3$ by the factors

$$\mathbf{A}_3^\top \equiv \left[\, S\sqrt{P^\star_{t+1}} \quad \sqrt{\Delta^\circ} \,\right], \quad \mathbf{B}_3^\top \equiv \left[\, \sqrt{P^\star_t} \quad \mathbf{0} \,\right]$$

where S_t and $\sqrt{\Delta^\circ_t}$ in $\mathbf{A}_3^\top$ are from the filter time-update. Again, an LQ or QR factorization can transform $\mathbf{A}_3^\top$ to the lower-triangular $\mathbf{B}_3^\top$.

Unlike its Kalman filter counterpart (7, 9), the smoother state-estimate recursion (18) is independent from the accompanying covariance recursion (19). The LQ or QR factorization for $\sqrt{P^\star_t}$ can then be skipped if only the smoothed estimate $x^\star_t$ is desired.

V. Filter and Smoother with Square-root Pair

The state-estimates can be co-updated with the square-root covariance if the pre-transform matrices $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ are augmented by a single vector each. A single QR factorization can hence update both the error covariance matrix and state-estimate vector at each recursion step. The quantity that can be updated in such an arrangement is the *square-root pair* defined as $(\sqrt{P}^\top, \bar{b})$ for a given estimate-covariance pair $(\bar{x}, P)$ where $\bar{x} = \sqrt{P}\,\bar{b}$. In particular, $b^\sim_t$ and b°_t, which will be called the *companion vectors*, are defined to encode the state estimates as $x^\sim_t = \sqrt{P^\sim_t}\, b^\sim_t$ and $x^\circ_t = \sqrt{P^\circ_t}\, b^\circ_t$.

Lemma V.1. *Given the system equations (1, 2), the companion vectors $b^\sim_t$ and b°_t can exist if $\bar{x}_1 \in \mathrm{range}(\bar{P}_1)$ and $\bar{q}_t \in \mathrm{range}(Q_t)$ for all t.*

> `Proof:` Existence of $b^\sim_t$ and b°_t is equivalent to $x^\sim_t \in \mathrm{range}(P^\sim_t)$ and $x^\circ_t \in \mathrm{range}(P^\circ_t)$. Suppose $x^\sim_t \in \mathrm{range}(P^\sim_t)$, then from (6, 7) $x^\circ_t \in \mathrm{range}(P^\sim_t)$ and from (6, 14) $\mathrm{range}(P^\circ_t) = \mathrm{range}(P^\sim_t)$. Thus $x^\circ_t \in \mathrm{range}(P^\circ_t)$. Then from (9) $x^\sim_{t+1} \in \mathrm{range}(F_t\sqrt{P^\sim_t}) \cup \mathrm{range}(Q_t)$ and from (10) $\mathrm{range}(P^\sim_{t+1}) = \mathrm{range}(F_t\sqrt{P^\sim_t}) \cup \mathrm{range}(Q_t)$. Thus $x^\sim_{t+1} \in \mathrm{range}(P^\sim_{t+1})$. The lemma must then be true by induction, noting that $x^\sim_1 = \bar{x}_1$, $P^\sim_1 = \bar{P}_1$, and $\bar{x}_1 \in \mathrm{range}(\bar{P}_1)$.

The requirements on $\bar{x}_1$ and $\bar{q}_t$ can be satisfied by the zero-mean condition assumed here (or, alternatively, by assuming that $\bar{P}_1$ and Q_t are all non-singular, a less flexible assumption in practice). Lemma V.1 guarantees the existence of the state square-root pair under the zero-mean condition ($\bar{x}_1 = 0$, $\bar{q}_t = 0$).

A. Data Update of Square-root Pair

The state estimate can be updated jointly with the square-root covariance in the form of the square-root pair. To this end, the joint covariance matrix is augmented by the given innovation $\eta_t \equiv y_t - H_t x_t^{\sim}$ and estimate $x_t^{\sim}$ vectors as

$$\begin{bmatrix} Y & HP^{\sim} & -\eta \\ P^{\sim}H^{\top} & P^{\sim} & x^{\sim} \end{bmatrix} \tag{26}$$

which can be factorized as both $\mathbf{A}_1^{\top}\mathbf{A}_{1x}$ and $\mathbf{B}_1^{\top}\mathbf{B}_{1x}$ where

$$\mathbf{A}_{1x} \equiv \begin{bmatrix} \sqrt{R}^{\top} & \mathbf{0} & -\sqrt{R}^{-1}y \\ \sqrt{P^{\sim}}^{\top}H^{\top} & \sqrt{P^{\sim}}^{\top} & b^{\sim} \end{bmatrix}, \quad \mathbf{B}_{1x} \equiv \begin{bmatrix} \sqrt{Y}^{\top} & \sqrt{Y}^{\top}K^{\top} & -\sqrt{Y}^{-1}\eta \\ \mathbf{0} & \sqrt{P^{\circ}}^{\top} & b^{\circ} \end{bmatrix}$$

which are vector augmented versions of the square-root matrices $\mathbf{A}_1$ and $\mathbf{B}_1$ respectively. The prior $(\sqrt{P_t^{\sim}}^{\top}, b_t^{\sim})$ and posterior $(\sqrt{P_t^{\circ}}^{\top}, b_t^{\circ})$ square-root pairs appear in the lower-right blocks of $\mathbf{A}_{1x}$ and $\mathbf{B}_{1x}$, respectively. Since $\mathbf{B}_{1x}$ is upper triangular, the data-update of the square-root pair could be performed by a single QR factorization that transforms $\mathbf{A}_{1x}$ to $\mathbf{B}_{1x}$. A multiplication by the triangular square-root matrix

$$[\, P^{\circ},\, x^{\circ} \,] = \sqrt{P^{\circ}} \cdot \left[\, \sqrt{P^{\circ}}^{\top},\, b^{\circ} \,\right] \tag{27}$$

would retrieve the filtered estimate-covariance pair from the square-root pair. This multiplication step is optional for applications seeking only the smoothed estimates (e.g., [11, 12]).

