Distributed Estimation in the Presence of Correlation

A Dissertation presented

by

Zhiyuan Weng

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Electrical Engineering

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Zhiyuan Weng

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Petar M. Djurić - Dissertation Advisor
Professor, Department of Electrical & Computer Engineering

Sangjin Hong - Chairperson of Defense
Professor, Department of Electrical & Computer Engineering

Mónica F. Bugallo
Associate Professor, Department of Electrical & Computer Engineering

Jiaqiao Hu
Associate Professor, Department of Applied Mathematics and Statistics

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Distributed Estimation in the Presence of Correlation

by

Zhiyuan Weng

Doctor of Philosophy

in

Electrical Engineering

Stony Brook University

2014

We study the problem of distributed estimation, where a group of nodes are required to cooperate with each other to estimate some parameter of interest from noisy measurements without a fusion center. Distributed estimation algorithms are useful in several areas, including wireless sensor networks, where robustness, scalability, flexibility, and low power consumption are desirable. In this dissertation, we mainly focus on the cases where the node measurements are correlated. First, we consider the problem of fusing multiple estimates from different nodes. Cases of both known and unknown correlation are investigated. A Bayesian approach and a convex optimization approach are proposed. Second, we study the sequential estimation problem where all the nodes in the network cooperate to estimate a static parameter recursively, and where the correlation between measurements from different nodes are known. We propose an efficient distributed algorithm and prove that it is optimal in the sense that the ratio of the variance of the proposed estimator to that of the centralized estimator approaches one in the long run. Last, we study the belief consensus problem in the networks. Instead of estimating a scalar or a vector, we are interested in the beliefs of nodes, which are represented as probability densities. The Chi-square information is used as the criterion to determine the optimal values of the weighting coefficients in the fusion of densities. We also prove that the optimization problem of minimizing the Chi-square information with respect to the weighting coefficients is convex, and therefore can be solved efficiently by existing methods.

Contents

List of Figures

Notation

$\mathbf{A} \succeq \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix
$\mathbf{A} \succ \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is a positive definite matrix
$\mathbf{A} \geq \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is element-wise nonnegative
$\mathbf{A} > \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is element-wise positive
$x \sim p(x)$	random variable x is distributed according to $p(x)$
$\mathrm{tr}\left[\mathbf{A}\right]$	trace of $\mathbf{A}$
$\mathbb{E}[\mathbf{x}]$	expectation of $\mathbf{x}$
$\mathrm{Cov}[\mathbf{x}]$	covariance of $\mathbf{x}$
$\otimes$	Kronecker product
$\mathbf{I}$	an identity matrix
$\mathbf{I}_M$	an identity matrix with size $M \times M$
$\mathbf{O}$	a matrix with all entries equal to zero
$\Gamma(\cdot)$	standard gamma function
$\Gamma_n(\cdot)$	multivariate gamma function
$\mathcal{N}(\mathbf{m}, \mathbf{C})$	normal distribution
$\mathcal{W}_M(n, \mathbf{\Sigma})$	Wishart distribution
$\|\mathbf{A}\|$	determinant of $\mathbf{A}$
$\|\mathbf{x}\|$	Euclidean norm of $\mathbf{x}$
$\|\mathbf{A}\|_F$	Frobenius norm of $\mathbf{A}$
$\|\mathbf{A}\|$	sum of Euclidean norm of columns of $\mathbf{A}$
$\mathbb{N}$	the set of natural numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^+$	the set of positive real numbers
$\mathbb{S}$	the set of symmetric matrices
$\mathbb{S}^+$	the set of positive definite matrices
$\mathcal{N}_i$	neighbors of node i
$\mathbf{A}^\top$	transpose of $\mathbf{A}$
$\mathbf{1}_M$	a $M \times 1$ column vector with all entries equal to 1
$\mathbf{0}_M$	a $M \times 1$ column vector with all entries equal to 0
$\delta_{i,j}$	Kronecker delta

Chapter 1

Introduction

1.1 Overview

In recent years, research in the area of distributed estimation has been growing quickly due to the increasing popularity of distributed systems, like the wireless sensor network (WSN). A WSN consists of many sensor nodes that cooperate with each other to perform an inference or monitoring task, where data are exchanged and shared between neighbors through wireless communication. The objective of distributed systems is to utilize the data at different locations to enhance their performance. With a centralized architecture, a fusion center collects data from all the sensors to perform the computation and processing tasks. In most cases, a decentralized approach is preferred, because it can provide a degree of scalability, flexibility and robustness which cannot be achieved with traditional centralized architectures.

