On the Approximation of the Quotient of Two Gaussian Densities for Multiple-Model Smoothing

Yi Liu, Xi Li, Le Yang, Lyudmila Mihaylova and Yanbo Xue

June 8, 2023

On the Approximation of the Quotient of Two Gaussian Densities for Multiple-Model Smoothing

Yi Liu[1◇], Xi Li[2◇], Le Yang[1*], Lyudmila Mihaylova[3], Yanbo Xue[4]
1. Department of Electrical and Computer Engineering, University of Canterbury, Christchurch NZ
2. State Key Lab of Complex Electromagnetic Environment Effects on Electronics and Information Systems, National University of Defense Technology, Changsha China
3. Department of Automatic Control and System Engineering, University of Sheffield, Sheffield UK
4. Career Science Lab, Beijing China
◇: Equal contributors, *: corresponding author, le.yang@canterbury.ac.nz

***Abstract*—The quotient of two multivariate Gaussian densities can be written as an unnormalized Gaussian density, which has been applied in some recently developed multiple-model fixed-interval smoothing algorithms. However, this expression is invalid if instead of being positive definite, the covariance of the unnormalized Gaussian density is indefinite (i.e., it has both positive and negative eigenvalues) or undefined (i.e., computing it requires inverting a singular matrix). This paper considers approximating the quotient of two Gaussian densities in this case using two different approaches to mitigate the caused numerical problems. The first approach directly replaces the indefinite covariance of the unnormalized Gaussian density with a positive definite matrix nearest to it. The second approach computes the approximation through solving, using the natural gradient, an optimization problem with a Kullback-Leibler divergence-based cost function. This paper illustrates the application of the theoretical results by incorporating them into an existing smoothing method for jump Markov systems and utilizing the obtained smoothers to track a maneuvering target.**

I. Introduction

Consider the multivariate Gaussian density $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ in $\mathbf{x}\in R^{n\times 1}$ with mean $\boldsymbol{\mu}$ and covariance Σ. n is the dimensionality of $\mathbf{x}$. The quotient of two multivariate Gaussian densities $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$ and $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)$ can be written as [1]–[3]

$$\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)}=\frac{|\boldsymbol{\Sigma}_a|}{|\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c|}\cdot\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_b,\boldsymbol{\Sigma}_b)}{\mathcal{N}(\boldsymbol{\mu}_a;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c)},\tag{1}$$

where

$$\boldsymbol{\Sigma}_b=(\boldsymbol{\Sigma}_c^{-1}-\boldsymbol{\Sigma}_a^{-1})^{-1}=\boldsymbol{\Sigma}_c+\boldsymbol{\Sigma}_c(\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c)^{-1}\boldsymbol{\Sigma}_c,\tag{2a}$$

$$\boldsymbol{\mu}_b=\boldsymbol{\Sigma}_b(\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c-\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a).\tag{2b}$$

It can be seen that the term on the right hand side of (1) is an unnormalized Gaussian density. The expression (1) has been used in distributed target tracking to decorrelate the information before fusion in order to avoid producing overconfident estimates [2], [4], [5]. It is also employed to remove the factor to be updated from the current approximation of the posterior and compute the updated factor in Gaussian expectation propagation (EP) for approximate Bayesian inference [1], [6], [7].

Recent work on fixed-interval smoothing for jump Markov systems [3], [8]–[11] utilizes the quotient of two Gaussian densities as well. The considered smoothing problem aims at finding the joint posterior $p(\mathbf{x}_t, M_t^j|\mathbf{z}_{1:k})$ for $1\leq t\leq k$ using all the k measurements collected within the interval [12]–[14]. Here, $\mathbf{x}_t$ is the system state at time t. M_t^j represents that the system follows the jth state-space model in the time interval $(t-1,t]$, where $j=1,2,...,r$ and r is the number of models admitted by the jump Markov system.

To obtain low-complexity algorithms, the joint posterior $p(\mathbf{x}_t, M_t^j|\mathbf{z}_{1:k})$ is computed recursively from $t=k-1$ to $t=1$ through approximating the optimal backward-time smoothing equation. For example, in [11], we evaluate

$$p(\mathbf{x}_t, M_t^j|\mathbf{z}_{1:k})=\sum_{i=1}^{r}\frac{p(\mathbf{x}_t|M_t^j,\mathbf{z}_{1:t})p(\mathbf{x}_t|M_{t+1}^i,\mathbf{z}_{1:k})}{p(\mathbf{x}_t|M_{t+1}^i,\mathbf{z}_{1:t})}\cdot h_{t|k}^{ji},\tag{3}$$

where $h_{t|k}^{ji}$ is a factor depending on the model probabilities. All the densities in $\mathbf{x}_t$ in the summands of (3) are Gaussian if Gaussian filtering [15] is adopted in the algorithm implementation. Thus, by exploring the product rule [12] and quotient rule (1) of Gaussian densities, we can express these summands as weighted Gaussian densities for merging to achieve multiple-model smoothing. [3] follows the same idea but tries to approximate a different form of (3). In [8]–[10], the quotient of two Gaussian densities in (3), $p(\mathbf{x}_t|M_{t+1}^i,\mathbf{z}_{1:k})$ and $p(\mathbf{x}_t|M_{t+1}^i,\mathbf{z}_{1:t})$, is needed to calculate the smoothed mixing probabilities to achieve model interaction in retrodiction.

The expression (1) is valid only if the covariance $\boldsymbol{\Sigma}_b$ of the unnormalized Gaussian density is positive definite (i.e., $\boldsymbol{\Sigma}_b\succ\mathbf{O}$), or equivalently, $\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c$ is positive definite (see (2a)). However, this may not always be satisfied in the backward-time recursion of the multiple-model smoothers [3], [8]–[11]. Empirical results show that both $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c$ could be indefinite, causing $1/\mathcal{N}(\boldsymbol{\mu}_a;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c)$ in (1) to be a complex number. Besides, $\boldsymbol{\Sigma}_b$ may be undefined when $\boldsymbol{\Sigma}_a-\boldsymbol{\Sigma}_c$ is singular. These numerical problems can render the smoothing output meaningless. We cannot simply discard the affected results at the current time as in [2] either, because this would completely interrupt the backward-time recursion.

To address the above problem, [10] considers the Gaussian density obtained by combining the result from (1) to be a flat prior to avoid the propagation of numerical problems. In [3], [11], we apply the uncertainty-injection (UI) technique [16], [17], where we keep increasing a factor $\lambda>1$ to scale up $\boldsymbol{\Sigma}_a$

until the approximation to $\mathbf{\Sigma}_b$ in (2a),

$$(\mathbf{\Sigma}_c^{-1} - (\lambda \cdot \mathbf{\Sigma}_a)^{-1})^{-1}, \tag{4}$$

becomes positive definite. Despite of their effectiveness in the empirical studies, these methods appear to be *ad hoc*.

This paper carries out further studies on approximating, for multiple-model smoothing, the quotient of two multivariate Gaussian densities when $\mathbf{\Sigma}_b$ is indefinite or undefined. We consider two approaches and focus on finding approximations that either have closed-form solutions or can be found through simple iterative search. In the first approach, we calculate $\mathbf{\Sigma}_b$ using (2a) and find a positive definite matrix $\hat{\mathbf{\Sigma}}_b$ nearest to it. Next, $\boldsymbol{\mu}_b$ in (2b) and the scaling factor in (1) are computed using $\hat{\mathbf{\Sigma}}_b$ in place of $\mathbf{\Sigma}_b$ to achieve the approximation.