B. Time Update of Square-root Pair

If the joint covariance matrix in (23) is augmented to its right by its accompanying joint mean vector, the augmented matrix could be factored as $\mathbf{A}_2^{\top}\mathbf{A}_{2x}$ and $\mathbf{B}_2^{\top}\mathbf{B}_{2x}$ where

$$\mathbf{A}_{2x} \equiv \begin{bmatrix} \sqrt{Q}^{\top} & \mathbf{0} & 0 \\ \sqrt{P^{\circ}}^{\top}F^{\top} & \sqrt{P^{\circ}}^{\top} & b^{\circ} \end{bmatrix}, \quad \mathbf{B}_{2x} \equiv \begin{bmatrix} \sqrt{P_{t+1}^{\sim}}^{\top} & \sqrt{P_{t+1}^{\sim}}^{\top}S^{\top} & b_{t+1}^{\sim} \\ \mathbf{0} & \sqrt{\Delta^{\circ}}^{\top} & d^{\circ} \end{bmatrix}$$

and $(\sqrt{\Delta_t^{\circ}}^{\top}, d_t^{\circ})$ is the square-root pair such that $\delta_t^{\circ} = \sqrt{\Delta_t^{\circ}} d_t^{\circ}$ for the boomerang prediction residual estimate-covariance pair $(\delta_t^{\circ}, \Delta_t^{\circ})$ given by (15, 24). As before, the prior $(\sqrt{P_t^{\circ}}^{\top}, b_t^{\circ})$ is used to construct $\mathbf{A}_{2x}$, which can be QR-factorized into the upper triangular $\mathbf{B}_{2x}$, from which the time-updated posterior $(\sqrt{P_{t+1}^{\sim}}^{\top}, b_{t+1}^{\sim})$ is extracted along with the smoother parameters S_t and $(\sqrt{\Delta_t^{\circ}}^{\top}, d_t^{\circ})$ as needed. The RTS smoother gain S_t needs to be computed via back-substitutions

$$S^{\top} = \left(\sqrt{P_{t+1}^{\sim}}\right)^{-1} \cdot \left(\sqrt{P_{t+1}^{\sim}} S^{\top}\right) \tag{28}$$

where the transposed form $S_t^{\top}$ will be seen to fit directly into the square-root smoother. If only the filter results are desired, the second block columns of $\mathbf{A}_{2x}$ and $\mathbf{B}_{2x}$ can be removed to bypass the computation of the smoother parameters.

C. Smoothing

If the covariance matrix from each side of the DMCS covariance recursion (19) is augmented to its right by the corresponding vector from each side of the DMCS estimate recursion (18), the augmented matrices can be factored as $\mathbf{A}_3^\top \mathbf{A}_{3x}$ and $\mathbf{B}_3^\top \mathbf{B}_{3x}$ where

$$\mathbf{A}_{3x} \equiv \begin{bmatrix} \sqrt{P^\star_{t+1}}^\top S^\top & b^\star_{t+1} \\ \sqrt{\Delta^\circ}^\top & d^\circ \end{bmatrix}, \quad \mathbf{B}_{3x} \equiv \begin{bmatrix} \sqrt{P^\star_t}^\top & b^\star_t \\ \mathbf{0} & 0 \end{bmatrix}$$

which are, as before, related via a QR factorization and used for recursion of the smoothed square-root pair. The initial square-root pair $(\sqrt{P^\star_T}^\top, b^\star_T)$ can directly be extracted from $\mathbf{B}_{1x}$ as $(\sqrt{P^\circ_T}^\top, b^\circ_T)$. Again, a triangular matrix multiplication

$$[\, P^\star, x^\star \,] = \sqrt{P^\star} \cdot \left[\, \sqrt{P^\star}^\top, b^\star \,\right] \tag{29}$$

is needed to retrieve the smoothed estimate-covariance pair from the square-root pair.

VI. Square Root Information Filter

The *square-root information filter* (SRIF) [16] is a different form of square-root Kalman filter. The recursions of SRIF are based not on the covariance matrix but on its inverse, called the *information matrix*. Consequently, SRIF does not require an initial condition for its recursions, a potential advantage in applications where the initial estimate-covariance pair is not available and is arbitrarily created for the sake of filtering [11, 12]. On the other hand, SRIF requires all system covariance matrices including Q_t and R_t to be invertible, potentially restricting its applicability.

The native form for stochastic dynamics in SRIF is the backward difference equation

$$B_t x_{t+1} \quad = \quad x_t + C_t w_t \tag{30}$$

which is typically derived from the original dynamics (1) by assuming that F_t is invertible [16] so that $B_t \equiv F_t^{-1}$ for all t. No such assumption is necessary, however, when the stochastic model is expressed directly in terms of the backward state transition matrix B_t. The noise modulator matrix C_t is useful for maintaining invertibility of Q_t when $C_t w_t$ has a singular covariance matrix.

Instead of the square-root pair, SRIF iterates the square-root information (SRI) pair consisting of an upper-triangular SRI matrix U and SRI vector b. For a random vector $x \sim (\,\bar{x}, P\,)$, the SRI pair defines its mean and covariance implicitly as $U\bar{x} = b$ and $UPU^\top = I$, from which the mean-covariance pair can be recovered as $P = (U^\top U)^{-1}$ and $\bar{x} = U^{-1} b = \sqrt{P}\, b$ if U is invertible. Such an implicit representation permits the filter recursion to continue even in situations where the information matrix $U^\top U$ is singular (rank deficient) so that P cannot exist numerically. While such situations occur often with incomplete observations (in absence of the initial condition $\bar{P}_1$), the singularity could eventually be resolved through integration of more observations if the filter recursion is allowed to proceed. Note that the SRI vector b is identical to the companion vector from the square-root pair representation when both $\sqrt{P}$ and U exist. By Lemma V.1, the zero-mean condition ($\bar{x}_1 = 0$, $\bar{q}_t = 0$) would guarantee existence of the SRI pair and hence the feasibility of using SRIF. The SRI pair constrains x through the observation equation

$$b \quad = \quad Ux + \nu, \quad \nu \sim (\ 0,\ I\ )$$

associated with the normal equation $U^\top U\bar{x} = U^\top b$ that specifies the maximum likelihood estimate $\bar{x}$ implicitly (Appendix II.B).

A. SRIF Data Update

In the data update step of the SRIF algorithm, the given SRI pair $(U_t^\sim,\ b_t^\sim)$ is updated with the new observation (2). A joint observation equation for x_t is formed by combining the observation equation implied by $(U_t^\sim,\ b_t^\sim)$ with (2):

$$\begin{bmatrix} b^\sim \\ y \end{bmatrix} \quad = \quad \begin{bmatrix} U^\sim \\ H \end{bmatrix} x + \begin{bmatrix} \nu \\ v \end{bmatrix} \tag{31}$$

where $\nu_t \sim (\ 0,\ I\ )$. The desired update x_t° is the maximum likelihood estimate based on the joint observation (31), which constrains the estimate through the normal equation

$$(U^{\sim\top} U^\sim + H^\top R^{\text{-}1} H)x^\circ = U^{\sim\top} b^\sim + H^\top R^{\text{-}1} y. \tag{32}$$

On the other hand, the updated SRI pair $(U_t^\circ,\ b_t^\circ)$ implies the normal equation

$$U^{\circ\top} U^\circ x^\circ = U^{\circ\top} b^\circ$$

which is required to be identical to the derived normal equation (32), leading to the equation in a matrix-vector aggregated form

$$\left[\ (U^{\sim\top} U^\sim + H^\top R^{\text{-}1} H),\ (U^{\sim\top} b^\sim + H^\top R^{\text{-}1} y)\ \right] \quad = \quad \left[\ (U^{\circ\top} U^\circ),\ (U^{\circ\top} b^\circ)\ \right]$$

whose two sides can be factored as $\mathbf{A}_4^\top \mathbf{A}_{4x} = \mathbf{B}_4^\top \mathbf{B}_{4x}$ where

$$\mathbf{A}_{4x} \equiv \begin{bmatrix} U^\sim & b^\sim \\ \sqrt{R}^{-1} H & \sqrt{R}^{-1} y \end{bmatrix}, \quad \mathbf{B}_{4x} \equiv \begin{bmatrix} U^\circ & b^\circ \\ \mathbf{0} & 0 \end{bmatrix},$$

and $\mathbf{A}_4$ and $\mathbf{B}_4$ are given by the first block columns of $\mathbf{A}_{4x}$ and $\mathbf{B}_{4x}$. Since $\mathbf{B}_{4x}$ is upper triangular, $(U_t^\sim, b_t^\sim)$ can be updated to (U_t°, b_t°) by transforming $\mathbf{A}_{4x}$ into $\mathbf{B}_{4x}$ via QR factorization after forming $\mathbf{A}_{4x}$ with the given parameters.