On the other hand, a decentralized approach comes with its disadvantages. One important issue associated with the distributed processing is the information redundancy. In other words, measurements from different sensors usually have correlation and cannot be treated and processed as statistically independent variables. Correlation arises for various reasons. The most common reason is the presence of correlated noise in the measurement. Besides, if we consider the underlying parameter of interest as a random variable, all the measurements of the parameter becomes statistically dependent if the parameter is unknown. If the parameter is known, then the measurements are no longer dependent. This can be easily understood using graphical models [1]. Another more complicated and challenging correlation appears when information is being spread and diffused over networks. This is because each time nodes exchange information with their neighbors, they have more information in common with their neighbors until finally all the nodes possess the same piece of information. As the correlation increases, it becomes more and more difficult to extract useful information for nodes through communication with others. In other words, it makes it difficult for nodes to collect all the information available in the network. How to properly diffuse the information in the network and allow the nodes to aggregate the information is a challenging topic.

1.2 Contributions

The thrust of our work is the solutions to the problems of distributed fusion and estimation in the presence of correlation. This is an important topic because correlation is prevalent in most practical applications. We consider cases with both known and unknown correlation.

Our first contribution is in the area of information fusion. Given multiple correlated estimates from different sensor nodes, we seek a "good" way to combine these estimates. Two methods are proposed to solve the problem. The first method considers the problem within the Bayesian framework and assumes that the covariance matrix of the concatenated estimate has a prior distribution. We then derive the conditional distribution of the off-diagonal blocks of the covariance matrix, which is the cross-correlation of our interest. We design a special algorithm to sample from this distribution and then use the Monte Carlo method to compute the minimum mean square error (MMSE) estimate for the fusion problem. In the second method, we try to estimate the cross-covariances rather than marginalize them. We consider two settings, one where we do not use priors for the covariance matrices of the model and another, where we use priors and engage the Bayesian machinery. Both formulations turn out to require convex optimization and they can be solved by existing techniques. When the cross-covariance estimates are obtained, the weighting coefficients can easily be calculated so that optimal fusion can take place.

For the second contribution, we consider distributed sequential estimation in a network in the presence of correlated noises. Unlike the former setting, here we assume that the correlation of the noises is known to each node. A distributed sequential estimation algorithm is proposed. At each iteration, a node exchanges information with its neighbors. The node also updates the estimate with its new local observation. Further, it is assumed that the noises have the Markov property with respect to the network topology. A doubly stochastic matrix is employed to average the sufficient statistics over the network. We compare the proposed method with the centralized one. We show that the ratio of the variances of the two estimators approaches one. Therefore, the proposed estimator is asymptotically efficient.

Last, we study the problem of belief consensus in the networks. At the beginning, each node has an initial belief based on its observations. Unlike traditional problems where the beliefs are scalars or vectors, here we assume that the beliefs are probability densities. Ideally, the nodes should reach consensus at the density that is equal to the product of all the initial densities in the network. The desired density can be considered as the Bayesian posterior. However, we show that without knowledge of the number of nodes in the network, optimal consensus cannot be achieved. Instead, we use the weighted product of the node densities to approximate the Bayesian posterior. We adopt the χ^2 information as the criterion to measure distance between the Bayesian posterior and the weighted product of the densities. We prove that the optimization problem of minimizing the χ^2 information with respect to the weighting coefficients is convex and therefore can be solved efficiently.

1.3 Dissertation Organization

The dissertation is organized as follows. In Chapter 2, we formulate the data fusion problem. We then use two approaches to address the problems, both of which are related to covariance estimation. The Bayesian approach is proposed in Chapter 3 and the convex optimization method is considered in Chapter 4. In Chapter 5, we consider distributed estimation in networks. Particularly, we consider the case when the noises are zero-mean white Gaussian noises with spatial correlation. We consider the problem of belief consensus in Chapter 6. We conclude the dissertation with Chapter 7.

