SDS-12

Pearson $\chi 2$ -divergence Approach to Gaussian Mixture Reduction and its Application to Gaussian-sum Filter and Smoother

Genshiro Kitagawa

October 2019

Statistics & Data Science Series back numbers:
http://www.mims.meiji.ac.jp/publications/datascience.html

Pearson χ^2-divergence Approach to Gaussian Mixture Reduction and its Application to Gaussian-sum Filter and Smoother

Genshiro Kitagawa
Mathematics and Informatics Center, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN

Abstract

The Gaussian mixture distribution is important in various statistical problems. In particular it is used in the Gaussian-sum filter and smoother for linear state-space model with non-Gaussian noise inputs. However, for this method to be practical, an efficient method of reducing the number of Gaussian components is necessary. In this paper, we show that a closed form expression of Pearson χ^2-divergence can be obtained and it can apply to the determination of the pair of two Gaussian components in sequential reduction of Gaussian components. By numerical examples for one dimensional and two dimensional distribution models, it will be shown that in most cases the proposed criterion performed almost equally as the Kullback-Libler divergence, for which computationally costly numerical integration is necessary. Application to Gaussian-sum filtering and smoothing is also shown.

Keywords: Gaussian mixture model (GMM), Gaussian mixture reduction, Kullback-Leibler Divergence, Pearson χ^2-divergence, Gaussian-sum filter.

1 Introduction

Reduction of the number of components in Gaussian mixture distribution is important in various field of statistical problems, data fusion, pattern recognition, supervised learning of multimedia and target tracking[8],[10]. As an example, consider a linear state space model

$$\begin{aligned} x_n &= F_n x_{n-1} + G_n v_n \\ y_n &= H_n x_n + w_n, \end{aligned} \tag{1}$$

where the system noise v_n and the observation noise w_n are distributed according to a mixture of several Gaussian components:

$$\begin{aligned} p(v_n) &= \sum_{i=1}^{q} \alpha_i \varphi(v_n|\mu_v, Q_i) \\ p(w_n) &= \sum_{j=1}^{r} \beta_j \varphi(w_n|\mu_w, R_j). \end{aligned} \tag{2}$$

q and r are the number of Gaussian components of $p(v)$ and $p(w)$, respectively, and $\varphi(x|\mu, V)$ denotes the Gaussian density with mean vector μ and the variance covariance matrix V.

Here assume that Y_n denotes the set of observations up to time n, i.e., $Y_n = \{y_1, \ldots, y_n\}$. The prediction problem is to obtain, $p(x_n|Y_{n-1})$, the conditional distribution of x_n given Y_{n-1}, and the filter problem is to obtain, $p(x_n|Y_n)$, the conditional distribution of x_n given Y_n. For the linear state-space model with Gaussian mixture noise, it is known that these conditional distributions are also given as the mixture of Gaussian densities[1],[4],[5],[9]:

$$\begin{aligned} p(x_n|Y_{n-1}) &= \sum_{i=1}^{q}\sum_{k=1}^{\ell_{n-1}} \alpha_i \gamma_{k,n-1} \varphi(x_n|x_{n|n-1}^{ik}, V_{n|n-1}^{ik}) = \sum_{j=1}^{m_n} \delta_{jn} \varphi(x_n|x_{n|n-1}^{j}, V_{n|n-1}^{j}) \\ p(x_n|Y_n) &= \sum_{j=1}^{r}\sum_{k=1}^{m_n} \gamma_{jk,n} \varphi(x_n|x_{n|n}^{jk}, V_{n|n}^{jk}) = \sum_{i=1}^{\ell_n} \gamma_{in} \varphi(x_n|x_{n|n}^{i}, V_{n|n}^{i}) \end{aligned} \tag{3}$$

where $m_n = q \times \ell_{n-1}$, $\delta_{jn} = \alpha_i \gamma_{k,n-1}$, $\ell_n = r \times m_n$ and $\gamma_{jk,n} = \beta_j \delta_{kn} \varphi(y_n|x_{n|n-1}^{jk}, V_{n|n-1}^{jk})$.

The Gaussian-sum filter is an algorithm to obtain these conditional densities recursively with time. The advantage of the Gaussian-sum filter is that the parameters of the state distributions such as δ_{jn}, γ_{in}, $x_{n|t}$, and $V_{n|t}$ are obtained by running the Kalman filters in parallel. Therefore, the computation is easy and can yield accurate results. However, there is a severe difficulties with this method. Namely, the numbers of Gaussian components, m_n and ℓ_n, increase by $q \times r$ times at each time step of the filtering. Therefore, the number of Gaussian components would increase exponentially over time, and for this filtering method to be practical, a computationally efficient method for the reduction of the number of Gaussian components is indispensable.

In principle, reduction of the number of Gaussian components can be realized by minimizing the Kullback-Leibler divergence of the full-order Gaussian mixture distribution with respect to the reduced-order Gaussian mixture distribution. However, as we discussed later in Section 2, two problems make this method impractical. Therefore, as a practical measure, we usually reduce the number of Gaussian components successively. In this paper, we refer to this method as the sequential reduction method and consider criteria for selecting a pair of Gaussian components to be merged.

Kitagawa[4][5] used a weighted Kullback-Leibler divergence of two candidate Gaussian components. Salmond[8] proposed a mixture reduction algorithm in which the number of components is reduced by repeatedly choosing the two components that appear to be most similar to each other. Williams and Maybeck[11] proposed a mixture reduction algorithm based on an integrated squared difference (ISD) similarity measure, which has the big advantage that the similarity between two arbitrary Gaussian mixtures can be expressed in closed form. Runnalls[7] proposed a measure of similarity between two components based on the upper bound of the increase of Kullback-Leibler (KL) discrimination measure when a pair of two Gaussian components are merged. In this paper, we propose use of Pearson χ^2-divergence of two Gaussian components for which we can derive a closed form expression for the criterion to select the pair of Gaussian components to be merged.

In section 2, we define the Gaussian mixture reduction problem and briefly show some reduction methods. In section 3, a sequential reduction method based on Pearson χ^2-divergence will be introduced, in which the criteria for selecting a pair of indices to be merged can be obtained in explicit analytical form. In section 4, emperical studies on the sequential reduction of the number of Gaussian components are shown, using one-dimensional and two-dimensional Gaussian mixture distributions. Section 5 deals with the application of the sequential Gaussian-mixture reduction method to the a Gaussianm-sum filtering and smoothing for linear state-space model with Gaussian-mixture noise inputs. We conclude in Section 6. Details of the derivation of the Pearson χ^2-divergence is shown in Appendix.

2 Reduction of Gaussian Components

2.1 Reduction based on Kullback-Leibler Discrimination

The Kullback-Leibler divergence is the most frequently used to evaluate the dissimilarity between true distribution and an approximated distribution, which is defined by

$$I(g(x);f(x)) = \int \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx = \int \log\{g(x)\}\, g(x)dx - \int \log\{f(x)\}\, g(x)dx, \tag{4}$$

where in the context of the Gaussian mixture approximation, $g(x)$ is the full-order mixture model and $f(x)$ is the reduced order model ($\ell < m$):

$$g(x) = \sum_{i=1}^{m} \alpha_i \varphi(x|\xi_i, V_i) \tag{5}$$

$$f_\ell(x) = \sum_{i=1}^{\ell} \beta_i \varphi(x|\mu_i, \Sigma_i). \tag{6}$$

Hereafter, for simplicity of the notation, the number of Gaussian components is referred to as the order.

In principle, the best reduced order model can be obtained by minimizing the Kullback-Leibler divergence. However, there are two problems with this method. Firstly, except for simple densities such as Gaussian density, the KL-divergence does not have a closed expression. So we need to apply numerical integration to evaluate the KL-divergence. Secondly, to estimate the parameters of the best reduced order model, we need to apply numerical optimization in high dimensional parameter space. Therefore, at least for recursive filtering in which this reduction process is repeated as long as a new observation is obtained, this method is impractical.