The second approach attempts to find a Gaussian density $\mathcal{N}(\mathbf{x}; \tilde{\boldsymbol{\mu}}_b, \tilde{\mathbf{\Sigma}}_b)$ such that multiplying it with $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)$ and normalizing the result yield a Gaussian density closest in the sense of Kullback-Leibler divergence (KLD) to $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$. We impose appropriate constraints on $\tilde{\mathbf{\Sigma}}_b$ when minimizing the KLD, and natural gradient is applied to derive a simple interative solution. Interestingly, we show that the UI techique in (4) is a special case of this approximation approach, which justifies, to some extent, the effectiveness of the UI technique seen in empirical studies. We incorporate the results of the theoretical developments into an existing multiple-model fixed-interval smoother [11] to demonstrate their performance in tracking a maneuvering target.

The rest of this paper is organized as follows. Section II gives methods for computing the positive definite matrix approximation to $\mathbf{\Sigma}_b$. Section III derives the KLD minimization-based approximation using natural gradient. Section IV gives some numerical results. Section V concludes the paper.

II. Positive Definite Matrix Approximation

A. Problem Formulation

By exploiting the product rule of Gaussian densities [12], we obain an alternative but equivalent form for the quotient rule in (1), which is [2], [3]

$$\frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)} = \frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_b, \mathbf{\Sigma}_b)}{\mathcal{N}(\boldsymbol{\mu}_a; \boldsymbol{\mu}_b, \mathbf{\Sigma}_a + \mathbf{\Sigma}_b)}, \tag{5}$$

where $\boldsymbol{\mu}_b$ and $\mathbf{\Sigma}_b$ are defined in (2). Thus, a straightforward idea for approximating the expression (1) is to directly replace $\mathbf{\Sigma}_b$ in (5) and (2b) with a positive definite matrix $\hat{\mathbf{\Sigma}}_b$ when $\mathbf{\Sigma}_b$ is indefinite. To establish a meaningful approximation and improve numerical stability, we expect that $\hat{\mathbf{\Sigma}}_b$ is nearest to $\mathbf{\Sigma}_b$ under certain criterion and may have its condition number $\text{cond}(\hat{\mathbf{\Sigma}}_b) = \lambda_{\text{max}}(\hat{\mathbf{\Sigma}}_b)/\lambda_{\text{min}}(\hat{\mathbf{\Sigma}}_b)$ upper-bounded by κ. Here, $\lambda_{\text{max}}(\mathbf{X})$ and $\lambda_{\text{min}}(\mathbf{X})$ denote the maximum and minimum eigenvalues of $\mathbf{X}$, and κ is a user-specified positive number.

Mathematically, with the above approach, when the covariance $\mathbf{\Sigma}_b$ is indefinite, the quotient for two multivariate Gaussian densities in (1) is approximated using

$$\frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)} \approx \frac{\mathcal{N}(\mathbf{x}; \hat{\boldsymbol{\mu}}_b, \hat{\mathbf{\Sigma}}_b)}{\mathcal{N}(\boldsymbol{\mu}_a; \hat{\boldsymbol{\mu}}_b, \mathbf{\Sigma}_a + \hat{\mathbf{\Sigma}}_b)}, \tag{6}$$

where

$$\hat{\boldsymbol{\mu}}_b = \hat{\mathbf{\Sigma}}_b(\mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c - \mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a). \tag{7}$$

$\hat{\mathbf{\Sigma}}_b$ is found by solving the following optimization problem:

$$\begin{aligned} &\min_{\hat{\mathbf{\Sigma}}_b} \mathcal{L}(\hat{\mathbf{\Sigma}}_b, \mathbf{\Sigma}_b) \\ &\text{subject to } \hat{\mathbf{\Sigma}}_b \succ \mathbf{O} \text{ and other constraints.} \end{aligned} \tag{8}$$

The cost function $\mathcal{L}(\hat{\mathbf{\Sigma}}_b, \mathbf{\Sigma}_b)$ quantifies the difference between $\mathbf{\Sigma}_b$ in (2a) and $\hat{\mathbf{\Sigma}}_b$. Besides the positive definiteness constraint $\hat{\mathbf{\Sigma}}_b \succ \mathbf{O}$, the condition number constraint $\text{cond}(\hat{\mathbf{\Sigma}}_b) \leq \kappa$ may be incorporated into (8) as well.

B. Low-Complexity Approximations

The problem of finding the nearest positive definite matrix approximation probably originates from finance [18]. For instance, the covariance to be approximated could be the sample covariance $\mathbf{S} = \frac{1}{p}\sum_{i=1}^{p}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$, where $\bar{\mathbf{y}} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{y}_i$. Some elements in $\mathbf{S}$ are modified to make them consistent with prior knowledge such as two variables usually being positively correlated. This may render the covariance indefinite and thus, approximation is needed.

There are a number of methods available for nearest positive definite matrix approximation. Many of them are not suitable for the problem considered in this paper, where the covariance $\mathbf{\Sigma}_b$ to be approximated is obtained *algebraically* (see (2a)), rather than from a set of data samples. As an example, when $\mathbf{\Sigma}_b$ is indefinite, we may denote its eigenvalue decomposition as $\mathbf{\Sigma}_b = \mathbf{U}\text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)\mathbf{U}^T$, where

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m > 0 > \lambda_{m+1} \geq \lambda_{m+1} \geq ... \geq \lambda_n, \tag{9}$$

and $\mathbf{U}$ is orthonormal. In words, $\mathbf{\Sigma}_b$ has m positive eigenvalues and $n-m$ negative eigenvalues, where $m < n$. The Stein's estimator [19] approximates $\mathbf{\Sigma}_b$ using

$$\hat{\mathbf{\Sigma}}_b = \mathbf{U}\text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_n)\mathbf{U}^T, \tag{10}$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_n > 0$. The eigenvalues $\hat{\lambda}_i$ are found by applying the isotonic regression [20] to λ_i/l_i, and $l_i = \frac{1}{p}\left(p - n + 1 + 2\lambda_i\sum_{j\neq i}\frac{1}{\lambda_i - \lambda_j}\right)$. Evaluating l_i requires knowledge on p, the number of samples used to generate $\mathbf{\Sigma}_b$, which is not known in our case. The Ledoit-Wolf (LW) estimator [21] finds the eigenvalues of $\hat{\mathbf{\Sigma}}_b$ in (10) using $\hat{\lambda}_i = (1-\alpha)\lambda_i + \alpha\gamma$. This corresponds to restraining $\hat{\mathbf{\Sigma}}_b$ to be $\hat{\mathbf{\Sigma}}_b = (1-\alpha)\mathbf{\Sigma}_b + \alpha\gamma\mathbf{I}$. The coefficients α and γ are computed by approximating the optimal solution that minimizes an expected quadratic loss function, which still requires using data samples unavailable in our problem.

In the following, we shall present several simple positive definite matrix approximations to $\mathbf{\Sigma}_b$ when it is indefinite. Each of them is obtained via solving an associated optimization problem with its generic formulation given in (8), and the resulting approximations all take the form in (10). In other words, they are found by manipulating the eigenvalues λ_i of the original covariance $\mathbf{\Sigma}_b$ with different techniques.