B. SRIF Time Update

The backward form (30) of the dynamics allows w_t and x_t to be expressed as linear combinations of w_t and x_{t+1} as

$$\begin{bmatrix} I & \mathbf{0} \\ -C & B \end{bmatrix} \begin{bmatrix} w_t \\ x_{t+1} \end{bmatrix} \quad = \quad \begin{bmatrix} w_t \\ x_t \end{bmatrix}$$

which in turn allows forming a joint observation equation for w_t and x_{t+1} using the known SRI pairs for w_t and x_t as

$$\begin{bmatrix} 0 \\ b^\circ \end{bmatrix} = \begin{bmatrix} \sqrt{Q}^{-1} & \mathbf{0} \\ \mathbf{0} & U^\circ \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -C & B \end{bmatrix} \begin{bmatrix} w_t \\ x_{t+1} \end{bmatrix} + \begin{bmatrix} \nu^w \\ \nu^x \end{bmatrix} \tag{33}$$

where $\nu_t^w \sim (0, I)$ and $\nu_t^x \sim (0, I)$ are mutually independent. The normal equation for (33) is

$$\begin{bmatrix} I & -C^\top \\ \mathbf{0} & B^\top \end{bmatrix} \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & U^{\circ\top} U^\circ \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -C & B \end{bmatrix} \begin{bmatrix} w_t \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} I & -C^\top \\ \mathbf{0} & B^\top \end{bmatrix} \begin{bmatrix} 0 \\ U^{\circ\top} b^\circ \end{bmatrix}$$

whose coefficients from both sides are used to form the matrix-vector aggregate:

$$\begin{bmatrix} Q^{-1} + C^\top U^{\circ\top} U^\circ C & -C^\top U^{\circ\top} U^\circ B & -C^\top U^{\circ\top} b^\circ \\ -B^\top U^{\circ\top} U^\circ C & B^\top U^{\circ\top} U^\circ B & B^\top U^{\circ\top} b^\circ \end{bmatrix} \tag{34}$$

which can be factored in two ways as $\mathbf{A}_5^\top \mathbf{A}_{5x}$ and $\mathbf{B}_5^\top \mathbf{B}_{5x}$ where

$$\mathbf{A}_{5x} \equiv \begin{bmatrix} \sqrt{Q}^{-1} & \mathbf{0} & 0 \\ -U^\circ C & U^\circ B & b^\circ \end{bmatrix}, \quad \mathbf{B}_{5x} \equiv \begin{bmatrix} \Gamma & \Xi & \gamma \\ \mathbf{0} & U_{t+1}^\sim & b_{t+1}^\sim \end{bmatrix}$$

and $\mathbf{A}_5$ and $\mathbf{B}_5$ are given by the first two block columns of $\mathbf{A}_{5x}$ and $\mathbf{B}_{5x}$. Due again to the upper triangular structure of $\mathbf{B}_{5x}$, the QR factorization can transform $\mathbf{A}_{5x}$ into $\mathbf{B}_{5x}$ from which the time-updated SRI pair $(U_{t+1}^\sim, b_{t+1}^\sim)$ can be extracted.

The familiar smoother recursion parameters would emerge if the top row of $\mathbf{B}_{5x}$, consisting of Γ_t, Ξ_t, and γ_t, is left-multipled by $\sqrt{\Delta_t^\circ}$ as

$$\sqrt{\Delta^\circ} \cdot [\,\Gamma,\ \Xi,\ \gamma\,] \ = \ [\,C,\ (S - B),\ -\delta^\circ\,]. \tag{35}$$

In particular, $\sqrt{\Delta_t^\circ}$ itself can be computed via inversion of the upper-triangular Γ_t as

$$\sqrt{\Delta^\circ} \ = \ C\,\Gamma^{-1} \tag{36}$$

and the remaining parameters required for the square-root covariance smoother (Section IV.C) can be derived as

$$S \ = \ \sqrt{\Delta^\circ}\,\Xi + B \tag{37}$$

$$\delta^\circ \ = \ -\sqrt{\Delta^\circ}\,\gamma \tag{38}$$

which are stored for the smoother recursions. The SRIF parameters $(\Gamma_t, \Xi_t, \gamma_t)$ can be ignored if smoothed estimates are not desired.

The remainder of this section focuses on derivation of the three components of the matrix equation (35). For simplicity, the filtered information matrix $L_t^\circ \equiv U_t^{\circ\top} U_t^\circ$ is assumed to be invertible so that $P_t^\circ = L_t^{\circ\,-1}$ exists. (Another derivation is necessary to include a singular L_t°.) The starting point is the backward state transition equation (30), which is rewritten using $F_t = B_t^{-1}$ as $x_{t+1} = F_t x_t + F_t C_t w_t$ so that the forecast covariance is given as

$P^{\sim}_{t+1} = F_t(P^{\circ}_t + C_tQ_tC_t^\top)F_t^\top$. Since $(P^{\sim}_{t+1})^{-1} = F_t^{-\top}(P^{\circ}_t + C_tQ_tC_t^\top)^{-1}B_t$, the RTS smoother gain formula (11) can be written as

$$\begin{aligned} S &= P^{\circ}(P^{\circ} + CQC^\top)^{-1}B \\ &= [I - CQC^\top(P^{\circ} + CQC^\top)^{-1}]B \\ &= [I - C(Q^{-1} + CL^{\circ}C^\top)^{-1}C^\top L^{\circ}]B \end{aligned} \tag{39}$$

where the last equality is due to formulas from the Block Matrix Inversion Lemma. On the other hand, the filtered covariance for the boomerang prediction residual in (24) can be written as $\Delta^{\circ} = P^{\circ} - SFP^{\circ} = (I - SF)P^{\circ}$, and substituting (39) into the last expression would lead to

$$\begin{aligned} \Delta^{\circ} &= [I - I + C(Q^{-1} + CL^{\circ}C^\top)^{-1}C^\top L^{\circ}]P^{\circ} \\ &= C(Q^{-1} + CL^{\circ}C^\top)^{-1}C^\top. \end{aligned} \tag{40}$$