Chapter 2

Covariance Estimation and Data Fusion

2.1 Introduction

In this Chapter, we formulate the fusion problem we will discuss in Chapter 3 and Chapter 4. In many applications, the information propagated through a sensor network is transformed to a form that provides the estimated state of interest. For example, in distributed Kalman filtering [2, 3, 4, 5], the information is converted into the first and second moment statistics. With the statistics from neighbors at hand, fusing the estimates to obtain a better estimate is expected. A serous problem arising in such setting is the effect of redundant information [6]. The estimates provided by different nodes have unknown cross-correlations. This is particularly true for networks with unknown topological structure. Pieces of information from two nodes cannot be simply combined by averaging and weighted averaging unless they are independent or have a known degree of correlation.

Many approaches have been proposed to mitigate the problem. Most of them fall into two categories. The first category is looking for an optimal linear combination of estimates in terms of some criterion, for example, weighted least squares or minimum variance [7, 8]. In [9, 10, 11, 12], a unified model is developed for estimation and fusion based on the best linear unbiased estimation (BLUE) or linear minimum variance approach. The second category tries to fuse the available estimates directly [13, 14, 15, 16, 17]. Algorithms for combining estimates of both the first and the second moments in linear systems have been proposed. It is a linear combination of estimates when the first two moments are given. However, none of the above methods investigated the situation where the covariance of each estimate is available while the cross-covariance is unknown. Consider the following problem. Given N estimates $\mathbf{x}_j$ for $j \in \{1, \cdots, N\}$ of the true state vector $\mathbf{x}_0 \in \mathbb{R}^{M \times 1}$ with their covariance matrices of the estimation error, $\mathbf{P}_{j,j}$, we seek a fusion scheme that combines the available information and provides an estimate $\hat{\mathbf{x}}_0$ with minimum mean square error. We denote by $\mathbf{P}_0$ the covariance matrix of the estimation error of $\hat{\mathbf{x}}_0$. A naive but simple method is to calculate the weighted average, where the weighting coefficients are proportional to the degrees of the nodes (the numbers of the neighbors of the nodes) [3]. The approach makes sense because the higher the degree, the more information the node collects and the better it does in estimation. A more complicated and popular one is the Covariance Intersection (CI) [18]. It provides a general framework for information fusion where there is a lack of knowledge about cross-correlation between noisy measurements, and it yields consistent estimates between the fused local estimates. Here “consistent” means that the resulting covariance is an upper-bound of the true covariance. The algorithm can be expressed as

$$\mathbf{P}_0^{-1} = \sum_{j=1}^{N} w_j \mathbf{P}_{j,j}^{-1} \tag{2.1}$$

$$\mathbf{P}_0^{-1}\hat{\mathbf{x}}_0 = \sum_{j=1}^{N} w_j \mathbf{P}_{j,j}^{-1}\mathbf{x}_j, \tag{2.2}$$

where the weighting coefficient $w_j \in [0, 1]$ and $\sum_{j=1}^{N} w_j = 1$ holds. Different performance criteria can be used to decide the values of w_j. Since the mean square error is of our interests, we use the trace of $\mathbf{P}_0$ as the criterion. The minimization of the trace requires iterative minimization of a nonlinear cost function with respect to the weight coefficients w_j. In order to reduce the computational complexity, several suboptimal non-iterative algorithms for fast Covariance Intersection have been developed [19, 20].

In [19], it was reasoned that a replacement of $\mathbf{P}_{i,i}$ by $\mathbf{P}_{j,j}$ and vice versa must lead to correspondingly switched coefficients w_i and w_j and that if $\mathrm{tr}\,[\mathbf{P}_{i,i}] \ll \mathrm{tr}\,[\mathbf{P}_{j,j}]$ for $j \neq i, j \in \{1, \cdots, N\}$ one would expect to get $w_i \approx 1$. Thus it was suggested to use the linear equations

$$\mathrm{tr}\,[\mathbf{P}_{i,i}]\, w_i - \mathrm{tr}\,[\mathbf{P}_{j,j}]\, w_j = 0, (i, j = 1, \cdots, N) \tag{2.3}$$

which leads to the solution:

$$w_i = \frac{1/\mathrm{tr}\,[\mathbf{P}_{i,i}]}{\sum_{j=1}^{N} 1/\mathrm{tr}\,[\mathbf{P}_{j,j}]}. \tag{2.4}$$

In [20], it was pointed out that the above approximation fails to consider the relative orientation of the estimation error variance matrices which may lead to a degraded performance in certain applications. Accordingly, an improved fast Covariance Intersection algorithm was proposed which comes with increased computational complexity while yielding better performance in some cases and comparable results.