2.2 Sequential Reduction

Therefore, we usually apply a sequential reduction method. Assume that the full-order model and an approximated reduced order model are respectively defined by

$$\begin{aligned} g(x) &= \sum_{i=1}^{m} w_i \varphi(x|\xi_i, U_i) \\ f_\ell(x) &= \sum_{i=1}^{\ell} \alpha_i \varphi(x|\mu_i, \Sigma_i). \end{aligned} \tag{7}$$

In the sequential reduction method, to further reduce the number of components, we select a pair of two components, say j and k, and pool these two densities. The reduced order model is defined by

$$h_{jk}(x) = \sum_{i \notin \{j,k\}} \alpha_i \varphi(x|\mu_i, \Sigma_i) + (\alpha_j + \alpha_k)\varphi(x|\zeta_{jk}, V_{jk}) \tag{8}$$

where $\varphi(x|\zeta_{jk}, V_{jk})$ is the merged density whose parameters are usually determined so that the first two moments of the distributions are preserved:

$$\begin{aligned} \xi_{jk} &= (\alpha_j + \alpha_k)^{-1}(\alpha_j \mu_j + \alpha_k \mu_k) \\ V_{jk} &= (\alpha_j + \alpha_k)^{-1}\left[\alpha_j\left\{\Sigma_j + (\mu_j - \xi_{jk})(\mu_j - \xi_{jk})^T\right\} + \alpha_k\left\{\Sigma_k + (\mu_j - \xi_{jk})(\mu_j - \xi_{jk})^T\right\}\right]. \end{aligned} \tag{9}$$

The indices of two pooled densities, j and k, are selected so that a properly determined criterion is minimized. By repeating this process, we can obtain a Gaussian mixture approximation of $g(x)$ with a smaller number of Gaussian components.

For selecting a pair of two densities, many ad hoc criteria have been proposed so far. Kitagawa(1989,1994) used the weighted KL-divergence of Gaussian components

$$D(k,j) = \alpha_k \alpha_j \left\{ \Sigma_k^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_k + (\mu_k - \mu_j)^T (\Sigma_k^{-1} + \Sigma_j^{-1})(\mu_k - \mu_j) \right\}. \tag{10}$$

Salmond(1990) proposed the increase of within-component variance

$$D_s^2(k,j) = \mathrm{tr}(\Sigma^{-1}\Delta W), \quad \Delta W(\varphi_k, \varphi_j) = \frac{\alpha_k \alpha_j}{\alpha_k + \alpha_j}(\mu_k - \mu_j)(\mu_k - \mu_j)^T. \tag{11}$$

Williams and Mayback (2003) used a squared difference of two densities

$$J(g,f) = \int (g(x) - f(x))^2 dx. \tag{12}$$

Runnalls(2006) used the upper bound of the increase of KL-divergence by pooling two densities:

$$B(k,j) = \frac{1}{2}\left\{ (\alpha_k + \alpha_j)\log\det(V_{kj}) - \alpha_k \log\det(\Sigma_k) - \alpha_j \log\det(\Sigma_j) \right\} \tag{13}$$

and it is reported that this criterion mitigated some anomalous behavior in certain circumstances of the ones by Williums and Mayback[11] and Salmond[8], and provide us with a reasonable reduction result[7].

3 Reduction Criterion based on Pearson χ^2-Divergence

3.1 Pearson χ^2-Divergence of Two gaussian Mixture Models

In this paper, we consider the use of Pearson χ^2-divergence:

$$D_{\chi^2}(q;p) = \int \left(\frac{q(x)}{p(x)} - 1 \right)^2 p(x) dx = \int \frac{q(x)^2}{p(x)} dx - 1. \tag{14}$$

Assume that $q(x)$ is a mixture of two Gaussian densities

$$q(x) = \alpha_j \varphi(x|\mu_j, \Sigma_j) + \alpha_k \varphi(x|\mu_k, \Sigma_k), \quad \alpha_j + \alpha_k = 1 \tag{15}$$

and $p(x)$ is a pooled Gaussian density, $p_{jk}(x) = \varphi(x|\zeta_{jk}, W_{jk})$, obtained by the moment preserving merge where ζ_{jk} and V_{jk} are given in (9). Then the Pearson χ^2-divergence $D_{\chi^2}(j,k)$ of the mixture of two Gaussian densities with respect to the merged density is obtained by

$$\begin{aligned} D_{\chi^2}(j,k) &= \int \frac{q(x)^2}{p_{jk}(x)} dx - 1 \\ &= \alpha_j^2 \int \frac{f_j(x)^2}{p_{jk}(x)} dx + 2\alpha_j \alpha_k \int \frac{f_j(x) f_k(x)}{p_{jk}(x)} dx + \alpha_k^2 \int \frac{f_k(x)^2}{p_{jk}(x)} dx - 1. \end{aligned} \tag{16}$$

Here, since the densities $f_j(x)$, $f_k(x)$ and p_{jk} are respectively defied by

$$\begin{aligned} f_j(x) &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_j\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\} \\ f_k(x) &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\} \\ p_{jk}(x) &= (2\pi)^{-\frac{k}{2}} \left|V_{jk}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\zeta_{jk})^T V_{jk}^{-1}(x-\zeta_{jk})\right\}, \end{aligned} \tag{17}$$

the integrand of the second term of the right hand side of the equation (16) is given by

$$\begin{aligned} \frac{f_j(x)f_k(x)}{p_{jk}(x)} &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_j\right|^{-\frac{1}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \left|V_{jk}\right|^{\frac{1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \frac{1}{2}(x-\xi_{jk})^T V_{jk}^{-1}(x-\xi_{jk})\right\} \\ &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_j\right|^{-\frac{1}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \left|V_{jk}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\zeta_{jk}-\eta_{jk})^T(V_{jk}-\Sigma_{jk})^{-1}(\zeta_{jk}-\eta_{jk})\right\} \exp\left\{-\frac{1}{2}(x-\eta_{jk})^T W_{jk}(x-\eta_{jk})\right\} \end{aligned} \tag{18}$$

where $\Sigma_{jk} = (\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}$, $W_{jk} = \Sigma_j^{-1}+\Sigma_k^{-1}-V_{jk}^{-1}$ and $\eta_{jk} = (\Sigma_j^{-1}+\Sigma_k^{-1}-V_{jk}^{-1})^{-1}((\Sigma_j^{-1}+\Sigma_k^{-1})\zeta_{jk} - V_{jk}^{-1}\xi_{jk})$. The details of the derivation of the last equality of (18) is given in the appendix.

Then, by integrating over the whole domain of the distribution, we obtain

$$\begin{aligned} \int \frac{f_j(x)f_k(x)}{p_{jk}(x)}dx &= \left|\Sigma_j\right|^{-\frac{1}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \left|V_{jk}\right|^{\frac{1}{2}} \left|W_{jk}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\zeta_{jk}-\eta_{jk})^T(V_{jk}-\Sigma_{jk})^{-1}(\zeta_{jk}-\eta_{jk})\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\}. \end{aligned} \tag{19}$$

The expression for the first and the third term of (16) is obtained by putting by $f_k(x) = f_j(x)$; namely, $\mu_k = \mu_j$ and $\Sigma_k = \Sigma_j$.

$$\int \frac{f_j(x)^2}{p_{jk}(x)}dx = \left|\Sigma_j\right|^{-1} \left|V_{jk}\right|^{\frac{1}{2}} \left|\bar{W}_j\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mu_j-\eta_{jk})^T\bar{W}_j^{-1}(\mu_j-\eta_{jk})\right\} \tag{20}$$

where $\bar{W}_j = 2\Sigma_j^{-1} - V_{jk}^{-1}$, $\eta_j = (2\Sigma_j^{-1}-V_{jk}^{-1})^{-1}(2\Sigma_j^{-1}\mu_j - V_{jk}^{-1}\xi_{jk})$.