Defining $Q^{\circ} \equiv (Q^{-1} + CL^{\circ}C^\top)^{-1}$ and $\sqrt{Q^{\circ}}\sqrt{Q^{\circ}}^\top = Q^{\circ}$ would then lead to

$$\Delta^{\circ} = CQ^{\circ}C^\top, \quad \sqrt{\Delta^{\circ}} = C\sqrt{Q^{\circ}}. \tag{41}$$

Also, substituting (40) back into (39) would reveal

$$S = B - \Delta^{\circ}L^{\circ}B. \tag{42}$$

Now, each block of $\mathbf{B}_5^\top \mathbf{B}_{5x}$ is matched with the corresponding block of (34). Matching the top-left blocks yields $\Gamma^\top\Gamma = Q^{\circ\,-1}$ leading to $\Gamma^{-\top} = \sqrt{Q^{\circ}}^\top$ and $\Gamma^{-1} = \sqrt{Q^{\circ}}$, the latter of which is substituted into (41) to prove that $\sqrt{\Delta^{\circ}}\Gamma = C$, the first of three objectives here. Matching the top-middle blocks yields $\Gamma^\top\Xi = -C^\top L^{\circ}B$, which leads to $\sqrt{\Delta^{\circ}}\Xi = -\sqrt{\Delta^{\circ}}\Gamma^{-\top}C^\top L^{\circ}B = -\Delta^{\circ}L^{\circ}B = S - B$ from (42), completing the second objective. Matching the top-right vectors yields $\Gamma^\top\gamma = -C^\top U^{\circ\top}b^{\circ}$, leading to $\sqrt{\Delta^{\circ}}\gamma = -\sqrt{\Delta^{\circ}}\Gamma^{-\top}C^\top U^{\circ\top}b^{\circ} = -\Delta^{\circ}U^{\circ\top}b^{\circ} = -\delta^{\circ}$ since from (15) and (42) $\delta^{\circ} = (I - SF)x^{\circ} = [I - (I - \Delta^{\circ}L^{\circ})BF]x^{\circ} = \Delta^{\circ}L^{\circ}x^{\circ} = \Delta^{\circ}U^{\circ\top}b^{\circ}$. This completes the third and final objective.[4]

C. Square-root DMCS

The DMCS algorithm was originally introduced to accompany SRIF [16]. The parameters S_t and δ°_t generated during the SRIF time-update via (37, 38) would enable the smoother estimate recursion (18). The parameters S_t and $\sqrt{\Delta^{\circ}_t}$, by (36), would prepare $\mathbf{A}_3^\top$ for the square-root smoother error covariance recursion. To initialize the recursions as $x^{\star}_T = U^{\circ\,-1}_T b^{\circ}_T$ and $\sqrt{P^{\star}_T} = U^{\circ\,-1}_T$, the upper-triangular U°_T must be invertible.

[4]To match the bottom-middle and bottom-right blocks, the Schur complement and associated vector transform formulas yield $U^{\sim\top}_{+}U^{\sim}_{+} = B^\top L^{\circ}B - B^\top L^{\circ}\Delta^{\circ}L^{\circ}B$ and $U^{\sim\top}_{+}b^{\sim}_{+} = B^\top U^{\circ\top}b^{\circ} - B^\top L^{\circ}\Delta^{\circ}U^{\circ\top}b^{\circ}$, which are then complemented by $\Xi^\top\Xi = B^\top L^{\circ}\Delta^{\circ}L^{\circ}B$ and $\Xi^\top\gamma = B^\top L^{\circ}\Delta^{\circ}U^{\circ\top}b^{\circ}$ so that $U^{\sim\top}_{+}U^{\sim}_{+} = B^\top L^{\circ}B - \Xi^\top\Xi$ and $U^{\sim\top}_{+}b^{\sim}_{+} = B^\top U^{\circ\top}b^{\circ} - \Xi^\top\gamma$.

VII. Combined Data and Time Update

The filtered square-root covariance can be updated directly from $\sqrt{P_t^\circ}$ to $\sqrt{P_{t+1}^\circ}$, along with the companion vector from b_t° to b_{t+1}° as desired, by combining the time and data update procedures in tandem. For such purposes, consider the joint distribution

$$\begin{bmatrix} -\eta_+ \\ x_+ \\ x \end{bmatrix} \sim \left(\begin{bmatrix} H_+ x_+^\sim - y_+ \\ x_+^\sim \\ x^\circ \end{bmatrix}, \begin{bmatrix} Y_+ & H_+ P_+^\sim & H_+ F P^\circ \\ P_+^\sim H_+^\top & P_+^\sim & F P^\circ \\ P^\circ F^\top H_+^\top & P^\circ F^\top & P^\circ \end{bmatrix} \right)$$

where the subscript "$+$" denotes "$t+1$" for brevity. The joint covariance matrix has the following two square-root factors

$$\mathbf{A}_{21}^\top \equiv \begin{bmatrix} \sqrt{R_+} & H_+\sqrt{Q} & H_+ F\sqrt{P^\circ} \\ \mathbf{0} & \sqrt{Q} & F\sqrt{P^\circ} \\ \mathbf{0} & \mathbf{0} & \sqrt{P^\circ} \end{bmatrix}, \quad \mathbf{B}_{21}^\top \equiv \begin{bmatrix} \sqrt{Y_+} & \mathbf{0} & \mathbf{0} \\ K_+\sqrt{Y_+} & \sqrt{P_+^\circ} & \mathbf{0} \\ SK_+\sqrt{Y_+} & S\sqrt{P_+^\circ} & \sqrt{\Delta^\circ} \end{bmatrix}$$

which can be used for updating the square-root covariance matrix. Also, the covariance-mean aggregate from the joint distribution can be factored as $\mathbf{A}_{21}^\top \mathbf{A}_{21x}$ and $\mathbf{B}_{21}^\top \mathbf{B}_{21x}$ where

$$\mathbf{A}_{21x} \equiv \begin{bmatrix} \mathbf{A}_{21} & \begin{matrix} -\sqrt{R_+}^{-1} y_+ \\ 0 \\ b^\circ \end{matrix} \end{bmatrix}, \quad \mathbf{B}_{21x} \equiv \begin{bmatrix} \mathbf{B}_{21} & \begin{matrix} -\sqrt{Y_+}^{-1} \eta_+ \\ b_+^\circ \\ d^\circ \end{matrix} \end{bmatrix},$$

which can be used for updating the square-root pair. The last rows of $\mathbf{A}_{21}^\top$ and $\mathbf{B}_{21}^\top$ can be removed if the smoothed estimates are not needed.