*1) **Approximation based on diagonal loading:*** We adopt the idea of the LW estimator [21] and set $\hat{\mathbf{\Sigma}}_b$ to be $\hat{\mathbf{\Sigma}}_b = (1-\alpha)\mathbf{\Sigma}_b + \beta\mathbf{I}$, where $0 \leq \alpha < 1$ and $\beta > 0$. From (9), the minimum eigenvalue of $\hat{\mathbf{\Sigma}}_b$ is $\hat{\lambda}_n = (1-\alpha)\lambda_n + \beta$. This converts the positive definiteness constraint in (8) into $\hat{\lambda}_n > 0$ or equivalently, $\beta > -(1-\alpha)\lambda_n$. Besides, we also include the condition number constraint cond$(\hat{\mathbf{\Sigma}}_b) \leq \kappa$ in (8), which requires $((1-\alpha)\lambda_1 + \beta)/((1-\alpha)\lambda_n + \beta) \leq \kappa$. Combining these inequalities with $\beta > 0$ yields

$$\beta \geq (1-\alpha)\frac{\lambda_1 - \kappa\lambda_n}{\kappa - 1}. \tag{11}$$

Finally, applying the cost function $\mathcal{L}(\hat{\mathbf{\Sigma}}_b, \mathbf{\Sigma}_b) = ||\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b||_F^2$ transforms the optimization problem (8) into one with two variables α and β:

$$\begin{aligned} &\min_{\alpha,\beta} \sum_{i=1}^{n}(\beta - \alpha\lambda_i)^2 \\ &\text{subject to } 0 \leq \alpha < 1 \text{ and } \beta \geq (1-\alpha)\frac{\lambda_1 - \kappa\lambda_n}{\kappa - 1}. \end{aligned} \tag{12}$$

Here, $||\mathbf{X}||_F = \sqrt{\text{tr}(\mathbf{X}\mathbf{X}^T)}$ and tr$(\mathbf{X})$ denote the Frobenius norm and trace of $\mathbf{X}$, and we have applied $||\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b||_F^2 = ||\beta\mathbf{I} - \alpha\mathbf{\Sigma}_b||_F^2 = \sum_{i=1}^{n}(\beta - \alpha\lambda_i)^2$, where the third equality comes from the orthogonal invariance of the Frobenius norm.

A suboptimal solution to (12) is $\alpha = 0$ and $\beta = \frac{\lambda_1 - \kappa\lambda_n}{\kappa - 1}$. It is a good solution when the minimum eigenvalue of $\mathbf{\Sigma}_b$ has a relatively small magnitude (i.e., $\lambda_1 \gg |\lambda_n|$). As a result, the obtained positive definite matrix approximation is

$$\hat{\mathbf{\Sigma}}_b = \mathbf{\Sigma}_b + \frac{\lambda_1 - \kappa\lambda_n}{\kappa - 1}\mathbf{I}, \tag{13}$$

which has the form of diagonal loading.

*2) **Approximation by lower-bounding** $\hat{\lambda}_n$:* This method replaces the positive definiteness constraint in (8) with n positive lower bounds on the eigenvalues of $\hat{\mathbf{\Sigma}}_b$, which are $\hat{\lambda}_i \geq \gamma > 0, i = 1, 2, ..., n$. We continue to employ the squared Frobenius norm of the difference matrix $\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b$ as the cost function such that the approximation to $\mathbf{\Sigma}_b$ is found by solving

$$\begin{aligned} &\min_{\hat{\mathbf{\Sigma}}_b} ||\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b||_F^2 \\ &\text{subject to } \hat{\lambda}_i \geq \gamma,\ i = 1, 2, ..., n. \end{aligned} \tag{14}$$

The above problem is an extension of the one in [22] formulated to compute the nearest positive semidefinite matrix to an arbitrary symmetric matrix. We can follow the same argument in [22] to show that the cost function in (14) satisfies

$$||\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b||_F^2 \geq \sum_{i=1}^{n}(\hat{\lambda}_i - \lambda_i)^2 \geq \sum_{\lambda_i < \gamma}(\gamma - \lambda_i)^2. \tag{15}$$

The two inequalities in (15) would become equalities if $\hat{\mathbf{\Sigma}}_b$ takes the form in (10) with its eigenvalues equal to

$$\hat{\lambda}_i = \begin{cases} \lambda_i, & \text{if } \lambda_i \geq \gamma \\ \gamma, & \text{if } \lambda_i < \gamma \end{cases}, \tag{16}$$

which is the optimal solution to (14). In words, this approximation $\hat{\mathbf{\Sigma}}_b$ is obtained by increasing the eigenvalues of the original covariance $\mathbf{\Sigma}_b$ that are lower than the threshold γ to γ while keeping other eigenvalues intact.

*3) **Spectral norm-based approximation:*** The approximation $\hat{\mathbf{\Sigma}}_b$ given in (10) and (16) implicitly assumes that only the small eigenvalues of $\mathbf{\Sigma}_b$ need to be corrected when it is indefinite. To simultaneously account for the possible overestimation of the large eigenvalues of $\mathbf{\Sigma}_b$, [23] finds $\hat{\mathbf{\Sigma}}_b$ via solving (8) with the cost function chosen to be the spectral norm $\max_i |\lambda_i(\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b)|$. Here, $\lambda_i(\mathbf{X})$ denotes the ith largest eigenvalue of $\mathbf{X}$. The constraints adopted are the positive definiteness constraint $\hat{\mathbf{\Sigma}}_b \succ \mathbf{O}$ and condition number constraint cond$(\hat{\mathbf{\Sigma}}_b) \leq \kappa$. The obtained approximation $\hat{\mathbf{\Sigma}}_b$ still takes the form in (10), whose eigenvalues are [23]

$$\hat{\lambda}_i = \begin{cases} \delta, & \text{if } \lambda_i < \delta \\ \lambda_i, & \text{if } \delta \leq \lambda_i \leq \kappa\delta \\ \kappa\delta, & \text{if } \lambda_i > \kappa\delta \end{cases}, \tag{17}$$

where $\delta = \frac{\lambda_1 + \lambda_n}{\kappa + 1}$. It is evident from (17) that this estimator clips the large eigenvalues of $\mathbf{\Sigma}_b$ if they are bigger than $\kappa\delta$.

*4) **Trace norm-based approximation:*** The approximation $\hat{\mathbf{\Sigma}}_b$ may also be found through solving the same optimization problem as in the spectral norm-based case but with its cost function replaced by the trace norm $\sum_{i=1}^{n}|\lambda_i(\hat{\mathbf{\Sigma}}_b - \mathbf{\Sigma}_b)|$ [23]. The optimal $\hat{\mathbf{\Sigma}}_b$ is again given in (10) and (16), where the threshold γ in (16) is now equal to $\gamma = \lambda_1/\kappa$.

*5) **Maximum-likelihood approximation:*** In [24], $\mathbf{\Sigma}_b$ is assumed to be the sample covariance of the data samples drawn independently from a zero-mean Gaussian density with unknown covariance $\hat{\mathbf{\Sigma}}_b$. The problem of approximating $\mathbf{\Sigma}_b$ is thus cast into the maximum likelihood estimation of $\hat{\mathbf{\Sigma}}_b$ under the positive definiteness and condition number constraints. The associated optimization problem is

$$\begin{aligned} &\min_{\hat{\mathbf{\Sigma}}_b} \text{tr}\left(\hat{\mathbf{\Sigma}}_b^{-1}\mathbf{\Sigma}_b\right) - \log\left|\hat{\mathbf{\Sigma}}_b^{-1}\right| \\ &\text{subject to } \hat{\mathbf{\Sigma}}_b \succ \mathbf{O} \text{ and cond}(\hat{\mathbf{\Sigma}}_b) \leq \kappa. \end{aligned} \tag{18}$$

$|\mathbf{X}|$ denotes the matrix determinant of $\mathbf{X}$.

An efficient method has been developed in [24] to solve (18) when $\mathbf{\Sigma}_b$ is positive semidefinite. We shall extend it to the scenario in consideration where $\mathbf{\Sigma}_b$ may be indefinite. In this case, we can show by following [24] and utilizing the results in Chapter 14.2 of [25] that the cost function tr$(\hat{\mathbf{\Sigma}}_b^{-1}\mathbf{\Sigma}_b) - \log|\hat{\mathbf{\Sigma}}_b^{-1}|$ attains its lower bound $\sum_{i=1}^{n}\lambda_i u_i - \log(u_i)$ when $\hat{\mathbf{\Sigma}}_b$ is given in (10) and $u_i = 1/\hat{\lambda}_i$. With the introduction of an auxiliary variable u, (18) can be transformed into [24]

$$\begin{aligned} &\min_{u>0, u_i} \sum_{i=1}^{n}\lambda_i u_i - \log(u_i) \\ &\text{subject to } u \leq u_i \leq \kappa u. \end{aligned} \tag{19}$$

Note that the condition number constraint on $\hat{\mathbf{\Sigma}}_b$ is now imposed on the inverse of its eigenvalues $u_i = 1/\hat{\lambda}_i$.