3.2 Proposed Reduction Criterion

Therefore the Pearson χ^2-divergence for the Gaussian mixture reduction is obtained by

$$\begin{aligned} D_{\chi^2}(j,k) &= \alpha_j^2 \left|\Sigma_j\right|^{-1} \left|V_{jk}\right|^{\frac{1}{2}} \left|\bar{W}_j\right|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}(\mu_j-\xi_{jk})^T(V_{jk}-\frac{1}{2}\Sigma_j)^{-1}(\mu_j-\xi_{jk})\right\} \\ &+ \alpha_k^2 \left|\Sigma_k\right|^{-1} \left|V_{jk}\right|^{\frac{1}{2}} \left|\bar{W}_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mu_k-\xi_{jk})^T(V_{jk}-\frac{1}{2}\Sigma_k)^{-1}(\mu_k-\xi_{jk})\right\} \\ &+ 2\alpha_i\alpha_j \left|\Sigma_j\right|^{-\frac{1}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \left|V_{jk}\right|^{\frac{1}{2}} \left|W_{jk}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\zeta_{jk}-\xi_{jk})^T(V_{jk}-\Sigma_{jk})^{-1}(\zeta_{jk}-\xi_{jk})\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\} - 1 \end{aligned} \tag{21}$$

In the sequential reduction based on this criterion, $D_{\chi^2}(j,k)$ are evaluated for $j = 1,...,\ell-1$ and $k = 2,...,\ell$ and find the pair (j^*,k^*) that satisfies

$$D_{\chi^2}(j^*,k^*) = \min_{j,k} D_{\chi^2}(j,k). \tag{22}$$

Then the two Gaussian components $\varphi(x|\mu_j^*,\Sigma_j^*)$ and $\varphi(x|\mu_k^*,\Sigma_k^*)$ are merged and we obtain the Gaussian mixture model with $\ell-1$ components. Repeating this process, it is possible to obtain Gaussian mixture distribution with a specific order.

The problem with this Pearson χ^2-divergence is that $q(x)/p(x)$ may become unbounded. Therefore, in using this as the criterion for selecting the pair for merging, we need a safe-guard in computation. Namely, we exclude the pair j and k from the merging candidate.

4 Empirical Study: Comparison of Reduction Methods

Many criteria have been proposed for selecting a pair of Gaussian components in sequential reduction of Gaussian components. In this section we compare the following criteria:

1. Weighted KL-divergence of Gaussian components, Kitagawa (1989, 1994):

$$D(j,k) = \alpha_j\alpha_k\left\{\Sigma_j^{-1}\Sigma_k + \Sigma_j^{-1}\Sigma_k + (\mu_j-\mu_k)^T(\Sigma_j^{-1}+\Sigma_k^{-1})(\mu_j-\mu_k)\right\} \tag{23}$$

2. Upper bound of the increase of KL-divergence, Runalls (2006):

$$B(j,k) = \frac{1}{2}\left\{(\alpha_j+\alpha_k)\log\det(V_{jk}) - \alpha_j\log\det(\Sigma_j) - \alpha_k\log\det(\Sigma_k)\right\} \tag{24}$$

3. χ^2-divergence proposed in this paper: $D_{\chi^2}(j,k)$

Beside these ad hoc criteria, we also considered the following two reduction methods based on the Kullback-Leibler divergence.

4. The sequential reduction based on the Kullback-Leibler divergence of the pooled model obtained by numerical integration:

$$I(g;f_{jk}) = \int \log g(x)g(x)dx - \int \log f_{jk}(x)g(x)dx. \tag{25}$$

5. The global Kullback-Leibler divergence minimization method. Note that this method requires both numerical integration and numerical optimization:

$$I(g;\hat{f}_{jk}) = \int \log g(x)g(x)dx - \int \log \hat{f}_{jk}(x)g(x)dx, \tag{26}$$

where the parameters of $f_{jk}(x)$ are estimated by minimizing $I(g;f_{jk})$. Therefore, this method is very computationally costly and is feasible only for very low dimensional distributions.

Table 1: Assumed one-dimensional Gaussian-mixture distribution with 16 components.

i	α_i	μ_i	Σ_i
1	0.30	0.0	0.5
2	0.15	5.0	1.0
3	0.15	-4.0	1.0
4	0.05	0.2	9.0
5	0.05	-1.5	2.0
6	0.0686	1.03982	4.39842
7	0.03472	-1.55209	3.78821
8	0.07578	-1.35090	2.78963
9	0.00101	-0.25711	1.18460
10	0.00011	2.00426	1.14186
11	0.01699	1.44357	1.00000
12	0.00003	-2.15010	1.02979
13	0.05787	-0.58808	1.21395
14	0.00039	1.57966	1.35196
15	0.02193	1.87170	1.12458
16	0.02257	0.55285	1.05299

4.1 One-dimensional Distributions

Table 1 shows the assumed full-order Gaussian mixture model with 16 Gaussian components. Table 2 and Figure 1 show the increase of KL-divergence when the reduced order models are obtained by five methods. In the figure, grey line shows the results by Runnalls, green one by Kitagawa, blue one by Pearson χ^2-divergence, yellow one sequential reduction by Kullback-Leibler divergence, and red one by global optimization of Kullback-Leibler divergence. It can be seen that the sequential reduction based on Pearson χ^2-divergence yields almost the same performance as the sequential reduction by Kullback-Leibler divergence.

The accuracy of the sequential reduction methods are worth by one or two digit than the optimal model. However, the figure also indicates that by using a larger order m, we can attain a similar accuracy as the optimal model.

Figure 2 shows the comparison of the densities obtained by the sequential reduction and the global optimization method. In these plots, the red curve shows the true full order density, the green one the optimal reduced order model obtained by minimizing the KL-divergence, and the purple one obtained by the sequential reduction based on the Pearson χ^2-divergence. It can be seen that for $m \geq 8$, the green curve and purple curve are visually indistinguishable. But for m=2 and 3, they are considerably different.

4.2 Two-dimensional Distributions

In this example, the true 2-dimensional density is expressed by 10 Gaussian distributions shown in Table 3. Table 4 and Figure 8 show the Kullback-Leibler divergence of the true mixture model with respect to the reduced order model obtained by 5 methods. It can be seen that, except for ℓ=2 and 3, the results by the Pearson χ^2-divergence is almost indistinguishable with the method based on Kullback-Leibler divergence.

Table 2: Change of KL-divergence of true with respect to the reduced order models by various reduction methods.

m	Runnalls	Kitagawa	Pearson	KL-div.	Optimal
15	1.43×10^{-12}	2.18×10^{-11}	9.80×10^{-14}	3.00×10^{-13}	3.80×10^{-14}
14	1.21×10^{-09}	5.98×10^{-10}	1.43×10^{-12}	4.63×10^{-12}	2.94×10^{-13}
13	1.17×10^{-07}	1.74×10^{-08}	4.85×10^{-11}	3.40×10^{-11}	2.63×10^{-12}
12	1.54×10^{-07}	7.55×10^{-08}	6.53×10^{-10}	3.24×10^{-11}	3.96×10^{-11}
11	1.24×10^{-06}	4.85×10^{-07}	1.54×10^{-07}	1.78×10^{-09}	1.82×10^{-09}
10	0.00010181	1.18×10^{-06}	7.45×10^{-07}	3.70×10^{-09}	4.82×10^{-09}
9	0.00010793	1.24×10^{-05}	2.60×10^{-06}	1.08×10^{-08}	6.15×10^{-09}
8	0.00013676	0.00022274	1.23×10^{-05}	1.67×10^{-08}	8.84×10^{-09}
7	0.00033167	0.00022197	0.0001042	2.88×10^{-07}	2.55×10^{-07}
6	0.00175442	0.00031239	6.90×10^{-05}	2.66×10^{-07}	2.57×10^{-07}
5	0.0040189	0.00110572	0.00035793	1.61×10^{-06}	2.57×10^{-07}
4	0.0060584	0.00076506	0.00076506	0.00024942	0.00024942
3	0.02886692	0.0331135	0.01810894	0.01650889	0.00435254
2	0.08941172	0.07007295	0.07938004	0.06884198	0.06884198
1	0.13589858	0.1304686	0.1304686	0.1304686	0.1304686

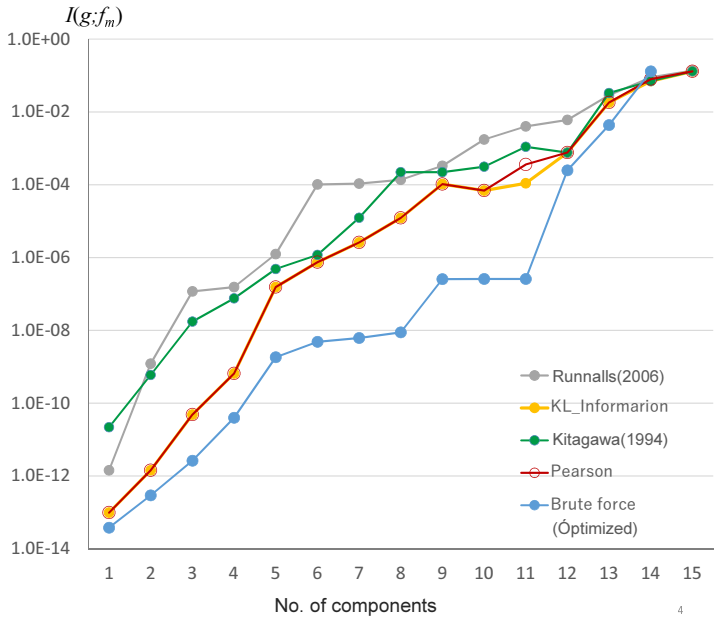

Figure 1: Change in KL-divergence of true, sequentially reduced and optimal reduced order models.