The SRI-pair can similarly be updated directly from (U_t°, b_t°) to $(U_{t+1}^\circ, b_{t+1}^\circ)$ using the observation equation that combines (33) with (31):

$$\begin{bmatrix} 0 \\ b^\circ \\ y_+ \end{bmatrix} = \begin{bmatrix} -I & \mathbf{0} \\ -U^\circ C & U^\circ B \\ \mathbf{0} & H_+ \end{bmatrix} \begin{bmatrix} w \\ x_+ \end{bmatrix} + \begin{bmatrix} w \\ \nu^\circ \\ v_+ \end{bmatrix}$$

leading to the normal equation:

$$\begin{bmatrix} Q^{-1} + C^\top U^{\circ\top} U^\circ C & -C^\top U^{\circ\top} U^\circ B \\ -B^\top U^{\circ\top} U^\circ C & B^\top U^{\circ\top} U^\circ B + H_+^\top R_+^{-1} H_+ \end{bmatrix} \begin{bmatrix} w^\circ \\ x_+^\circ \end{bmatrix} = \begin{bmatrix} -C^\top U^{\circ\top} b^\circ \\ B^\top U^{\circ\top} b^\circ + H_+^\top R_+^{-1} y_+ \end{bmatrix}$$

whose coefficients can form the matrix-vector aggregate that can be factored as $\mathbf{A}_{54}^\top \mathbf{A}_{54x}$ and $\mathbf{B}_{54}^\top \mathbf{B}_{54x}$ where

$$\mathbf{A}_{54x} \equiv \begin{bmatrix} \sqrt{Q}^{-1} & \mathbf{0} & 0 \\ -U^\circ C & U^\circ B & b^\circ \\ \mathbf{0} & \sqrt{R_+}^{-1} H_+ & \sqrt{R_+}^{-1} y_+ \end{bmatrix}, \quad \mathbf{B}_{54x} \equiv \begin{bmatrix} \Gamma & \Xi & \gamma \\ \mathbf{0} & U_+^\circ & b_+^\circ \\ \mathbf{0} & \mathbf{0} & 0 \end{bmatrix},$$

and $\mathbf{A}_{54}$ and $\mathbf{B}_{54}$ are the first two block-columns of $\mathbf{A}_{54x}$ and $\mathbf{B}_{54x}$, respectively. The transformation of $\mathbf{A}_{54x}$ to the upper-triangular $\mathbf{B}_{54x}$ via QR-factorization would update the SRI pair and yield the triplet $(\Gamma_t, \Xi_t, \gamma_t)$ needed to generate the smoother parameters $(\sqrt{\Delta_t^\circ}, S_t, \delta_t^\circ)$.

VIII. Summary and Conclusion

A self-contained derivation of square-root versions of the discrete Kalman filter and RTS smoother is presented. Introduction of the "boomerang prediction residual" process along with Lemma III.1 has clarified the connections between the filter and smoother equations, leading to a stabilized RTS smoother and simplified square-root formulas. The boomerang prediction residual process is also prominent in linking the square-root information filter (SRIF) with Dyer-McReynolds Covariance Smoother (DMCS), a lesser known but square-root compatible version of the RTS smoother. Table 4 summarizes these square-root filter and smoother algorithms.

Besides the extra steps needed to convert between the square-root forms and estimate-covariance pair, the square-root formulas for the Kalman filter and RTS smoother seem to offer only advantages over the standard formulas in regards to computational complexity, practical realizations, and numerical stability. The key advantages are numerical guarantee for positive semi-definite covariance matrix, absence of general matrix inverse (except for back-substitution of triangular matrix), and reduced number of matrix multiplications. The requirement for large memory to stage the aggregate matrix for QR factorization is no longer a limiting factor in contemporary computers for most applications.

QR factorization almost single-handedly handles the filter and smoother recursions. Performance of QR factorization can vary among different algorithmic realizations with respect to numerical reliability in near-singular cases, for example, by providing high-speed execution at the cost of reduced checks on singularity. The readers should note that both variations of the matrix square-roots are used in the presentation: one variant used exclusively for the covariance matrices ($\sqrt{P}\sqrt{P}^{\top}$) for consistency with their definitions and the other used for all other matrices ($U^{\top}U$, $\mathbf{A}^{\top}\mathbf{A}$, etc) for compatibility with QR factorization. The square roots of the covariance and information matrices are still the inverse of each other as $\sqrt{P} = U^{-1}$.

Existence of the Kalman filter and RTS smoother solutions depends on invertibility of Y in (6) and $P^{\sim}$ in (11), respectively. Quality of the filter and smoother estimates can be monitored during recursions by checking the numerical condition of these matrices. The square-roots of these matrices ($\sqrt{Y}$, $\sqrt{P^{\sim}}$) are accessible for monitoring during the filter recursion of the square-root covariance matrix. The filter recursion for the square-root pair has an additional requirement for $\sqrt{R_t}$ to be invertible for all t to form $\mathbf{A}_{1x}$ or $\mathbf{A}_{21x}$, although most applications in practice satisfy this requirement by having invertible observation error covariance matrices. SRIF also requires invertibility of $\sqrt{R_t}$ for all t and of U_t for t when the filter estimates are desired. If only the smoother estimates are needed (e.g., [11, 12]), only U_T needs to be invertible. SRIF has a unique flexibility among the presented filters in that the initial condition (estimate-covariance pair) is optional.

Table 4. Basic square-root algorithms for the standand Kalman filter and RTS smoother in Table 3.