2.2 Problem Formulation

Consider that a node in a network has $N-1$ nodes in its neighborhood. By communication with its neighbors, it has N available measurements, including its own. Each measurement $\mathbf{x}_j$ for $j \in \{1, \cdots, N\}$ is an $M \times 1$ vector, with covariance matrices of the estimation error $\mathbf{P}_{j,j}$. We concatenate the N vectors and let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \tag{2.5}$$

where $\mathbf{x} \in \mathbb{R}^{MN\times 1}$. We assume the mean of $\mathbf{x}_j$ is the true state $\mathbf{x}_0$. Therefore, the covariance matrix of $\mathbf{x}$ is also the covariance matrix of the estimation error of $\mathbf{x}$. We denote by $\mathbf{P}_x$ the covariance matrix of $\mathbf{x}$.

$$\mathbf{P}_x = \begin{bmatrix} \mathbf{P}_{1,1} & \mathbf{P}_{1,2} & \cdots & \mathbf{P}_{1,N} \\ \mathbf{P}_{1,2}^{\top} & \mathbf{P}_{2,2} & \cdots & \mathbf{P}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{1,N}^{\top} & \mathbf{P}_{2,N}^{\top} & \cdots & \mathbf{P}_{N,N} \end{bmatrix}. \tag{2.6}$$

We start by considering linear and unbiased estimators in the form

$$\hat{\mathbf{x}}_0 = \mathbf{W}^\top \mathbf{x}. \tag{2.7}$$

$\mathbf{W}$ is the weighting coefficient matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1^\top \\ \mathbf{W}_2^\top \\ \vdots \\ \mathbf{W}_N^\top \end{bmatrix} \tag{2.8}$$

where $\mathbf{W}_j \in \mathbb{R}^{M \times M}$. Since the estimate should be unbiased, we require

$$\mathbf{W}_1 + \mathbf{W}_2 + \cdots + \mathbf{W}_N = \mathbf{I}. \tag{2.9}$$

Let $\mathbf{I}_{(N)}$ be a $NM \times M$ matrix concatenated vertically by N identity matrices with sizes $M \times M$,

$$\mathbf{I}_{(N)} = \begin{bmatrix} \mathbf{I}_M \\ \mathbf{I}_M \\ \vdots \\ \mathbf{I}_M \end{bmatrix}. \tag{2.10}$$

Then (2.9) becomes

$$\mathbf{W}^\top \mathbf{I}_{(N)} = \mathbf{I}. \tag{2.11}$$

Let $\mathbf{P}_0$ be the covariance matrix of $\hat{\mathbf{x}}_0$, which can be expressed as

$$\mathbf{P}_0 = \mathbf{W}^\top \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{W} = \mathbf{W}^\top \mathbf{P}_x \mathbf{W}. \tag{2.12}$$

The minimization of the mean square error is equivalent to the minimization of $\operatorname{tr}[\mathbf{P}_0]$, This can be carried out by using the method of Lagrange multipliers. Let $\boldsymbol{\lambda}$ be the matrix of Lagrange multipliers. Define the Lagrangian Λ as

$$\Lambda(\mathbf{W}) = \operatorname{tr}\left[\mathbf{W}^\top \mathbf{P}_x \mathbf{W}\right] + \operatorname{tr}\left[\boldsymbol{\lambda}\left(\mathbf{W}^\top \mathbf{I}_{(N)} - \mathbf{I}\right)\right]. \tag{2.13}$$