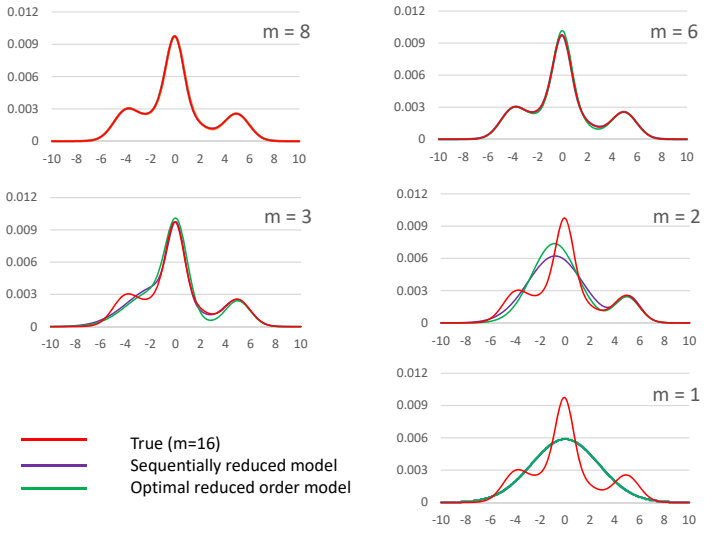

Figure 2: The comparison of the densities obtained by the sequential reduction and the global optimization method.

Table 3: Assumed twe-dimensional Gaussian-mixture distribution with 16 terms.

i	α_i	$\mu_i(1)$	$\mu_i(2)$	$\Sigma_i(1,1)$	$\Sigma_i(2,2)$	$\Sigma_i(2,1)$
1	0.30	0	0	1	1	0
2	0.20	2	0	4	2	0
3	0.16	3	3	2	2	-0.5
4	0.11	-4	-4	4	4	2
5	0.08	-1	1	9	9	4.0
6	0.06	2	-4	4	9	2
7	0.04	0	2	4	1	-0.5
8	0.03	-2	4	9	9	0
9	0.01	-2	0	2	1	0
10	0.01	1	-2	1	1	0

Table 4: Change of KL-divergence of true model with respect to the reduced order models by various reduction methods: Two dimensional case.

m	Runnalls	Kitagawa	Pearson	KL-div.	Optimal
9	0.000220	0.000143	0.000163	0.000143	0.000022
8	0.000656	0.000849	0.000300	0.000300	0.000093
7	0.002367	0.001812	0.001051	0.001051	0.000258
6	0.004783	0.003920	0.002010	0.002010	0.000496
5	0.006878	0.023910	0.005754	0.005754	0.002862
4	0.029877	0.029670	0.014775	0.014775	0.004916
3	0.056387	0.034783	0.079955	0.039786	0.029775
2	0.099586	0.099586	0.122572	0.091505	0.084608
1	0.180119	0.180119	0.180119	0.180119	0.180119

Figures 4 and 5 show the contour and the bird's-eye views of the reduced order Gaussian mixture models obtained by the Pearson χ^2-divergence.

Summarizing the two examples, there are three types of reduction methods, namely the sequential reduction by ad-hoc criterion, Sequential reduction by KL-divergence and global KL-divergence minimization. Obviously the accuracy increases in this order, but computational cost increases. So the suggestion is to estimate a mixture model with a slightly larger number of components by the sequential reduction method.

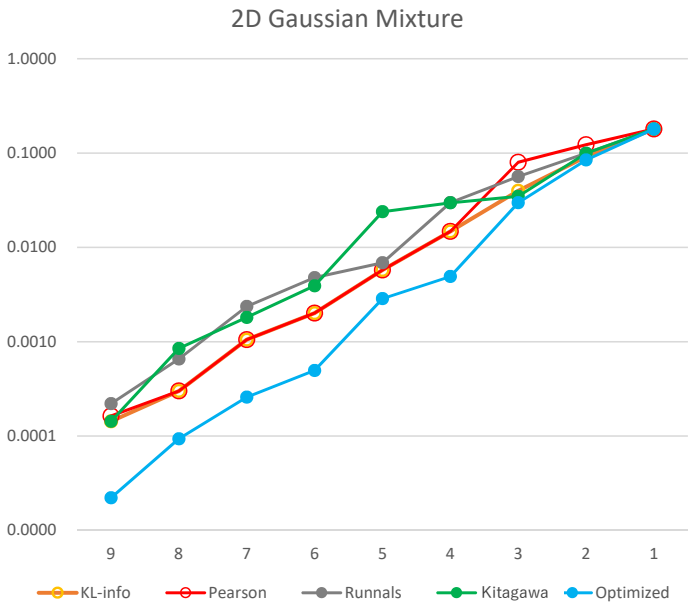

Figure 3: Change in KL-divergence of true and optimal reduced order model.

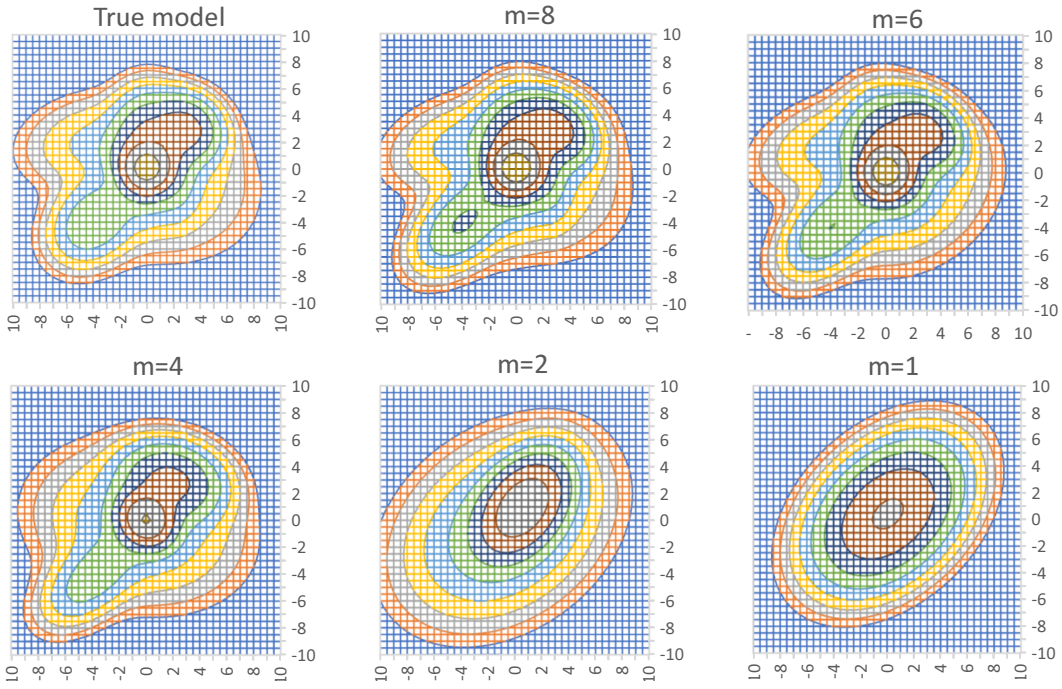

Figure 4: Contour of 2D densities obtained from the full-order Gaussian-mixture and reduced order Gaussian-mixture models.