Square-root Covariance	Square-root Pair	SRI Pair (SRIF/DMCS)
— Filter Initialization		
$\sqrt{P_1^{\sim}} = \sqrt{P_1},\ x_1^{\sim} = 0$	$\left(\sqrt{P_1^{\sim}}^{\top},\ b_1^{\sim}\right) = (\sqrt{P_1}^{\top},\ 0)$	$(U_1^{\sim},\ b_1^{\sim}) = (\sqrt{P_1}^{-1},\ 0)$
— Filter Recursion $(t = 1, 2, \ldots, T)$		
$\mathbf{A}_1^{\top} \longleftarrow \sqrt{P_t^{\sim}}$	$\mathbf{A}_{1x} \longleftarrow \left(\sqrt{P_t^{\sim}}^{\top},\ b_t^{\sim}\right)$	$\mathbf{A}_{4x} \longleftarrow (U_t^{\sim},\ b_t^{\sim})$
$\mathbf{B}_1^{\top}$ = LQ($\mathbf{A}_1^{\top}$)	$\mathbf{B}_{1x}$ = QR($\mathbf{A}_{1x}$)	$\mathbf{B}_{4x}$ = QR($\mathbf{A}_{4x}$)
$\sqrt{P_t^{\circ}} \Leftarrow \mathbf{B}_1^{\top}$	$\left(\sqrt{P_t^{\circ}}^{\top},\ b_t^{\circ}\right) \Leftarrow \mathbf{B}_{1x}$	$(U_t^{\circ},\ b_t^{\circ}) \Leftarrow \mathbf{B}_{4x}$
$K\sqrt{Y},\ \sqrt{Y} \Leftarrow \mathbf{B}_1^{\top}$	† $(P_t^{\circ},\ x_t^{\circ}) = \sqrt{P_t^{\circ}} \cdot \left(\sqrt{P_t^{\circ}}^{\top},\ b_t^{\circ}\right)$	† $(\sqrt{P_t^{\circ}},\ x_t^{\circ}) = {U_t^{\circ}}^{-1} \cdot (I, b_t^{\circ})$
$x_t^{\circ} = x_t^{\sim} + K\sqrt{Y} \cdot \sqrt{Y}^{-1}(y_t - H_t x_t^{\sim})$		
• if $t = T$, end recursion.	• if $t = T$, end recursion.	• if $t = T$, end recursion.
$\mathbf{A}_2^{\top} \longleftarrow \sqrt{P_t^{\circ}}$	$\mathbf{A}_{2x} \longleftarrow \left(\sqrt{P_t^{\circ}}^{\top},\ b_t^{\circ}\right)$	$\mathbf{A}_{5x} \longleftarrow (U_t^{\circ},\ b_t^{\circ})$
$\mathbf{B}_2^{\top}$ = LQ($\mathbf{A}_2^{\top}$)	$\mathbf{B}_{2x}$ = QR($\mathbf{A}_{2x}$)	$\mathbf{B}_{5x}$ = QR($\mathbf{A}_{5x}$)
$\sqrt{P_{t+1}^{\sim}} \Leftarrow \mathbf{B}_2^{\top}$	$\left(\sqrt{P_{t+1}^{\sim}}^{\top},\ b_{t+1}^{\sim}\right) \Leftarrow \mathbf{B}_{2x}$	$(U_{t+1}^{\sim},\ b_{t+1}^{\sim}) \Leftarrow \mathbf{B}_{5x}$
$x_{t+1}^{\sim} = F_t x_t^{\circ}$		
‡ $\sqrt{P_{t+1}^{\sim}},\ S_t\sqrt{P_{t+1}^{\sim}} \Leftarrow \mathbf{B}_2^{\top}$	‡ $\sqrt{P_{t+1}^{\sim}},\ \sqrt{P_{t+1}^{\sim}} S_t^{\top} \Leftarrow \mathbf{B}_{2x}$	‡ $\Gamma_t,\ \Xi_t,\ \gamma_t \Leftarrow \mathbf{B}_{5x}$
‡ $S_t = S_t\sqrt{P_{t+1}^{\sim}}\sqrt{P_{t+1}^{\sim}}^{-1}$	‡ $S_t^{\top} = \sqrt{P_{t+1}^{\sim}}^{-1}\sqrt{P_{t+1}^{\sim}} S_t^{\top}$	‡ $\sqrt{\Delta_t^{\circ}} = C_t \Gamma_t^{-1}$
‡ $\sqrt{\Delta_t^{\circ}} \Leftarrow \mathbf{B}_2^{\top}$	‡ $\left(\sqrt{\Delta_t^{\circ}}^{\top},\ d_t^{\circ}\right) \Leftarrow \mathbf{B}_{2x}$	‡ $S_t = \sqrt{\Delta_t^{\circ}}\ \Xi_t + B_t$
‡ $\delta_t^{\circ} = x_t^{\circ} - S_t x_{t+1}^{\sim}$	‡ store $S^{\top}$ and $\left(\sqrt{\Delta_t^{\circ}}^{\top},\ d_t^{\circ}\right)$	‡ $\delta_t^{\circ} = -\sqrt{\Delta_t^{\circ}}\ \gamma_t$
‡ store S_t, $\sqrt{\Delta_t^{\circ}}$, and δ_t°		‡ store S_t, $\sqrt{\Delta_t^{\circ}}$, and δ_t°
— Smoother Initialization		
$\sqrt{P_T^{\star}} = \sqrt{P_T^{\circ}},\ x_T^{\star} = x_T^{\circ}$	$\left(\sqrt{P_T^{\star}}^{\top},\ b_T^{\star}\right) = \left(\sqrt{P_T^{\circ}}^{\top},\ b_T^{\circ}\right)$	$(\sqrt{P_T^{\star}},\ x_T^{\star}) = {U_T^{\circ}}^{-1} \cdot (I, b_T^{\circ})$
— Smoother Recursion $(t = T-1,\ T-2,\ \ldots, 1)$		
$\mathbf{A}_3^{\top} \longleftarrow \sqrt{P_{t+1}^{\star}}$	$\mathbf{A}_{3x} \longleftarrow \left(\sqrt{P_{t+1}^{\star}}^{\top},\ b_{t+1}^{\star}\right)$	$\mathbf{A}_3^{\top} \longleftarrow \sqrt{P_{t+1}^{\star}}$
$\mathbf{B}_3^{\top}$ = LQ($\mathbf{A}_3^{\top}$)	$\mathbf{B}_{3x}$ = QR($\mathbf{A}_{3x}$)	$\mathbf{B}_3^{\top}$ = LQ($\mathbf{A}_3^{\top}$)
$\sqrt{P_t^{\star}} \Leftarrow \mathbf{B}_3^{\top}$	$\left(\sqrt{P_t^{\star}}^{\top},\ b_t^{\star}\right) \Leftarrow \mathbf{B}_{3x}$	$\sqrt{P_t^{\star}} \Leftarrow \mathbf{B}_3^{\top}$
$P_t^{\star} = \sqrt{P_t^{\star}}\sqrt{P_t^{\star}}^{\top}$	$(P_t^{\star},\ x_t^{\star}) = \sqrt{P^{\star}} \cdot \left(\sqrt{P_t^{\star}}^{\top},\ b_t^{\star}\right)$	$P_t^{\star} = \sqrt{P_t^{\star}}\sqrt{P_t^{\star}}^{\top}$
$x_t^{\star} = S_t x_{t+1}^{\star} + \delta_t^{\circ}$		$x_t^{\star} = S_t x_{t+1}^{\star} + \delta_t^{\circ}$

† optional steps for generating the filter estimates.

‡ optional steps for preparing the smoother parameters.

"$\mathbf{X} \longleftarrow Y$" denotes to "form $\mathbf{X}$ including Y as a block".

"$X \Leftarrow \mathbf{Y}$" denotes to "extract X from a block in $\mathbf{Y}$".

"QR" is the unitary transformation using QR factorization.

"LQ" is the unitary transformation using the transposed version of QR factorization.

If smoother is not used, ignore the second columns of $\mathbf{A}_2$, $\mathbf{B}_2$, $\mathbf{A}_{2x}$, $\mathbf{B}_{2x}$ for efficiency.

Any initial condition for SRIF should be converted to additional observations at $t = 1$.

Acknowledgments

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018F0591). The author thanks Andre Tkacenko for valuable feedback on this manuscript.

References

[1] M. S. Grewal and A. P. Andrews, "Applications of Kalman filtering in aerospace 1960 to the present," *IEEE Control Systems Magazine*, vol. 30, no. 3, pp. 69–78, 2010.

[2] P. G. Kaminski, A. E. Bryson, and S. F. Schmidt, "Discrete square root filtering: A survey of current techniques," *IEEE Transaction on Automatic Control*, vol. 16, pp. 727–736, 1971.