Taking derivative with respect to $\mathbf{W}$ and $\boldsymbol{\lambda}$ and using the identities

$$\frac{\partial \operatorname{tr}\left(\mathbf{X}\mathbf{A}\mathbf{X}^\top\right)}{\partial \mathbf{X}} = \mathbf{X}\mathbf{A} + \mathbf{X}\mathbf{A}^\top \tag{2.14}$$

$$\frac{\partial \operatorname{tr}\left(\mathbf{A}\mathbf{X}\mathbf{B}\right)}{\partial \mathbf{X}} = \mathbf{A}^\top \mathbf{B}^\top, \tag{2.15}$$

we obtain the stationary points by the following equations

$$2\mathbf{W}^\top \mathbf{P}_x + \boldsymbol{\lambda}^\top \mathbf{I}_{(N)}^\top = \mathbf{O} \tag{2.16}$$

$$\mathbf{W}^\top \mathbf{I}_{(N)} = \mathbf{I}. \tag{2.17}$$

Combining all of the three equations, we obtain

$$\mathbf{W}^\top = \left(\mathbf{I}_{(N)}^\top \mathbf{P}_x^{-1} \mathbf{I}_{(N)}\right)^{-1} \mathbf{I}_{(N)}^\top \mathbf{P}_x^{-1} \tag{2.18}$$

$$\mathbf{P}_0 = \mathbf{W}^\top \mathbf{P}_x \mathbf{W} = \left(\mathbf{I}_{(N)}^\top \mathbf{P}_x^{-1} \mathbf{I}_{(N)}\right)^{-1}. \tag{2.19}$$

By substituting (2.18) into (2.7), we have

$$\hat{\mathbf{x}}_0 = \left(\mathbf{I}_{(N)}^\top \mathbf{P}_x^{-1} \mathbf{I}_{(N)}\right)^{-1} \mathbf{I}_{(N)}^\top \mathbf{P}_x^{-1} \mathbf{x}. \tag{2.20}$$

However, in many situations we do not have information about $\mathbf{P}_{i,j}$ for $i \neq j$. We develop a Bayesian approach and a convex optimization approach in the following two chapters.

Chapter 3

Bayesian Approach

3.1 Introduction

Our strategy to solving the problem is to put it into a Bayesian framework. We assume that $\mathbf{P}_x$ has a prior and that the prior is a Wishart distribution. The Wishart distribution is a family of probability distributions defined over symmetric, nonnegative-definite matrix-valued random matrices. These distributions are of great importance in the estimation of covariance matrices in multivariate statistics [21]. The Wishart distribution is defined as follows: an $M \times M$ random matrix $\mathbf{A}$ is said to have a Wishart distribution if its probability distribution function (pdf) is given by

$$p\left(\mathbf{A}\right) = \frac{\left|\mathbf{A}\right|^{\frac{n-M-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\mathbf{A}\right]\right)}{2^{\frac{Mn}{2}} \left|\mathbf{\Sigma}\right|^{\frac{n}{2}} \Gamma_M\left(\frac{n}{2}\right)}, \tag{3.1}$$

where $\mathbf{\Sigma}$ is a positive definite matrix, $n \geq M$ is the degree of freedom and $\Gamma_M(n)$ is the multivariate gamma function defined as [22]

$$\Gamma_M\left(n\right) = \pi^{M(M-1)/4} \prod_{j=1}^{M} \Gamma\left(n - \frac{1}{2}(j-1)\right). \tag{3.2}$$

We denote by $\mathcal{W}_M\left(n, \Sigma\right)$ the Wishart distribution. The degree of freedom n also plays an important role in our Bayesian framework as later we will see. We will omit M and write simply $\mathcal{W}\left(n, \Sigma\right)$ if the size of the matrix is obvious from the context. The Wishart distribution is closely related to the multivariate Gaussian distribution. The pdf of the multivariate Gaussian distribution of an $M \times 1$ vector is

$$p\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{M/2} \left|\mathbf{C}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}\right)^{\top} \mathbf{C}^{-1}\left(\mathbf{x}-\mathbf{m}\right)\right), \tag{3.3}$$

where $\mathbf{m}$ is the mean and $\mathbf{C}$ is the covariance matrix, denoted by $\mathcal{N}\left(\mathbf{m}, \mathbf{C}\right)$. Suppose $\mathbf{X}$ is an $n \times M$ matrix, the rows of which have M-variate normal distribution $\mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right)$. Then the $M \times M$ random matrix $\mathbf{A} = \mathbf{X}^{\top}\mathbf{X}$ has a Wishart distribution, i.e., $\mathcal{W}\left(n, \mathbf{\Sigma}\right)$. This property makes it easy to generate Wishart random matrices.