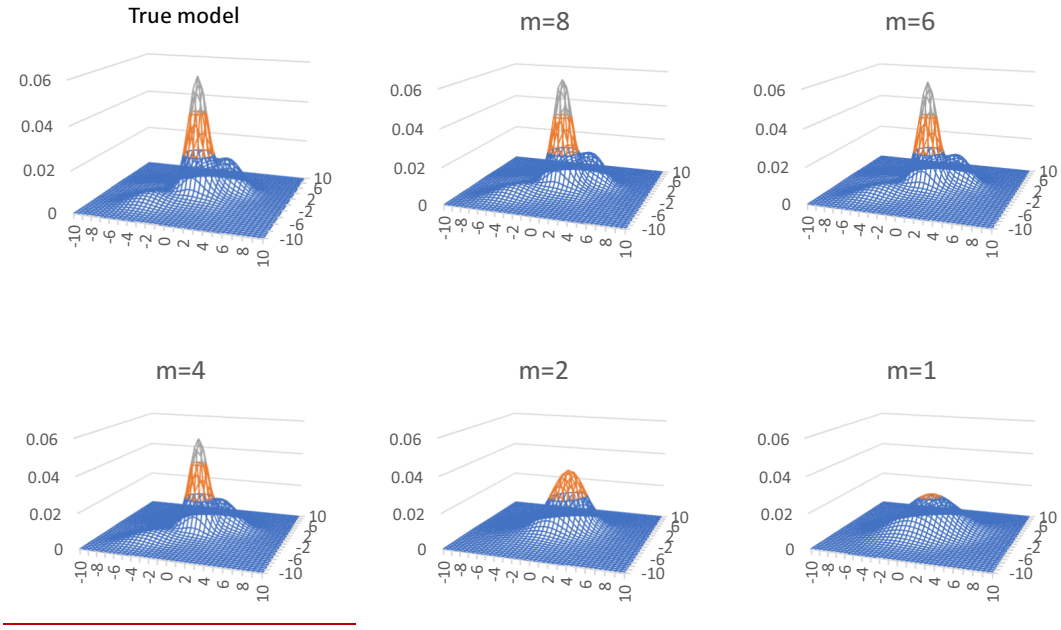

Figure 5: Bird's-eye-views of 2D densities obtained from the full-order Gaussian-mixture and reduced order Gaussian-mixture models.

5 Non-Gaussian Smoothing

We consider the application of Gaussian-sum filter and smoother to the detection of the level shift in the time series. The top-left plot of Figure 6 shows the example data analyzed in Kitagawa[3]. For estimation of the trend of the series, we consider a simple state-space model.

$$\begin{aligned} x_n &= x_{n-1} + v_n \\ y_n &= x_n + w_n. \end{aligned} \tag{27}$$

Here we assume that the observation noise is Gaussian but the system noise is a mixture of two Gaussian distributions:

$$\begin{aligned} v_n &\sim \alpha N(0,\tau^2) + (1-\alpha)N(0,\xi^2) \\ w_n &\sim N(0,\sigma^2), \end{aligned} \tag{28}$$

where $\sigma^2 = 1.027$, $\tau^2 = 0.000254$, $\xi^2 = 1.189$ and $\alpha = 0.989$.

Figure 6 show the estimates of the trend by the Non-Gaussian smoother [3] and the particle smoother [6]. Table 5 shows the log-likelihoods and the cpu-times for various number of the maximum number of Gaussian components approximating the state densities. At least in this case $M = 8$ or 16 looks sufficient. The cpu-time is less than 1 second for filtering.

Figure 7 shows the smoothed distribution of the trend obtained by the Gaussian-sum smoother for the number of components m=1, 2, 4 and 128. The top-left plot shows the case $m = 1$, bottom-left shows case $m = 2$, top-right $m = 4$ and bottom-right $m = 128$. At least visually the results by $m = 4$ and 128 are almost

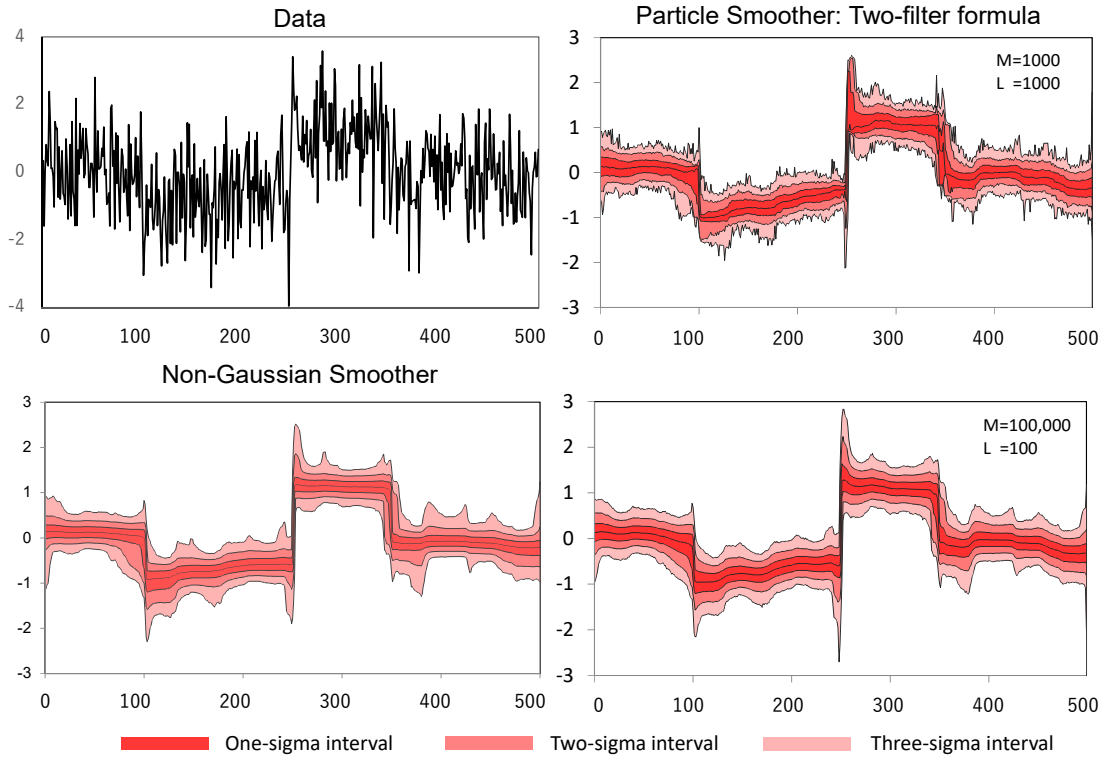

Figure 6: Test data and the estimated trends obtained by the non-Gaussian smoother and the particle filter with m=1000 and 100,000.

Table 5: Gaussian-sum filters and smoothers for various number of Gaussian components.

		cpu time (in second)	
m	log-lk	Filtering	Smoothing
1	-741.930	0.00	0.08
2	-741.047	0.02	0.23
4	-740.816	0.02	0.94
8	-740.748	0.05	3.70
16	-740.702	0.27	14.85
32	-740.704	1.86	59.53
64	-740.704	14.26	243.47
128	-740.704	112.51	1018.20

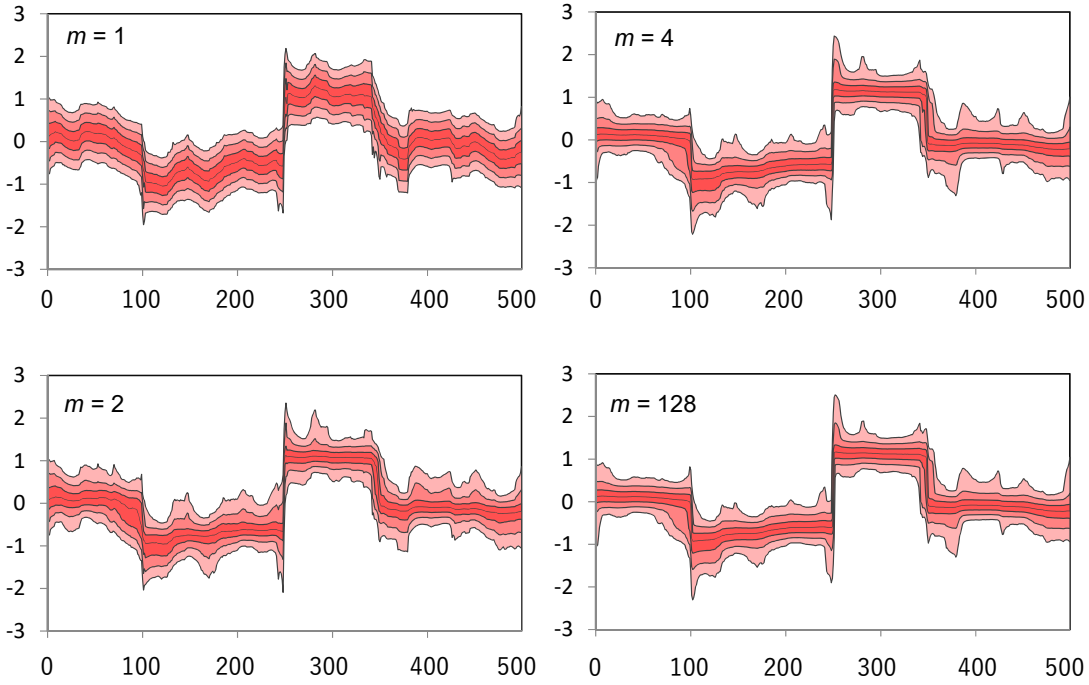

Figure 7: Estimated trends by the Gaussian-sum smoother with number of Gaussian components, m=1,2,4 and 128.

indistinguishable. This indicates that the Gaussian-sum filter is very efficints for linear state space model with Gaussian-mixture noise inputs in the sense that it can provide a very accurate approximation to the posterior distribution of the state.