[3] M. Morf and T. Kailath, "Square-root algorithms for least-squares estimation," *IEEE Transactions on Automatic Control*, vol. 20, no. 4, pp. 487–497, 1975.

[4] C. Thornton and G. Bierman, "Filtering and error analysis via the UDU^T covariance factorization," *IEEE Transactions on Automatic Control*, vol. 23, no. 5, pp. 901–907, 1978.

[5] F. Gaston and G. Irwin, "Systolic approach to square root information Kalman filtering," *International Journal of Control*, vol. 50, no. 1, pp. 225–248, 1989.

[6] T. M. Chin, W. C. Karl, and A. S. Willsky, "A distributed and iterative method for square root filtering in space-time estimation," *Automatica*, vol. 31, no. 1, pp. 67–82, 1995.

[7] Y. Oshman and I. Y. Bar-Itzhack, "Square root filtering via covariance and information eigenfactors," *Automatica*, vol. 22, no. 5, pp. 599–604, 1986.

[8] D. Fuertes Roncero, "A study of QR decomposition and Kalman filter implementations," Master's thesis, Universitat Politècnica de Catalunya, 2014.

[9] K. Tracy, "A square-root Kalman filter using only QR decompositions," arXiv preprint arXiv:2208.06452, 2022.

[10] H. E. Rauch, C. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[11] X. Wu, C. Abbondanza, Z. Altamimi, T. M. Chin, X. Collilieux, R. S. Gross, M. B. Heflin, Y. Jiang, and J. W. Parker, "KALREF-a Kalman filter and time series approach to the International Terrestrial Reference Frame realization," *Journal of Geophysical Research: Solid Earth*, vol. 120, no. 5, pp. 3775–3802, 2015.

[12] C. Abbondanza, T. M. Chin, R. S. Gross, M. B. Heflin, J. W. Parker, B. S. Soja, T. van Dam, and X. Wu, "JTRF2014, the JPL Kalman filter and smoother realization of the

International Terrestrial Reference System," *Journal of Geophysical Research: Solid Earth*, vol. 122, no. 10, pp. 8474–8510, 2017.

[13] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice-Hall, 1979.

[14] P. Park and T. Kailath, "New square-root algorithms for Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 40, no. 5, pp. 895–899, 1995.

[15] ——, "New square-root smoothing algorithms," *IEEE Transactions on Automatic Control*, vol. 41, no. 5, pp. 727–732, 1996.

[16] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*. Academic Press, 1977.

[17] A. Gelb *et al.*, *Applied Optimal Estimation*. MIT press, 1974.

[18] F. L. Lewis, *Optimal Estimation*. Wiley, 1986.

[19] H. W. Sorenson, "Least-squares estimation: from Gauss to Kalman," *IEEE Spectrum*, vol. 7, no. 7, pp. 63–68, 1970.

[20] A. Björck and S. Hammarling, "A Schur method for the square root of a matrix," *Linear Algebra and its Applications*, vol. 52/53, pp. 127–140, 1983.

[21] G. H. Golub and C. F. van Loan, *Matrix Computations*. The Johns Hopkins University Press, 1989.

[22] D. Fraser and J. Potter, "The optimum linear smoother as a combination of two optimum linear filters," *IEEE Transaction on Automatic Control*, vol. 14, pp. 387–390, 1969.

[23] D. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Automatica*, vol. 4, no. 2, pp. 73–92, 1966.

APPENDICES

I. Notes from Matrix Algebra

A. Matrix Square-roots and QR Factorization

Any symmetric matrix M has a square-root matrix A such that $A^\top A = M$ and A is not unique. The upper triangular matrix B produced by the QR factorization $A = \Theta B$, where Θ is unitary, is also a square-root of M since $M = A^\top A = B^\top \Theta^\top \Theta B = B^\top B$ due to the defining characteristic of an unitary matrix that $\Theta^\top \Theta = I$. Thus, any square-root of a symmetric matrix can be converted into an upper-triangular square-root of the same symmetric matrix using the QR factorization.

The so-called LQ factorization is a transposed version of the QR factorization: $A^\top = B^\top \Theta^\top$ where $B^\top$ is lower triangular.

B. Schur Complement and Block Matrix Square-root

If a square matrix M has block partitions

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

such that M_{11} is square and invertible, then the matrix $N_{22} = M_{22} - M_{21} M_{11}^{-1} M_{12}$ is called the *Schur complement* of M_{11}. A *block LU decomposition* of M exposes N_{22} [21] as

$$M = \begin{bmatrix} L_{11} & \mathbf{0} \\ L_{21} & I \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & N_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ \mathbf{0} & I \end{bmatrix}$$

noting that $L_{21} U_{12} = L_{21} U_{11} U_{11}^{-1} L_{11}^{-1} L_{11} U_{12} = M_{21} (L_{11} U_{11})^{-1} M_{12} = M_{21} M_{11}^{-1} M_{12}$. If in addition M is symmetric, the block triangular square-root factors

$$M = \begin{bmatrix} \sqrt{M}_{11}^\top & \mathbf{0} \\ U_{12}^\top & \sqrt{N}_{22}^\top \end{bmatrix} \begin{bmatrix} \sqrt{M}_{11} & U_{12} \\ \mathbf{0} & \sqrt{N}_{22} \end{bmatrix}$$

would expose the square-root $\sqrt{N}_{22}$ of the Schur complement such that $\sqrt{N}_{22}^\top \sqrt{N}_{22} = N_{22}$.

II. Notes from Estimation Theory

A. Linear Least-squares Estimation and Conditional Gaussian Distribution

If two random vectors x and y have a joint mean-covariance pair

$$\begin{bmatrix} y \\ x \end{bmatrix} \sim \left(\begin{bmatrix} \bar{y} \\ \bar{x} \end{bmatrix}, \begin{bmatrix} P_{yy} & P_{yx} \\ P_{xy} & P_{xx} \end{bmatrix} \right),$$

then the optimal linear estimate of x by y is $\widehat{x} = Ay + b$ that satisfies $\mathcal{E}[x - \widehat{x}] = 0$ (unbiased estimate) and minimizes $\mathcal{E}[(x - \widehat{x})^\top (x - \widehat{x})]$ (least mean-squares error). The zero bias requirement leads to $\mathcal{E}[x - \widehat{x}] = \bar{x} - A\bar{y} - b = 0$ or $b = \bar{x} - A\bar{y}$, which makes the form of the estimate be

$$\widehat{x} = \bar{x} + A(y - \bar{y}).$$

The associated estimation error becomes $x - \widehat{x} = (x - \bar{x}) + A(y - \bar{y})$ and the estimation error covariance becomes $\widehat{P} \equiv \mathcal{E}[(x - \widehat{x})(x - \widehat{x})^\top] = P_{xx} - P_{xy}A^\top - AP_{yx} + AP_{yy}A^\top$. To determine A, the mean-squares error is minimized with respect to A:

$$\begin{aligned} \frac{d}{dA}\mathcal{E}[(x - \widehat{x})^\top (x - \widehat{x})] &= \frac{d}{dA}\mathrm{trace}(\widehat{P}) \\ &= -P_{xy} - P_{yx}^\top + 2AP_{yy} \\ &= \mathbf{0} \end{aligned}$$

yielding $A = P_{xy}P_{yy}^{-1}$. Thus the optimal linear estimator and the associated estimation error are

$$\widehat{x} = \bar{x} + P_{xy}P_{yy}^{-1}(y - \bar{y}) \quad (43)$$

$$\widehat{P} = P_{xx} - P_{xy}P_{yy}^{-1}P_{yx} \quad (44)$$

indicating that $\widehat{P}$ is the Schur complement of P_{yy}.