We denote by $\mathbf{P}_o$ and $\mathbf{P}_d$ the off-diagonal block matrices and the diagonal block matrices, respectively, i.e.,

$$\mathbf{P}_d = \left\{\mathbf{P}_{j,j}: \ j \in \{1, \cdots, N\}\right\} \tag{3.4}$$

$$\mathbf{P}_o = \left\{\mathbf{P}_{i,j}: \ i \neq j; \ i, j \in \{1, \cdots, N\}\right\}. \tag{3.5}$$

In our problem, we know $\mathbf{P}_d$. To fuse the data, we would like to have information of $\mathbf{P}_o$ conditioned on $\mathbf{P}_d$. We express this by the conditional

$$p\left(\mathbf{P}_o|\mathbf{P}_d\right) = \frac{p\left(\mathbf{P}_x\right)}{p\left(\mathbf{P}_d\right)}. \tag{3.6}$$

Since $\mathbf{P}_d$ is known, the weighting matrix $\mathbf{W}$, and therefore $\hat{\mathbf{x}}_0$ are uniquely determined by $\mathbf{P}_o$ as in (2.20). We think of it as a function of the matrix variable $\mathbf{P}_o$ and use $f\left(\mathbf{P}_o\right)$ to denote it. Note that $\mathbf{P}_o$ cannot be an arbitrary matrix. $\mathbf{P}_o$ must lie in the set $\mathcal{P}_o$ defined by

$$\mathcal{P}_o = \left\{\mathbf{P}_o : \mathbf{P}_x > 0\right\}, \tag{3.7}$$

where $\mathbf{P}_x$ is defined in (2.6). We express the MMSE estimator by

$$\hat{\mathbf{x}}_{mmse} = \int_{\mathcal{P}_o} f\left(\mathbf{P}_o\right) p\left(\mathbf{P}_o|\mathbf{P}_d\right) d\mathbf{P}_o. \tag{3.8}$$

Unfortunately, the above integral is computationally intractable.

In order to approximate the integral, we have to resort to the Monte Carlo method. We sample K independent random matrices, $\mathbf{P}_o^{(j)} \sim p\left(\mathbf{P}_o|\mathbf{P}_d\right)$ for $j = 1, \cdots, K$. Then the Monte Carlo method approximates $\hat{\mathbf{x}}_{mmse}$ by the following expression

$$\hat{\mathbf{x}}_{mmse} \approx \frac{1}{K}\sum_{j=1}^{K} f\left(\mathbf{P}_o^{(j)}\right). \tag{3.9}$$

An immediate question is how we can sample from the conditional distribution $p\left(\mathbf{P}_o|\mathbf{P}_d\right)$. We answer the question in the next two sections.

3.2 Fusion for Two Nodes

In this section, we discuss the sampling method for the conditional distribution of the off-diagonal blocks when there are two nodes. In the case of known $\mathbf{P}_o$, the weight matrix $\mathbf{W}_1$ and $\mathbf{W}_2$ can be expressed as

$$\mathbf{W}_1 = \left(\mathbf{P}_{2,2} - \mathbf{P}_{1,2}^\top\right)\left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2} - \mathbf{P}_{1,2}^\top + \mathbf{P}_{2,2}\right)^{-1} \tag{3.10}$$

$$\mathbf{W}_2 = \left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2}\right)\left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2} - \mathbf{P}_{1,2}^\top + \mathbf{P}_{2,2}\right)^{-1}, \tag{3.11}$$

which are the weights for the optimal fusion in the mean square error sense. When we substitute (3.10) and (3.11) back into (2.7), we have

$$\hat{\mathbf{x}}_0 = \mathbf{W}_1\hat{\mathbf{x}}_1 + \mathbf{W}_2\hat{\mathbf{x}}_2. \tag{3.12}$$

$$\begin{aligned} &= \left(\mathbf{P}_{2,2} - \mathbf{P}_{1,2}^\top\right)\left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2} - \mathbf{P}_{1,2}^\top + \mathbf{P}_{2,2}\right)^{-1}\hat{\mathbf{x}}_1 \\ &\quad + \left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2}\right)\left(\mathbf{P}_{1,1} - \mathbf{P}_{1,2} - \mathbf{P}_{1,2}^\top + \mathbf{P}_{2,2}\right)^{-1}\hat{\mathbf{x}}_2. \end{aligned} \tag{3.13}$$