It is interesting to note that as seen in Figure 8 the Gaussian-sum smoother with $m = 1$ is different from the Kalman smoother.

6 Conclusion

Pearson χ^2-divergence of two Gaussian components with respect to the merged single Gaussian distribution has an explicit analytical form. According to the empirical studies, sequential reduction method based on the Pearson χ^2-divergence performed almost similarly as the one based on the Kullback-Leibler divergence for which computationally costly numerical integration is necessary. Application to Gaussian-sum filter and smoother is shown and it is shown that Gaussian-sum filtering method is very efficient for linear state-space model with Gaussian mixture noise inputs.

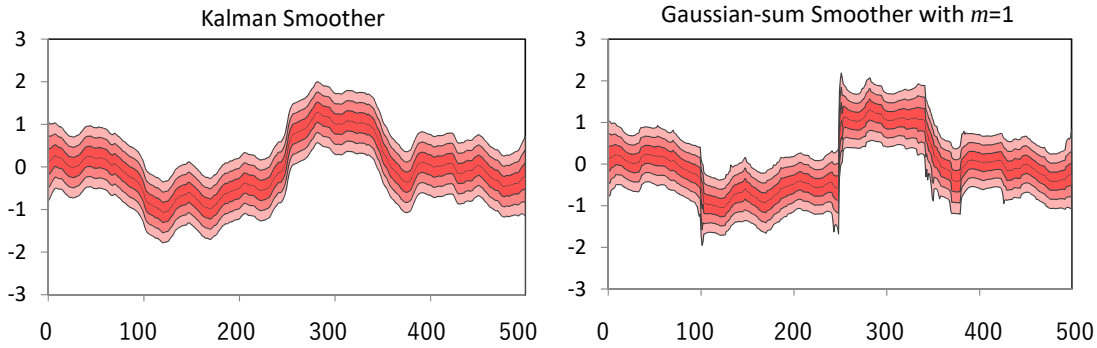

Figure 8: Comparison with Kalman smoother and Gaussian-sum smoother with one component (m=1).

7 Appendix

In this appendix, it will be shown that

$$\begin{aligned}\int \frac{f_j(x)f_k(x)}{p_{jk}(x)}dx &= (2\pi)^{-\frac{k}{2}}\left|\Sigma_j\right|^{-\frac{1}{2}}\left|\Sigma_k\right|^{-\frac{1}{2}}\left|V_{jk}\right|^{\frac{1}{2}}\left|W_{jk}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\} \\ &\quad\times\exp\left\{-\frac{1}{2}(\zeta_{jk}-\eta_{jk})^T(V_{jk}-\Sigma_{jk})^{-1}(\zeta_{jk}-\eta_{jk})\right\} \qquad (29)\end{aligned}$$

which is used in the derivation of the equation (19).

Notations

$$\begin{aligned}\Sigma_{jk}^{-1} &= \Sigma_j^{-1}+\Sigma_k^{-1}, \quad \Sigma_{jk}=(\Sigma_{jk}^{-1})^{-1}=(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}, &(30)\\ \xi_{jk} &= (\alpha_j+\alpha_k)^{-1}(\alpha_j\mu_j+\alpha_k\mu_k) &(31)\\ V_{jk} &= (\alpha_j+\alpha_k)^{-1}\left[\alpha_j\left\{\Sigma_j+(\mu_j-\xi_{jk})(\mu_j-\xi_{jk})^T\right\}+\alpha_k\left\{\Sigma_k+(\mu_j-\xi_{jk})(\mu_j-\xi_{jk})^T\right\}\right] &(32)\\ W_{jk} &= \Sigma_j^{-1}+\Sigma_k^{-1}-V_{jk}^{-1}=\Sigma_{jk}^{-1}-V_{jk}^{-1}, &(33)\\ \zeta_{jk} &= (\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k) &(34)\\ \Sigma_{jk}^{-1}\zeta_{jk} &= \Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k, &(35)\\ \eta_{jk} &= \left(\Sigma_j^{-1}+\Sigma_k^{-1}-V_{jk}^{-1}\right)^{-1}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k-V_{jk}^{-1}\xi_{jk}\right), \\ &= \left(\Sigma_{jk}^{-1}-V_{jk}^{-1}\right)^{-1}\left(\Sigma_{jk}^{-1}\zeta_{jk}-V_{jk}^{-1}\xi_{jk}\right), &(36)\\ \bar{W}_j &= 2\Sigma_j^{-1}-V_{jk}^{-1}, &(37)\\ \zeta_j &= (2\Sigma_j^{-1})^{-1}(2\Sigma_j^{-1}\mu_j)=\Sigma_j\Sigma_j^{-1}\mu_j=\mu_j &(38)\\ \eta_j &= (2\Sigma_j^{-1}-V_{jk}^{-1})^{-1}(2\Sigma_j^{-1}\mu_j-V_{jk}^{-1}\xi_{jk}). &(39)\end{aligned}$$

Hereafter in this appendix, for the simplicity of the notation, the suffix $_{jk}$ is omitted, namely we denote $\xi_{jk}=\xi$, $V_{jk}\equiv V$, $\Sigma_{jk}\equiv\Sigma$, $\zeta_{jk}\equiv\zeta$, $\xi_{jk}\equiv\xi$, $\eta_{jk}=\eta$.

Matrix Lemma

$$
\begin{aligned}
(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1} &= \Sigma_j-\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\Sigma_j, && (40)\\
(\Sigma^{-1}-V^{-1})^{-1} &= \Sigma(V-\Sigma)^{-1}V, && (41)\\
(\Sigma_j^{-1}+\Sigma_k^{-1}-V^{-1})^{-1} &= (\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}(V-(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1})^{-1}V \\
&= \left\{\Sigma_j-\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\Sigma_j\right\}(V-\Sigma)^{-1}V && (42)\\
V^{-1}-V^{-1}\Sigma(V-\Sigma)^{-1} &= V^{-1}(V-\Sigma)(V-\Sigma)^{-1}-V^{-1}\Sigma(V-\Sigma)^{-1}=(V-\Sigma)^{-1} && (43)\\
\Sigma^{-1}-(V-\Sigma)^{-1}V\Sigma^{-1} &= (V-\Sigma)^{-1}(V-\Sigma)\Sigma^{-1}-(V-\Sigma)^{-1}V\Sigma^{-1}=-(V-\Sigma)^{-1} && (44)\\
V^{-1}\Sigma(V-\Sigma)^{-1}V\Sigma^{-1} &= \{\Sigma V^{-1}(V-\Sigma)\Sigma^{-1}V\}^{-1}=(V-\Sigma)^{-1} && (45)
\end{aligned}
$$