If the joint distribution of x and y is Gaussian, $\widehat{x}$ and $\widehat{P}$ can be shown to be the mean and covariance of the conditional Gaussian distribution [13, 18], or

$$x|y \sim \left(\bar{x} + P_{xy}P_{yy}^{-1}(x - \bar{x}),\ P_{xx} - P_{xy}P_{yy}^{-1}P_{yx} \right).$$

If the joint distribution is non-Gaussian, the optimal linear estimator may not be equal to the conditional mean, which is the optimal least-squares estimator among all (linear or non-linear) estimators.

B. Maximum-likelihood Estimation and Normal Equation

The linear least-squares estimation formulas (43, 44) cannot be applied to cases where the distribution (mean and covariance) of the variable x is not known prior to observation. In many of such cases, the distribution function $p(\, y \,|\, x\,)$ of the observation y given x is available (often as Gaussian), and the *maximum likelihood* estimate $\hat{x} = \arg\max_x p(\, y \,|\, x\,)$ is sought. In particular, for the linear Gaussian observation equation

$$y = Hx + v, \quad v \sim (0,\, R),$$

the estimate that minimizes the observation error

$$e(x) \equiv (y - Hx)^\top R^{-1}(y - Hx)$$

is sought. Finding the minimum via $\partial e(\hat{x})/\partial x = 0$ would lead to the so-called *normal equation*

$$(H^\top R^{-1} H)\, \hat{x} = H^\top R^{-1} y \quad (45)$$

that constrains the optimal estimate $\hat{x}$ linearly. Since $H^\top R^{-1} y = H^\top R^{-1} H x + H^\top R^{-1} v$, the estimation error can similarly be given as $L(x - \hat{x}) = -H^\top R^{-1} v$ where $L = H^\top R^{-1} H$ is the *(Fisher) information matrix*. The estimation error covariance $\widehat{P} = \mathcal{E}[(x - \hat{x})(x - \hat{x})^\top]$ would then be constrained as

$$L\widehat{P}L = H^\top R^{-1} \mathcal{E}[vv^\top] R^{-1} H = H^\top R^{-1} H = L. \tag{46}$$

If the information matrix is invertible, we have $\widehat{P} = L^{-1} = (H^\top R^{-1} H)^{-1}$ and $\hat{x} = \widehat{P} H^\top R^{-1} y$. If, on the other hand, L is singular, additional observations are needed for (45, 46) to uniquely specify $\hat{x}$ and $\widehat{P}$.

C. Sequential Solutions for the Fixed-interval Smoothing Problem

The solution $x_t^\star$ for the fixed-interval smoothing problem can be derived by combining the Kalman filter estimates with estimates from another filter that iterates backwards in time [22]. The Markov property of the linear system (1, 2) allows decomposition of the conditional dependence on the entire observation set $Y_{1:T} \equiv \{y_1, y_2, \ldots, y_T\}$ into two sets: the past plus current observations $Y_{1:t}$ and the future observations $Y_{t+1:T}$. In particular, the original maximum-likelihood formulation for the RTS smoother [10] isolates the filtering problem $p(x_t | Y_{1:t})$ from the smoothing problem $p(x_t, x_{t+1}, Y_{1:T})$ as

$$\begin{aligned} p(x_t, x_{t+1}, Y_{1:T}) &= p(Y_{t+1:T} | x_{t+1})\, p(x_{t+1} | x_t)\, p(x_t | Y_{1:t})\, p(Y_{1:t}) \\ &= p(Y_{t+1:T} | x_{t+1})\, p(x_{t+1}, x_t, Y_{1:t}) \end{aligned}$$

where both $x_t^\star$ and $x_{t+1}^\star$ would be the solutions for maximization of $p(x_t, x_{t+1}, Y_{1:T})$. Marginalizing-out x_t from the two sides would focus only on $x_{t+1}^\star$ as

$$p(x_{t+1}, Y_{1:T}) = p(Y_{t+1:T} | x_{t+1})\, p(x_{t+1}, Y_{1:t})$$

which decomposes the smoothing problem into two estimation problems for $x_{t+1}^b \equiv \mathcal{L}[x_{t+1} | Y_{t+1:T}]$ and $x_{t+1}^\sim$. As shown in [23], a weighted average of these two estimates is the smoothed estimate

$$x_{t+1}^\star = W_{t+1}^f\, x_{t+1}^\sim + W_{t+1}^b\, x_{t+1}^b \tag{47}$$

where the weights are given by the normalized information (inverse covariance) matrices associated with the corresponding estimation errors

$$W_{t+1}^f = N_{t+1} (P_{t+1}^\sim)^{-1}, \quad W_{t+1}^b = N_{t+1} (P_{t+1}^b)^{-1}, \quad N_{t+1}^{-1} \equiv (P_{t+1}^\sim)^{-1} + (P_{t+1}^b)^{-1}. \tag{48}$$

Since $W_{t+1}^f + W_{t+1}^b = I$, the smoother error can be decomposed similarly to (47) as

$$x_{t+1} - x_{t+1}^\star = W_{t+1}^f\, (x_{t+1} - x_{t+1}^\sim) + W_{t+1}^b\, (x_{t+1} - x_{t+1}^b). \tag{49}$$

From the previous section (II.B), the estimation error $x_{t+1} - x_{t+1}^b$ would be a linear combination of the observation noise processes $v_{t+1}, \ldots, v_T$ since x_{t+1}^b is the maximum-likelihood estimate based on the objective $p(Y_{t+1:T} | x_{t+1})$ (which requires co-estimation of $w_{t+1}, \ldots, w_{T-1}$ to reach every observation). Similarly, the estimation error

$x_{t+1} - x^{\sim}_{t+1}$ (and $x_t - x^{\circ}_t$) is a linear combination of $v_1, \ldots, v_t$. Because the two sets of observation noise processes are disjoint,

$$\mathcal{E}[(x_{t+1} - x^{\sim}_{t+1})(x_{t+j} - x^{b}_{t+j})^{\top}] = \mathbf{0} \tag{50}$$

$$\mathcal{E}[(x_t - x^{\circ}_t)(x_{t+j} - x^{b}_{t+j})^{\top}] = \mathbf{0} \tag{51}$$

for $j = 1, \ldots, T - t$.