To solve (19), we first observe that $\lambda_i u_i - \log(u_i)$ is convex with respect to u_i, and it has a critical point $1/\lambda_i$. Besides, if $\lambda_i \leq 0$, $\lambda_i u_i - \log(u_i)$ would decrease monotonically as u_i increases. Therefore, given $u > 0$, the ith summand of the cost function in (19) reaches its minimum value

$$J_i(u) = \begin{cases} \lambda_i \kappa u - \log(\kappa u), & \text{if } \lambda_i \leq 0 \text{ or } \frac{1}{\lambda_i} > \kappa u \\ 1 + \log(\lambda_i), & \text{if } u \leq \frac{1}{\lambda_i} \leq \kappa u \\ \lambda_i u - \log(u), & \text{if } \frac{1}{\lambda_i} < u \end{cases}, \quad (20)$$

when the optimization variable u_i is equal to

$$u_i = \frac{1}{\hat{\lambda}_i} = \begin{cases} \kappa u, & \text{if } \lambda_i \leq 0 \text{ or } \frac{1}{\lambda_i} > \kappa u \\ 1/\lambda_i, & \text{if } u \leq \frac{1}{\lambda_i} \leq \kappa u \\ u, & \text{if } \frac{1}{\lambda_i} < u \end{cases}. \quad (21)$$

With the above results, (19) reduces to a univariate optimization problem: $\min\limits_{u>0} \sum_{i=1}^{n} J_i(u)$.

To find the minimizer for $\sum_{i=1}^{n} J_i(u)$ under $u > 0$, we note that $J_i(u)$ can take different forms (see (20)), depending on the relationship between u, $1/\lambda_i$ and $1/(\kappa\lambda_i)$, but they are still convex with respect to u. Therefore, we can sort in an ascending order $1/\lambda_i$ and $1/(\kappa\lambda_i)$, $i = 1, 2, ..., m$, which are proportional to the inverse of the positive eigenvalues of $\mathbf{\Sigma}_b$ (see (9)), and obtain $0 < b_1 \leq b_2 \leq ... \leq b_{2m}$. In this way, $2m - 1$ intervals $(b_j, b_{j+1}]$, $j = 1, 2, ..., 2m - 1$, are generated. We then minimize the cost function $\sum_{i=1}^{n} J_i(u)$ over each interval and output the minimizer that produces the smallest function value as the optimal solution u^*. Finally, u^* is substituted back into (21) to find $\hat{\lambda}_i$, which is put into (10) to generate the positive definite matrix approximation $\hat{\mathbf{\Sigma}}_b$.

III. KLD Minimization-based Approximation

A. Problem Formulation

The alternative expression for the quotient of two multivariate Gaussian densities in (5) can be further re-written as

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c) = \frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_b, \mathbf{\Sigma}_b)}{\mathcal{N}(\boldsymbol{\mu}_a; \boldsymbol{\mu}_b, \mathbf{\Sigma}_a + \mathbf{\Sigma}_b)}. \quad (22)$$

The term on the right hand side of (22) can be considered as the measurement update stage of a linear Kalman filter, with $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)$ being the *likelihood*, $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_b, \mathbf{\Sigma}_b)$ being the *prior* and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$ being the *posterior* [12]–[14].

Thus, another approach for approximating (5) or the original expression (1) when $\mathbf{\Sigma}_b$ is not positive definite can be established. In particular, we aim at finding the prior $\mathcal{N}(\mathbf{x}; \tilde{\boldsymbol{\mu}}_b, \tilde{\mathbf{\Sigma}}_b)$ such that multiplying it with the likelihood $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \mathbf{\Sigma}_a)$ and normalizing the result as in (22) produce the posterior $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)$ closest in terms of KLD to $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$. This approach is different from the methods in Section II as it utilizes KLD minimization to reduce the amount of *distortion* when approximating the quotient of two Gaussian densities. The corresponding optimization problem with positive definiteness and condition number constraints on $\tilde{\mathbf{\Sigma}}_b$ is

$$\begin{aligned} &\min_{\tilde{\boldsymbol{\mu}}_b, \tilde{\mathbf{\Sigma}}_b} \; \mathcal{D}(\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)||\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)) \\ &\text{subject to } \tilde{\mathbf{\Sigma}}_b \succ \mathbf{O} \text{ and } \mathrm{cond}(\tilde{\mathbf{\Sigma}}_b) \leq \kappa, \end{aligned} \quad (23)$$

where by the product rule of Gaussian densities [12], we have

$$\mathbf{\Sigma}_d = (\mathbf{\Sigma}_a^{-1} + \tilde{\mathbf{\Sigma}}_b^{-1})^{-1}, \; \boldsymbol{\mu}_d = \mathbf{\Sigma}_d(\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a + \tilde{\mathbf{\Sigma}}_b^{-1}\tilde{\boldsymbol{\mu}}_b). \quad (24)$$

$\mathcal{D}(\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)||\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c))$ is the reverse KLD between $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)$ and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$[1], which is [26]

$$\begin{aligned} &\mathcal{D}(\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)||\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)) \\ &= \int \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d) \log \frac{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d)}{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)} \mathrm{d}\mathbf{x} \\ &\propto \mathrm{tr}\left(\mathbf{\Sigma}_c^{-1}((\boldsymbol{\mu}_d - \boldsymbol{\mu}_c)(\boldsymbol{\mu}_d - \boldsymbol{\mu}_c)^T + \mathbf{\Sigma}_d)\right) - \log|\mathbf{\Sigma}_d|. \end{aligned} \quad (25)$$

The desired KLD minimization-based approximation to the quotient of two multivariate Gaussian densities given in (1) can thus be achieved by solving (23) for $\tilde{\boldsymbol{\mu}}_b$ and $\tilde{\mathbf{\Sigma}}_b$, and putting them into (6) in place of $\hat{\boldsymbol{\mu}}_b$ and $\hat{\mathbf{\Sigma}}_b$.

Note from (25) that the cost function reaches the lower bound when $\boldsymbol{\mu}_d = \boldsymbol{\mu}_c$. We immediately have that the optimal $\tilde{\boldsymbol{\mu}}_b$ is given by, according to (24),

$$\tilde{\boldsymbol{\mu}}_b = \tilde{\mathbf{\Sigma}}_b \left((\mathbf{\Sigma}_a^{-1} + \tilde{\mathbf{\Sigma}}_b^{-1})\boldsymbol{\mu}_c - \mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right). \quad (26)$$

We can substitute (26) into (25) and convert the optimization problem (23) into one with a single variable $\tilde{\mathbf{\Sigma}}_b$. However, estimating the covariance $\tilde{\mathbf{\Sigma}}_b$ from this newly obtained problem is still cumbersome.

In the remaining of this section, we shall first simplify the problem in (23) and then derive a simple iterative method to find an estimate of $\tilde{\mathbf{\Sigma}}_b$ to accomplish the KLD minimization-based approximation.