Lemma 1

$$
\begin{aligned}
&\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)^T\left(\Sigma_j^{-1}+\Sigma_k^{-1}\right)^{-1}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)\\
&\qquad=(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k) \qquad (46)
\end{aligned}
$$

proof

$$
\begin{aligned}
&\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)^T\left(\Sigma_j^{-1}+\Sigma_k^{-1}\right)^{-1}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)\\
&= \mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)^T\left\{\Sigma_j-\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\Sigma_j\right\}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)\\
&= \mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k\\
&\quad-\left\{\mu_j^T-\mu_j^T(\Sigma_j+\Sigma_k)^{-1}\Sigma_j+\mu_k^T\Sigma_k^{-T}\Sigma_j-\mu_k^T\Sigma_k^{-1}\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\Sigma_j\right\}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)\\
&= \mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\mu_j^T\Sigma_j^{-1}\mu_j-\mu_j^T\Sigma_k^{-1}\mu_k+\mu_j^T(\Sigma_j+\Sigma_k)^{-1}\mu_j+\mu_j^T(\Sigma_j+\Sigma_k)^{-1}\Sigma_j\Sigma_k^{-1}\mu_j\\
&\quad-\mu_k^T\Sigma_k^{-T}\mu_j-\mu_k^T\Sigma_k\Sigma_j^{-1}\Sigma_k\mu_j+\mu_k^T\Sigma_k^{-1}\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\Sigma_j\Sigma_k^{-1}\mu_k\\
&= \mu_k^T\Sigma_k^{-1}\mu_k-\mu_j^T(\Sigma_j+\Sigma_k)^{-1}\mu_k-\mu_k^T(\Sigma_j+\Sigma_k)^{-1}\mu_j+\mu_j^T(\Sigma_j+\Sigma_k)^{-1}\mu_j-\mu_k^T\Sigma_k^{-T}\Sigma_j(\Sigma_j+\Sigma_k)^{-1}\mu_k\\
&= (\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k) \qquad (47)
\end{aligned}
$$

Lemma 2

$$
\begin{aligned}
&(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)+(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\\
&\qquad=(x-\zeta)^T\Sigma^{-1}(x-\zeta)+(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k) \qquad (48)
\end{aligned}
$$

Proof

Using Lemma 1, we have

$$
\begin{aligned}
&(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)+(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\\
&= x^T\left(\Sigma_j^{-1}+\Sigma_k^{-1}\right)x-x^T\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k\right)^Tx+\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k\\
&= \left\{x-(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k)\right\}^T(\Sigma_j^{-1}+\Sigma_k^{-1})\left\{x-(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_kt)\right\}\\
&\quad+\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k)^T(\Sigma_j^{-1}+\Sigma_k^{-1})^{-1}(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k)\\
&= (x-\zeta)^T\Sigma^{-1}(x-\zeta)+(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k) \qquad (49)
\end{aligned}
$$

Lemma 3

$$
\begin{aligned}
&\zeta^T\Sigma^{-1}\zeta-\xi^TV^{-1}\xi-\left(\Sigma^{-1}\zeta-V^{-1}\xi\right)^T\left(\Sigma^{-1}-V^{-1}\right)^{-1}\left(\Sigma^{-1}\zeta-V^{-1}\xi\right)\\
&\qquad=-(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi) \qquad (50)
\end{aligned}
$$

Proof

$$\begin{aligned}
&\zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi - \left(\Sigma^{-1}\zeta - V^{-1}\xi\right)^T \left(\Sigma^{-1} - V^{-1}\right)^{-1} \left(\Sigma^{-1}\zeta - V^{-1}\xi\right) \\
&= \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi - \left(\Sigma^{-1}\zeta - V^{-1}\xi\right)^T \Sigma(V-\Sigma)^{-1}V \left(\Sigma^{-1}\zeta - V^{-1}\xi\right) \\
&= \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi - (\zeta^T - \xi^t V^{-1}\Sigma)(V-\Sigma)^{-1}(V\Sigma^{-1}\zeta - \xi) \\
&= \zeta^T\Sigma^{-1}\zeta + \xi^T V^{-1}\xi - \zeta^T(V-\Sigma)^{-1}V\Sigma^{-1}\zeta - \zeta^T(V-\Sigma)^{-1}\xi \\
&\quad + \xi^T V^{-1}\Sigma(V-\Sigma)^{-1}V\Sigma^{-1}\zeta + \xi^T V^{-1}\Sigma(V-\Sigma)^{-1}\xi \\
&= -(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi) \qquad (51)
\end{aligned}$$

Lemma 4

$$(x-\zeta)^T\Sigma^{-1}(x-\zeta) - (x-\xi)^T V^{-1}(x-\xi) = (x-\eta)^T W(x-\eta) - (\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi) \qquad (52)$$

Proof

Using Lemma 3,

$$\begin{aligned}
&(x-\zeta)^T\Sigma^{-1}(x-\zeta) - (x-\xi)^T V^{-1}(x-\xi) \\
&= x^T(\Sigma^{-1} - V^{-1})x - x^T(\Sigma^{-1}\zeta - V^{-1}\xi) - (\Sigma^{-1}\zeta - V^{-1}\xi)^T x + \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi \\
&= \left\{x - \left(\Sigma^{-1} - V^{-1}\right)^{-1}\left(\Sigma^{-1}\zeta - V^{-1}\xi\right)\right\}^T \left(\Sigma^{-1} - V^{-1}\right) \left\{x - \left(\Sigma^{-1} - V^{-1}\right)^{-1}\left(\Sigma^{-1}\zeta - V^{-1}\xi\right)\right\} \\
&\quad + \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi - \left(\Sigma^{-1}\zeta - V^{-1}\xi\right)^T \left(\Sigma^{-1} - V^{-1}\right)^{-1} \left(\Sigma^{-1}\zeta - V^{-1}\xi\right) \\
&= (x-\eta)^T\left(\Sigma^{-1} - V^{-1}\right)(x-\eta) + \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi \\
&\quad - \left(\Sigma^{-1}\zeta - V^{-1}\xi\right)^T \left(\Sigma^{-1} - V^{-1}\right)^{-1} \left(\Sigma^{-1}\zeta - V^{-1}\xi\right) \\
&= (x-\eta)^T W(x-\eta) + \zeta^T\Sigma^{-1}\zeta - \xi^T V^{-1}\xi - \left(\Sigma^{-1}\zeta - V^{-1}\xi\right)^T W^{-1} \left(\Sigma^{-1}\zeta - V^{-1}\xi\right) \\
&= (x-\eta)^T W(x-\eta) - (\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi) \qquad (53)
\end{aligned}$$

Lemma 5

$$\begin{aligned}
&\mu_j^T\Sigma_j^{-1}\mu_j + \mu_k^T\Sigma_k^{-1}\mu_k - \xi^T V^{-1}\xi \\
&= (\mu_j - \mu_k)^T(\Sigma_j + \Sigma_k)^{-1}(\mu_j - \mu_k) - (\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi). \qquad (54)
\end{aligned}$$

Proof

$$\begin{aligned}
&\mu_j^T\Sigma_j^{-1}\mu_j + \mu_k^T\Sigma_k^{-1}\mu_k - \xi^T V^{-1}\xi \\
&\quad - \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right)^T \left(\Sigma_j^{-1} + \Sigma_k^{-1} - V^{-1}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right) \\
&= \mu_j^T\Sigma_j^{-1}\mu_j + \mu_k^T\Sigma_k^{-1}\mu_k - \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right)^T \left(\Sigma_j^{-1} + \Sigma_k^{-1}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right) \\
&\quad - \xi^T V^{-1}\xi + \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right)^T \left(\Sigma_j^{-1} + \Sigma_k^{-1}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right) \\
&\quad - \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right)^T \left(\Sigma^{-1} - V^{-1}\right) \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right) \\
&= (\mu_j - \mu_k)^T(\Sigma_j + \Sigma_k)^{-1}(\mu_j - \mu_k) \qquad (55) \\
&\quad - \xi^T V^{-1}\xi + \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right)^T \left(\Sigma_j^{-1} + \Sigma_k^{-1}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right) \\
&\quad - \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right)^T \left(\Sigma^{-1} - V^{-1}\right) \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right)
\end{aligned}$$