B. Problem Simplification

The theoretical development begins with noticing that the multivariate Gaussian density is a member of the exponential family, and the density function $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma})$ can be expressed as [26], [27]

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \exp\left(\boldsymbol{\eta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\eta})\right), \quad (27)$$

where $\boldsymbol{\eta}$ is the natural parameter and $\boldsymbol{\phi}(\mathbf{x})$ is the sufficient statistic for $\boldsymbol{\eta}$. With slight abuse of notations, we have [28]

$$\boldsymbol{\eta}^T \boldsymbol{\phi}(\mathbf{x}) = \mathrm{tr}\left(\left(\boldsymbol{\eta}^{(1)}\right)^T \boldsymbol{\phi}_1(\mathbf{x}) + \boldsymbol{\eta}^{(2)} \boldsymbol{\phi}_2(\mathbf{x})\right),$$

where

$$\boldsymbol{\eta}^{(1)} = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \qquad \boldsymbol{\eta}^{(2)} = -\frac{1}{2}\mathbf{\Sigma}^{-1}, \quad (28a)$$

$$\boldsymbol{\phi}_1(\mathbf{x}) = \mathbf{x}, \qquad \boldsymbol{\phi}_2(\mathbf{x}) = \mathbf{x}\mathbf{x}^T, \quad (28b)$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}^{(1)} \\ \mathrm{vec}(\boldsymbol{\eta}^{(2)}) \end{bmatrix}, \quad \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\phi}_1(\mathbf{x}) \\ \mathrm{vec}(\boldsymbol{\phi}_2(\mathbf{x})) \end{bmatrix}, \quad (28c)$$

and $\mathrm{vec}(\mathbf{X})$ is the column-vectorised version of the matrix $\mathbf{X}$.

$A(\boldsymbol{\eta})$ in (27) is the cumulant function, which is equal to $A(\boldsymbol{\eta}) = \frac{1}{2}\log|2\pi\mathbf{\Sigma}| + \frac{1}{2}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$. It has two useful properties.

[1] The forward KLD $\mathcal{D}(\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c)||\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_d, \mathbf{\Sigma}_d))$ [26] may be used as the cost function in (23) as well. Developing a gradient-based solution for this newly formulated optimization problem and investigating its performance when being incorporated into multiple-model fixed-interval smoothers will be subject to future study.

Specifically, its first-order and second-order partial derivatives with respect to the natural parameter $\boldsymbol{\eta}$ are equal to [27]

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = E(\boldsymbol{\phi}(\mathbf{x})), \tag{29a}$$

$$\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \text{cov}(\boldsymbol{\phi}(\mathbf{x})), \tag{29b}$$

where $\text{cov}(\boldsymbol{\phi}(\mathbf{x})) = E(\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}^T(\mathbf{x})) - E(\boldsymbol{\phi}(\mathbf{x}))E(\boldsymbol{\phi}^T(\mathbf{x}))$ is the covariance of the sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$.

The Fisher information matrix (FIM) [29] of the natural parameter $\boldsymbol{\eta}$, denoted by $\text{FIM}(\boldsymbol{\eta})$, is equal to $\partial^2 A(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T$. This can be verified by noting that

$$\begin{aligned}\text{FIM}(\boldsymbol{\eta}) &= E\left(\frac{\partial \log \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\eta}} \cdot \frac{\partial \log \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\eta}^T}\right) \\ &= E\left(\left(\boldsymbol{\phi}(\mathbf{x}) - \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)\left(\boldsymbol{\phi}(\mathbf{x}) - \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)^T\right) \\ &= \text{cov}(\boldsymbol{\phi}(\mathbf{x})) = \partial^2 A(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T,\end{aligned} \tag{30}$$

where (27) and (29) have been substituted to arrive at the last three equalities. We would like to point out that the two properties of $A(\boldsymbol{\eta})$, given in (29) and (30), are valid for *any* member of the exponential family as long as its distribution function is written in the form $\exp(\boldsymbol{\eta}^T\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\eta}))$.

We explore these results to simplify the optimization problem (23). Let $\boldsymbol{\eta}_d$ and $\boldsymbol{\eta}_c$ denote the natural parameters of Gaussian densities $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)$ and $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$. We then apply (27) and (29) to transform the cost function in (25) into

$$\begin{aligned}&\mathcal{D}(\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)||\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)) \\ &= \int \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)\left((\boldsymbol{\eta}_d - \boldsymbol{\eta}_c)^T\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\eta}_d) + A(\boldsymbol{\eta}_c)\right)\mathrm{d}\mathbf{x} \\ &= (\boldsymbol{\eta}_d - \boldsymbol{\eta}_c)^T\frac{\partial A(\boldsymbol{\eta}_d)}{\partial \boldsymbol{\eta}_d} - A(\boldsymbol{\eta}_d) + A(\boldsymbol{\eta}_c).\end{aligned} \tag{31}$$

Besides, from the definition of $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)$ given above (23) and in (24), the natural parameter $\boldsymbol{\eta}_d$ can be expressed in terms of $\boldsymbol{\eta}_a$ and $\tilde{\boldsymbol{\eta}}_b$, the natural parameters of $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)$ and $\mathcal{N}(\mathbf{x};\tilde{\boldsymbol{\mu}}_b,\tilde{\boldsymbol{\Sigma}}_b)$, as

$$\boldsymbol{\eta}_d = \boldsymbol{\eta}_a + \tilde{\boldsymbol{\eta}}_b. \tag{32}$$

In other words, $\boldsymbol{\eta}_d$ is just a shifted version of $\tilde{\boldsymbol{\eta}}_b$.

Finally, note from (28) that the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of a Gaussian density $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be deduced from its natural parameter $\boldsymbol{\eta}$ uniquely via

$$\boldsymbol{\mu} = -\frac{1}{2}\left(\boldsymbol{\eta}^{(2)}\right)^{-1}\boldsymbol{\eta}^{(1)}, \tag{33a}$$

$$\boldsymbol{\Sigma} = -\frac{1}{2}\left(\boldsymbol{\eta}^{(2)}\right)^{-1}. \tag{33b}$$

As a result, through applying (33), (28) and (32), we are able to express the covariance of $\mathcal{N}(\mathbf{x};\tilde{\boldsymbol{\mu}}_b,\tilde{\boldsymbol{\Sigma}}_b)$ in terms of $\boldsymbol{\eta}_d$ as

$$\tilde{\boldsymbol{\Sigma}}_b = -\frac{1}{2}\left(\tilde{\boldsymbol{\eta}}_b^{(2)}\right)^{-1} = \frac{1}{2}\left(\boldsymbol{\eta}_a^{(2)} - \boldsymbol{\eta}_d^{(2)}\right)^{-1}. \tag{34}$$

With (31), (32) and (34), the optimization problem (23) could be transformed into an equivalent one given by

$$\begin{aligned}&\min_{\boldsymbol{\eta}_d} \mathcal{C}(\boldsymbol{\eta}_d) = (\boldsymbol{\eta}_d - \boldsymbol{\eta}_c)^T\frac{\partial A(\boldsymbol{\eta}_d)}{\partial \boldsymbol{\eta}_d} - A(\boldsymbol{\eta}_d) + A(\boldsymbol{\eta}_c) \\ &\text{subject to } \tilde{\boldsymbol{\Sigma}}_b \succ \mathbf{O} \text{ and } \text{cond}(\tilde{\boldsymbol{\Sigma}}_b) \leq \kappa.\end{aligned} \tag{35}$$

C. Natural Gradient-based Approximation

We shall develop a simple gradient algorithm to solve (35) for the covariance $\tilde{\boldsymbol{\Sigma}}_b$. For this purpose, the gradient of the cost function $\mathcal{C}(\boldsymbol{\eta}_d)$ can be shown to be

$$\nabla_{\boldsymbol{\eta}_d}\mathcal{C}(\boldsymbol{\eta}_d) = \frac{\partial \mathcal{C}(\boldsymbol{\eta}_d)}{\partial \boldsymbol{\eta}_d} = \frac{\partial^2 A(\boldsymbol{\eta}_d)}{\partial \boldsymbol{\eta}_d \partial \boldsymbol{\eta}_d^T}(\boldsymbol{\eta}_d - \boldsymbol{\eta}_c). \tag{36}$$

Evaluating (36) is difficult due to the presence of the second-order partial derivative $\partial^2 A(\boldsymbol{\eta}_d)/\partial\boldsymbol{\eta}_d\partial\boldsymbol{\eta}_d^T$. To circumvent this difficulty, we resort to the natural gradient defined as [30], [31]

$$\bar{\nabla}_{\boldsymbol{\eta}_d}\mathcal{C}(\boldsymbol{\eta}_d) = (\text{FIM}(\boldsymbol{\eta}_d))^{-1}\nabla_{\boldsymbol{\eta}_d}\mathcal{C}(\boldsymbol{\eta}_d) = \boldsymbol{\eta}_d - \boldsymbol{\eta}_c, \tag{37}$$

where (30) and (36) have been applied. The natural gradient incorporates the FIM of $\boldsymbol{\eta}_d$, $\text{FIM}(\boldsymbol{\eta}_d)$, to define a Riemannian metric so that the information geometry of $\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d)$ parameterized over $\boldsymbol{\eta}_d$ is taken into account [31], [32]. Note that (37) can also be established using expectation parameter-based reparameterization [32], [33], and it has been adopted in [32]–[35] to develop scalable Bayesian deep learning techniques.

We can obtain a simple rule from (37) for estimating $\boldsymbol{\eta}_d$ iteratively based on natural gradient descent, which is

$$\boldsymbol{\eta}_{d,k+1} = \boldsymbol{\eta}_{d,k} - \rho_k\bar{\nabla}_{\boldsymbol{\eta}_{d,k}}\mathcal{C}(\boldsymbol{\eta}_d) = \boldsymbol{\eta}_{d,k} - \rho_k(\boldsymbol{\eta}_{d,k} - \boldsymbol{\eta}_c), \tag{38}$$

where $\rho_k > 0$ is the step size and $\boldsymbol{\eta}_{d,k}$ is the estimate of $\boldsymbol{\eta}_d$ in the kth iteration. Putting (32) and (28) reveals that (38) estimates simultaneously the two components of the natural parameter $\tilde{\boldsymbol{\eta}}_b$, $\tilde{\boldsymbol{\eta}}_b^{(1)}$ and $\tilde{\boldsymbol{\eta}}_b^{(2)}$, using

$$\tilde{\boldsymbol{\eta}}_{b,k+1}^{(i)} = (1 - \rho_k)\tilde{\boldsymbol{\eta}}_{b,k}^{(i)} + \rho_k(\boldsymbol{\eta}_c^{(i)} - \boldsymbol{\eta}_a^{(i)}), \ i = 1, 2, \tag{39}$$

where $\tilde{\boldsymbol{\eta}}_{b,k}^{(i)}$ is the estimate of $\tilde{\boldsymbol{\eta}}_b^{(i)}$ in iteration k.

Through substituting (33) and (34) into (39) with $i = 2$, we obtain the following rule for updating the estimate of the desired covariance $\tilde{\boldsymbol{\Sigma}}_b$

$$\tilde{\boldsymbol{\Sigma}}_{b,k+1}^{-1} = (1 - \rho_k)\tilde{\boldsymbol{\Sigma}}_{b,k}^{-1} + \rho_k(\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}). \tag{40}$$

Here, $\tilde{\boldsymbol{\Sigma}}_{b,k}$ denotes the estimate of $\tilde{\boldsymbol{\Sigma}}_b$ in iteration k.

According to (2a), $\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}$ would not be a positive definite matrix when $\boldsymbol{\Sigma}_b$ in the original expression (1) for the quotient of two multivariate Gaussian densities is not positive definite. Thus, the necessary condition for the updated estimate $\tilde{\boldsymbol{\Sigma}}_{b,k+1}$ in (40) to satisfy the positive definiteness constraint in (35) is $0 < \rho_k < 1$, since $\rho_k \geq 1$ will lead to an indefinite or even negative definite result.

Under the condition $0 < \rho_k < 1$, the update rule (40) indeed keeps shrinking the estimate of $\tilde{\boldsymbol{\Sigma}}_b^{-1}$ toward the fixed target matrix $\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}$. As such, it can be shown by mathematical induction that the output of (40) is always able to be expressed as the linear combination of the initial guess for $\tilde{\boldsymbol{\Sigma}}_b^{-1}$, denoted by $\tilde{\boldsymbol{\Sigma}}_{b,0}^{-1}$, and $\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}$, which is

$$\tilde{\boldsymbol{\Sigma}}_{b,k}^{-1} = (1-\rho)\tilde{\boldsymbol{\Sigma}}_{b,0}^{-1} + \rho(\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}),\ 0 < \rho < 1. \qquad (41)$$

If $\rho = 1$, the estimator (41) reduces to the algebraic solution in (2a). This corresponds to setting $\rho_k = 1$ in (38), which results in the *unconstrained* optimal solution to (35). In this case, we have that $\boldsymbol{\eta}_d = \boldsymbol{\eta}_c$ and the cost function, which is indeed the KLD, attains its lower bound of 0 [26], [27].

With the theoretical results derived so far, a simple iterative method for finding an estimate of $\tilde{\boldsymbol{\Sigma}}_b$ using (41) is developed. It starts with an initial guess $\tilde{\boldsymbol{\Sigma}}_{b,0}$ and keeps increasing the value of ρ within (0,1) until the output of (41) no longer satisfies the positive definiteness and condition number constraints in (35). The purpose of searching for the maximum possible ρ is to decrease the KLD-based cost function and improve the chance of producing a good solution. This can be carried out efficiently by using e.g, the bisection technique with an upper bound on the number of allowed iterations.

The obtained estimate of $\tilde{\boldsymbol{\Sigma}}_b$ will be substituted into (26) to compute the mean $\tilde{\boldsymbol{\mu}}$. They together serve as a suboptimal solution to the original optimization problem (23). Applying them in (6) in place of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}_b$ yields the KLD minimization-based approximation of the quotient of two multivariate Gaussian densities when $\boldsymbol{\Sigma}_b$ in (2a) is indefinite or undefined.

Two important observations can be made by carefully examining (41). First, if the initial guess $\tilde{\boldsymbol{\Sigma}}_{b,0}$ is set to be $\boldsymbol{\Sigma}_c$, the covariance of the Gaussian density in the numerator of the quotient rule (see (1)), (41) becomes

$$\tilde{\boldsymbol{\Sigma}}_{b,k}^{-1} = (\boldsymbol{\Sigma}_c^{-1} - (\rho^{-1}\boldsymbol{\Sigma}_a)^{-1}). \qquad (42)$$

This estimate has the same functional form as the one in (4), which indicates that the UI-based approximation adopted in our previous work [3], [11] is in fact a special case of the KLD minimization-based approximation developed here. The main difference is that the former tends to scale up $\boldsymbol{\Sigma}_a$ to guarantee the positive definiteness of the output. On the contrary, the latter tries to maximize ρ (or equivalently, minimize $1/\rho$) in (42) so as to reduce the deviation of the estimated covariance from the algebraic solution (2a). This is due to the use of KLD as the cost function (see (24) and (25)).

Besides, setting the initial guess $\tilde{\boldsymbol{\Sigma}}_{b,0}$ to be $\epsilon\mathbf{I}$, where ϵ is sufficiently large, reduces the KLD minimization-based approximation (41) to the nearest positive definition matrix approximation in Section II. But this is a poor choice. This can be justified by considering the scenario where $\boldsymbol{\Sigma}_b$ is indefinite and its eigenvalues are given in (9). According to (2a), the eigenvalues of $\boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_a^{-1}$, u_i, satisfy $u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n < 0 < u_1 \leq u_2 \leq \ldots \leq u_m$, where $u_i = 1/\lambda_i$, $i = 1, 2, \ldots, n$. Therefore, we have from (41) that the largest eigenvalue of the estimated $\tilde{\boldsymbol{\Sigma}}_b$ would be $1/((1-\rho)\epsilon + \rho u_{m+1})$, which depends on a negative eigenvalue of $\boldsymbol{\Sigma}_b$, rather than the largest eigenvalue λ_1. This is not desired, as $\tilde{\boldsymbol{\Sigma}}_b$ will be significantly different from the original covariance $\boldsymbol{\Sigma}_b$.