Here, the terms after the second term of the above equation can be expressed in a signle term as follows:

$$\begin{aligned}
&-\xi^T V^{-1}\xi + \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right)^T \left(\Sigma_j^{-1} + \Sigma_k^{-1}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k\right) \\
&\quad - \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right)^T (\Sigma^{-1} - V^{-1}) \left(\Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k - V^{-1}\xi\right) \\
&= -\xi V^{-1}\xi + \zeta^T\Sigma^{-1}\zeta - (\Sigma^{-1}\zeta - V^{-1}\xi)^T(\Sigma^{-1} - V^{-1})(\Sigma^{-1}\zeta - V^{-1}\xi) \\
&= -\xi V^{-1}\xi + \zeta^T\Sigma^{-1}\zeta - (\Sigma^{-1}\zeta - V^{-1}\xi)^T\Sigma(V-\Sigma)^{-1}V(\Sigma^{-1}\zeta - V^{-1}\xi) \\
&= -\xi V^{-1}\xi + \zeta^T\Sigma^{-1}\zeta - (\zeta^T - \xi^T V^{-1}\Sigma)^T(V-\Sigma)^{-1}(V\Sigma^{-1}\zeta - \xi) \\
&= \zeta^T\Sigma^{-1}\zeta - \xi V^{-1}\xi - \zeta^T(V-\Sigma)^{-1}V\Sigma^{-1}\zeta + \xi^T V^{-1}\Sigma(V-\Sigma)^{-1}V\Sigma^{-1}\zeta \\
&\quad + \zeta^{-1}(V-\Sigma)^{-1}\xi - \xi^T V^{-1}\Sigma(V-\Sigma)^{-1}\xi \\
&= -\zeta^T(V-\Sigma)^{-1}\zeta - \xi(V-\Sigma)^{-1}\xi + \zeta^T(V-\Sigma)^{-1}\xi + \xi^T(V-\Sigma)^{-1}\zeta \\
&= -(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi). \qquad (56)
\end{aligned}$$

Proposition

Assume that $f_j(x)$, $f_j(x)$, $f_k(x)$ and $p_{jk}(x)$ are respectively given by $f_j(x) \sim N(\mu_j, \Sigma_j)$, $f_k(x) \sim N(\mu_k, \Sigma_k)$ and $p_{jk}(x) \sim N(\zeta, V)$, then integral of $f_j(x)f_k(x)/p_{jk}(x)$ over the whole domain is given by

$$\begin{aligned}
\int \frac{f_j(x)f_k(x)}{p_{jk}(x)}dx &= \left|\Sigma_j\right|^{-\frac{1}{2}} |\Sigma_k|^{-\frac{1}{2}} |V|^{\frac{1}{2}} |W|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\} \qquad (57)
\end{aligned}$$

Proof

Since $f_j(x)$ and $f_k(x)$ are defined by

$$\begin{aligned}
f_j(x) &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_j\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\} \\
f_k(x) &= (2\pi)^{-\frac{k}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}, \qquad (58)
\end{aligned}$$

respectively, $f_j(x)f_k(x)$ is given by

$$f_j(x)f_k(x) = (2\pi)^{-k} \left|\Sigma_j\right|^{-\frac{1}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}. \qquad (59)$$

Then by Lemma 2

$$\begin{aligned}
f_j(x)f_k(x) &= (2\pi)^{-k} \left|\Sigma_j\right|^{-\frac{1}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\zeta)^T\Sigma^{-1}(x-\zeta)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\}.
\end{aligned}$$

Since $p_{jk}(x)$ is defined by

$$p_{jk}(x) = (2\pi)^{-\frac{k}{2}} |V|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\xi)^T V^{-1}(x-\xi)\right\}, \qquad (60)$$

$f_j(x)f_k(x)/p_{jk}(x)$ is given by

$$\begin{aligned}
\frac{f_j(x)f_k(x)}{p_{jk}(x)} &= (2\pi)^{-\frac{k}{2}} \left|\Sigma_j\right|^{-\frac{1}{2}} |\Sigma_k|^{-\frac{1}{2}} |V|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(x-\zeta)^T\Sigma^{-1}(x-\zeta) + \frac{1}{2}(x-\xi)^T V^{-1}(x-\xi)\right\}. \qquad (61)
\end{aligned}$$

Then by Lemma 4, it can be expressed as

$$\begin{aligned}\frac{f_j(x)f_k(x)}{p_{jk}(x)} &= (2\pi)^{-\frac{k}{2}}\left|\Sigma_j\right|^{-\frac{1}{2}}\left|\Sigma_k\right|^{-\frac{1}{2}}|V|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\}\\ &\quad\times\exp\left\{\frac{1}{2}(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi)\right\}\exp\left\{-\frac{1}{2}(x-\eta)^TW(x-\eta)\right\}. \qquad (62)\end{aligned}$$

By integrating whole domain of x, we obtain

$$\begin{aligned}\int\frac{f_j(x)f_k(x)}{p_{jk}(x)}dx &= \left|\Sigma_j\right|^{-\frac{1}{2}}\left|\Sigma_k\right|^{-\frac{1}{2}}|V|^{\frac{1}{2}}|W|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi)\right\}\\ &\quad\times\exp\left\{-\frac{1}{2}(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)\right\}, \qquad (63)\end{aligned}$$

which complete the proof of the proposition.

By putting $\mu_k=\mu_j$,$\Sigma_k=\Sigma_j$ in the Proposition we obtain the following

Corollary

$$\int\frac{f(x)_j^2}{p_{jk}(x)}dx = \left|\Sigma_j\right|^{-1}|V|^{\frac{1}{2}}\left|W_j\right|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}(\mu_j-\xi)^T(V-\frac{1}{2}\Sigma_j)^{-1}(\mu_j-\xi)\right\}. \qquad (64)$$

Note

The equation (62) can be directly obtained by considering the expression of the $f_j(x)f_k(x)/p_{jk}(x)$ as follows:

$$\begin{aligned}\frac{f_j(x)f_k(x)}{p_{jk}(x)} &= (2\pi)^{-\frac{k}{2}}\left|\Sigma_j\right|^{-\frac{1}{2}}\left|\Sigma_k\right|^{-\frac{1}{2}}|V|^{\frac{1}{2}} \qquad (65)\\ &\times\ \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)+\frac{1}{2}(x-\xi)^TV^{-1}(x-\xi)\right\}.\end{aligned}$$

Here the terms in the brace of the right hand side of the above equation is given by

$$\begin{aligned}&(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)+(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)-(x-\xi)^TV^{-1}(x-\xi)\\ &= x^T\left(\Sigma_j^{-1}+\Sigma_k^{-1}-V^{-1}\right)x-x^T\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k-V^{-1}\xi\right)\\ &\quad-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k-V^{-1}\xi\right)^Tx+\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\xi^TV^{-1}\xi\\ &= (x-\zeta)^{-1}W(x-\zeta)^{-1}+\mu_j^T\Sigma_j^{-1}\mu_j+\mu_k^T\Sigma_k^{-1}\mu_k-\xi^TV^{-1}\xi\\ &\quad-\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k-V^{-1}\xi\right)^T\left(\Sigma_j^{-1}+\Sigma_k^{-1}-V^{-1}\right)^{-1}\left(\Sigma_j^{-1}\mu_j+\Sigma_k^{-1}\mu_k-V^{-1}\xi\right). \qquad (66)\end{aligned}$$

Then by Lemma 5, it can be expressed as

$$(x-\zeta)^{-1}W(x-\zeta)^{-1}+(\mu_j-\mu_k)^T(\Sigma_j+\Sigma_k)^{-1}(\mu_j-\mu_k)-(\zeta-\xi)^T(V-\Sigma)^{-1}(\zeta-\xi). \qquad (67)$$

Therefore we obtain the equation (62).

Aknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 18H03210.

References

[1] Alspach, D. and Sorenson, H. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. IEEE transactions on automatic control, 17(4), 439–448.

[2] Crouse, D. F., Willett, P., Pattipati, K. and Svensson, L. (2011). A look at Gaussian mixture reduction algorithms. In *14th International Conference on Information Fusion, IEEE*, 1–8.

[3] Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400), 1032–1041.

[4] Kitagawa, G. (1989). Non-Gaussian seasonal adjustment, *Computers & Mathematics with Applications*, Vol.18, No.6/7, pp. 503–514.

[5] Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother, *Annals of the Institute of Statistical Mathematics*, Vol. 46, No.4, pp. 605–623.

[6] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics*, Vol.5, no.1, pp. 1–25.

[7] Runnalls, A.R. (2007). A Kullback-Leibler approach to Gaussian mixture reduction, *IEEE Trans. Aerospace and Electronics Systems*, Vol. 43, No. 3, pp. 989–999.

[8] Salmond, D.L. (1990). Mixture reduction algorithms for target tracking in clutter, in *Signal and Data Processing of Small Targets 1990, Proc. of SPIE*, 1305, 434–445.

[9] Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. Automatica, 7(4), 465–479.

[10] West, M. (1993). Approximate posterior distributions by mixture, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 55, No. 2, pp. 409–422.

[11] Williams, J.L. and Maybeck, P.S. (2003). Cost-function-based Gaussian mixture reduction, in *Sixth Int. Conf. on Information Fusion*, Vol. 2, pp. 1047–1054, Piscataway, NJ: IEEE Publ.