IV. Numerical Results

A. Simulation Setup

The simulation scenario is very similar to the one used in [3], [11]. We shall give a brief description here. For more details, readers are referred to Sections IV of [3], [11]. Specifically, a stationary sensor is deployed at the origin to estimate the smoothed trajectory of a moving target using 200 pairs of bearing and range measurements. The measurements are obtained with a sampling period of 3s (i.e., they are collected within an interval of 600s). The standard deviations of bearing and range noises are 0.5^o and 50m.

At the beginning of the tracking process, the target is 250km away from the sensor with a true bearing of 70^o. It moves in the southwest direction with an initial speed of 200m/s. Its trajectory has four segments. The first segment (0s to 200s) and third segment (219s to 479s) follow a constant velocity (CV) model whose process noise has a standard deviation of 1m/s^2. The target makes two turns, one in the second segment (201s to 218s) with an acceleration of 1g and the other in the fourth segment (480s to 599s) with an acceleration of 0.5g. Both turns are modeled by the constant turn (CT) model with process noise having a standard deviation of 1m/s^2.

B. Smoothing Algorithm Implementation

We incorporate the proposed approximations to the quotient of two multivariate Gaussian densities into the fixed-interval smoother developed in [11] for jump Markov nonlinear systems. The resulting smoothers are applied to identify the target trajectory described in the previous subsection.

At sampling index t, the considered multiple-model smoothing algorithm evaluates (3), where every density in $\mathbf{x}_t$ in the summand is approximated using the Gaussian density found by the cubature Kalman filter (CKF) [36]. The summand in (3) can be written as an unnormalized Gaussian density by applying the product rule [12] to its numerator and (1) to the resulting quotient of two multivariate Gaussian densities. The approximation to the quotient rule (1) will be invoked once the covariance $\boldsymbol{\Sigma}_b$ computed using (2a) is indefinite or undefined.

To cope with the presence of target maneuvers, we employ $r = 7$ motion models, one CV model and six CT models with different turn rates, in the simulated smoothing algorithm [11]. The state $\mathbf{x}_t$ is composed of the target position and velocity at sampling index t, $t = 1, 2, \ldots, 200$. The model index M_t^j indicates that the target follows the jth assumed motion model during $(t-1, t]$. See Sections IV of [11] for more details on the smoothing algorithm realization.

C. Approximation Algorithm Implementation

When implementing the proposed approximations to the quotient of two multivariate Gaussian densities, we need to select κ, the bound on the condition number of the covariance of the approximated unnormalized Gaussian density, $\hat{\boldsymbol{\Sigma}}_b$ or $\tilde{\boldsymbol{\Sigma}}_b$. In particular, we set κ to be the maximum of cond($\boldsymbol{\Sigma}_c$) and cond($\boldsymbol{\Sigma}_a$) for the approximations based on diagonal loading ('DL'), spectral norm ('SN'), trace norm ('TN') and KLD minimization ('KLD') (see Sections II.B and III.C). Besides, we set the initial guess $\tilde{\mathbf{\Sigma}}_{b,0}$ in the KLD minimization-based approximation to be $\mathbf{\Sigma}_c$. The number of allowed iterations is 5, which is equivalent to performing a grid search for the shrinkage factor ρ with a resolution of $1/32$ (see (42)).

There exist occasions where the smallest eigenvalue of the original covariance $\mathbf{\Sigma}_b$, λ_n, is negative but with a magnitude larger than the biggest eigenvalue λ_1 ($|\lambda_n| > \lambda_1$). As a result, the spectral norm-based approximation cannot yield a positive definite matrix approximation to $\mathbf{\Sigma}_b$, because the threshold $\delta = \frac{\lambda_1+\lambda_n}{\kappa+1}$, also the smallest eigenvalue of the approximation $\hat{\mathbf{\Sigma}}_b$, would be negative (see (17)). We avoid this problem by setting $\delta = \lambda_1/\kappa$ (i.e., the spectral norm-based approximation reduces to the trace norm-based one in this case).

For the maximum-likelihood approximation ('ML') presented in Section II.B, κ is configured to be $0.95\frac{\lambda_1}{\lambda_m}$ [24], the down-scaled ratio of the largest and smallest positive eigenvalues of the original covariance $\mathbf{\Sigma}_b$.

D. Results and Discussions

The performance of the smoothers with different approximations to the quotient of two Gaussian densities is quantified using the target position estimation root mean square error (RMSE) and target velocity estimation RMSE. The estimation RMSEs from three benchmark techniques are given for comparison. They include the forward-time interacting multiple model (IMM) filter ('IMM Filter'), the smoother in Algorithm 1 of [10] ('Method from [10]') and the original smoothing algorithm in [11] with UI-based approximation ('UI') to the quotient rule (see (4)). The results are obtained through averaging over 2000 ensemble runs.

The simulation study is conducted on a desktop running MATLAB® and Windows® 10 with a 3GHz Intel® Core i7-9700 CPU and 32GB RAM. The implemented fixed-interval smoothers with different techniques for approximating the quotient of two Gaussian densities have similar CPU run time. They all take about 0.72s to achieve fixed-interval smoothing using the collected 200 bearing and range measurements.

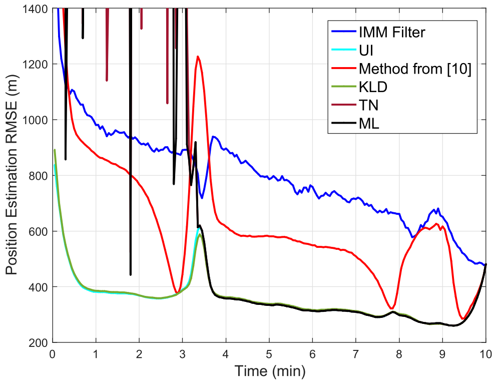

Fig. 1. Comparison of target position estimation RMSEs.

Figs. 1 and 2 plot as function of time the estimation RMSEs for the target position and velocity. It can be observed that as expected, the smoothing algorithms such as the ones from [10] ('Method from [10]') and [11] ('UI') provide better estimation accuracy over the forward-time IMM filter.

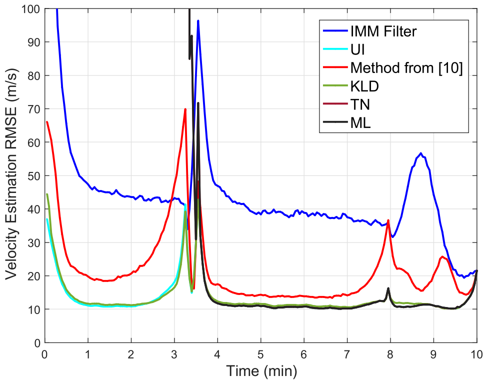

Fig. 2. Comparison of target velocity estimation RMSEs.

More importantly, the approximations developed in Section II, which try to find the nearest positive definite matrix to replace the indefinite covariance in the quotient of two Gaussian densities when it is present, perform poorly. For the sake of clarity, we only include in the figures the results from the trace norm-based ('TN') and maximum likelihood ('ML') approximations. The performance of the diagonal loading-based ('DL') and spectral norm-based ('SN') approximations have similar trend. In this simulation, they all suffer from divergence near the time instant 201s when the target starts making its first sharp turn (see Section IV.A). The underlying reason is probably that the approximations derived using this approach focus on finding, under different *analytical* criteria, the nearest positive definite matrix approximation $\hat{\mathbf{\Sigma}}_b$ to the original covariance $\mathbf{\Sigma}_b$ only. The approximation error in $\hat{\mathbf{\Sigma}}_b$ may lead to significantly deviated estimate of the mean $\hat{\boldsymbol{\mu}}_b$ (see (7)), which leads to large and greatly fluctuating estimation RMSEs observed in Figs. 1 and 2.

On the contrary, the proposed KLD minimization-based approximation ('KLD') and its special case ('UI' from [11]) offer the best estimation accuracy over the whole tracking interval. They are numerically stable as well. The greatly improved performance could be due to two aspects. First, these approximations indeed utilize a KLD-dependent cost function to calculate the approximation covariance $\tilde{\mathbf{\Sigma}}_b$ (see the discussion under (26), and Sections III.B and III.C). Second, they compute the mean $\tilde{\boldsymbol{\mu}}_b$ using (26), and $\boldsymbol{\mu}_c$ is now weighted by $(\mathbf{\Sigma}_a^{-1} + \tilde{\mathbf{\Sigma}}_b^{-1})$, rather than $\mathbf{\Sigma}_c^{-1}$ as in (7), which also comes from minimizing the KLD in (25). In other words, the KLD minimization-based approach considers the impact of the approximation error on $\tilde{\boldsymbol{\mu}}_b$ and $\tilde{\mathbf{\Sigma}}_b$ simultaneously by exploring the relationship between the product rule [12] and quotient rule (1) of Gaussian densities (see (22)).

Last, the proposed KLD minimization-based approximation ('KLD') given in (41) has similar performance as the UI-based technique ('UI') originally adopted in [3] and [11]. The reason may be found by examining (3). As each summand in (3) is an unnormalized Gaussian density, the term on the right hand side of (3) can be roughly considered as a Gaussian mixture model (GMM) with r components. To maintain computational tractability, the smoothing algorithm in [11] merges these Gaussian components into a single Gaussian density, subject to certain scaling, using the method of moment matching [29]. As a result, the difference between the KLD minimization-based and UI-based approximations can no longer be easily observed in the simulation results.

V. Conclusions

In recently proposed multiple-model smoothing algorithms, the quotient of two multivariate Gaussian densities was employed in realizing the backward-time recursion. It is expressed as an unnormalized Gaussian density if the latter has a positive definite covariance. Otherwise, approximations are needed to avoid numerical problems. This paper presented several low-complexity approximations derived using two approaches. The first approach replaces the indefinite covariance with the positive definite matrix nearest to it. The second approach finds the approximation through solving a KLD minimization problem using natural gradient. We proved that the UI technique adopted in our previous work on multiple-model smoothing is a special case of the second approach. The obtained approximations were integrated into an existing smoother to estimate the trajectory of a maneuvering target. The KLD minimization-based approach was shown to be able to provide the best empirical performance. This is because it achieves the approximation by attempting, through minimizing the KLD, to conform to the relationship between the product and quotient rules of Gaussian densities as a whole.

References

[1] J. M. Hernández-Lobato, "Balancing flexibility and robustness in machine learning: Semiparametric methods and sparse linear models," Ph.D. dissertation, Universidad Autónoma de Madrid, 2010.

[2] D. Acar and U. Orguner, "Information decorrelation for an interacting multiple model filter," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Cambridge, UK, Jul. 2018, pp. 1527–1534.

[3] X. Li, Y. Liu, L. Yang, L. S. Mihaylova, and B. Deng, "Enhanced fixed-interval smoothing for Markovian switching systems," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Sun City, South Africa, Nov. 2021.

[4] M. E. Liggins, C.-Y. Chong, I. Kardar, M. G. Alford, V. Vannicola, and S. Thomopoulos, "Distributed fusion architectures and algorithms for target tracking," *Proc. IEEE*, vol. 85, pp. 95–107, Jan. 1997.

[5] C.-Y. Chong, S. Mori, W. H. Barker, and K.-C. Chang, "Architectures and algorithms for track association and fusion," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 15, pp. 5–13, Jan. 2000.

[6] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. Conf. Uncertainty in Artificial Intelligence (UAI)*, Seattle, Washington, USA, 2001.

[7] M. Seeger, "Expectation propagation for exponential families," University of California at Berkely, Department of EECS, Tech. Rep., April 2008.

[8] R. Lopez and P. Danès, "Exploiting Rauch-Tung-Striebel formulae for IMM-based smoothing of Markovian switching systems," in *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 3953–3956.

[9] ——, "A fixed-interval smoother with reduced complexity for jump Markov nonlinear systems," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Salamanca, Spain, Jul. 2014, pp. 1–8.

[10] ——, "Low-complexity IMM smoothing for jump Markov nonlinear systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, pp. 1261–1272, Jun. 2017.

[11] Y. Liu, X. Li, L. Yang, L. S. Mihaylova, and Y. Xue, "On the fixed-interval smoothing for jump Markov nonlinear systems," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Linköping, Sweden, July 2022.

[12] S. Challa, M. R. Morelande, D. Mušicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[13] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: Theory, algorithms and software*. New York: Wiley, 2001.

[14] S. Särkkä, *Bayesian Filtering and Smoothing*. New York: Cambridge University Press, 2013.

[15] S. Särkkä and J. Hartikainen, "On Gaussian optimal smoothing of non-linear state space models," *IEEE Trans. Autom. Control*, vol. 55, no. 8, pp. 1938–1941, Aug. 2010.

[16] S. V. Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaría, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Trans. Neural Networks and Learning Systems*, vol. 23, pp. 1313–1326, Aug. 2012.

[17] L. Yang, K. Wang, and L. S. Mihaylova, "Online sparse multi-output Gaussian process regression and learning," *IEEE Trans. Signal and Information Process. over Networks*, vol. 5, pp. 258–272, Jun. 2019.

[18] G. Cornuéjols, J. Peña, and R. Tütüncü, *Optimization methods in Finance*. Cambridge University Press, 2018.

[19] C. Stein, "Lectures on the theory of estimation of many parameters," *J. of Mathematical Sciences*, vol. 34, pp. 1373–1403, July 1986.

[20] S. P. Lin and M. Perlman, "A Monte-Carlo comparison of four estimators of a covariance matrix," *Multivariate Analysis*, vol. 6, pp. 411–429, 1985.

[21] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, Feb. 2004.

[22] N. J. Higham, "Computing a nearest symmetric positive semidefinite matrix," *Linear Algebra and Its Applications*, vol. 103, pp. 103–118, May 1988.

[23] M. Tanaka and K. Nakata, "Positive definite matrix approximation with condition number constraint," *Optimization Letters*, vol. 8, pp. 939–947, March 2014.

[24] J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam, "Condition-number-regularized covariance estimation," *Journal of the Royal Statistical Society, Series B*, vol. 75, pp. 427–450, June 2013.

[25] R. H. Farrell, *Multivariate Calculation*. Springer-Verlag, 1985.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[27] K. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[28] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Prentice Hall, 1995.

[29] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[30] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, Feb. 1998.

[31] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, May 2013.

[32] M. E. Khan and D. Nielsen, "Fast yet simple natural-gradient descent for variational inference in complex models," in *Proc. Intl. Symposium on Information Theory and Its Applications (ISITA)*, Singapore, Oct. 2018.

[33] M. E. Khan and W. Lin, "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models," in *Proc. Intl. Conf. Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, April 2017.

[34] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava, "Fast and scalable Bayesian deep learning by weight-perturbation in Adam," in *Proc. Intl. Conf. Machine Learning (ICML)*, Stockholm, Sweden, July 2018, pp. 2611–2620.

[35] W. J. Wilkinson, S. Särkkä, and A. Solin, "Bayes-Newton methods for approximate Bayesian inference with PSD guarantees," *arXiv preprint*, Dec. 2022. [Online]. Available: arXiv.org/abs/2111.01721.

[36] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.