Hereafter, we use the notation $\mathbf{A}$ to represent the large covariance matrix. Suppose that a random matrix $\mathbf{A}$ is distributed according to $\mathcal{W}\left(n, \mathbf{\Sigma}\right)$. Let the partitions of the two positive definite matrices $\mathbf{A}$ and $\mathbf{\Sigma}$ be denoted by

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{1,2}^\top & \mathbf{A}_{2,2} \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{1,2}^\top & \mathbf{\Sigma}_{2,2} \end{bmatrix}. \tag{3.14}$$

Here we assume $\mathbf{\Sigma}_{1,2} = \mathbf{O}$. Recall that a Wishart matrix variate $\mathbf{A}$ can be expressed as $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$. $\mathbf{X}$ is a Gaussian random matrix, each column of which has the multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$. Therefore $\mathbf{\Sigma}_{1,2} = \mathbf{O}$ means that the upper part of each column in $\mathbf{X}$ is independent from the lower part. Our objective is to derive the expression for $p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right)$. We will need two properties of the Wishart distribution in our derivation [21]. To make it general enough, we assume that $\mathbf{A}_{1,1}$ is with size $L_1 \times L_1$ and $\mathbf{A}_{2,2}$ is with size $L_2 \times L_2$, $L_1 + L_2 = L$.

Lemma 1 *Let* $\mathbf{A}$ *and* $\mathbf{\Sigma}$ *be partitioned into* L_1 *and* L_2 *rows and columns as shown in* (3.14)*. If* $\mathbf{A}$ *is distributed according to* $\mathcal{W}_L\left(n,\mathbf{\Sigma}\right)$*, then* $\mathbf{A}_{1,1}$ *is distributed according to* $\mathcal{W}_{L_1}\left(n,\mathbf{\Sigma}_{1,1}\right)$*.*

Lemma 2 *If* $\mathbf{\Sigma}_{1,2} = \mathbf{O}$ *and* $\mathbf{A}$ *is distributed according to* $\mathcal{W}\left(n,\mathbf{\Sigma}\right)$*, then* $\mathbf{A}_{1,1}$ *and* $\mathbf{A}_{2,2}$ *are independently distributed.*

Lemma 1 provides the marginal distributions of $p\left(\mathbf{A}_{1,1}\right)$ and $p\left(\mathbf{A}_{2,2}\right)$ (they are $\mathcal{W}\left(n,\Sigma_{1,1}\right)$ and $\mathcal{W}\left(n,\Sigma_{2,2}\right)$, respectively). Lemma 2 maintains that $\mathbf{A}_{1,1}$ and $\mathbf{A}_{2,2}$ are independent. Therefore, $p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right)$ becomes

$$\begin{aligned} p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right) &= \frac{p\left(\mathbf{A}\right)}{p\left(\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right)} && (3.15)\\ &= \frac{p\left(\mathbf{A}\right)}{p\left(\mathbf{A}_{1,1}\right)p\left(\mathbf{A}_{2,2}\right)}. && (3.16)\end{aligned}$$

With a little algebraic manipulation, we have

$$\begin{aligned} p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right) =& Z\cdot\left|\mathbf{A}\right|^{\frac{n-L-1}{2}} && (3.17)\\ =& Z\cdot\left(\left|\mathbf{A}_{1,1}\right|\left|\mathbf{A}_{2,2}-\mathbf{A}_{1,2}^{\top}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\right|\right)^{\frac{n-L-1}{2}} && (3.18)\\ =& Z\cdot\left(\left|\mathbf{A}_{1,1}\mathbf{A}_{2,2}\right|\left|\mathbf{I}-\mathbf{A}_{2,2}^{-1}\mathbf{A}_{1,2}^{\top}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\right|\right)^{\frac{n-L-1}{2}}, && (3.19)\end{aligned}$$

where $n \geq L$, and the constant Z equals

$$\begin{aligned}
Z = & \frac{\left(\prod_{i=1}^{L_1} \Gamma\left(\frac{1}{2}(n+1-i)\right) \prod_{h=1}^{L_2} \Gamma\left(\frac{1}{2}(n+1-h)\right)\right)}{\prod_{j=1}^{L} \Gamma\left(\frac{1}{2}(n+1-j)\right)} \cdot \\
& \frac{1}{\pi^{\frac{L_1 L_2}{2}} \left|\mathbf{A}_{1,1}\right|^{\frac{n-L_1-1}{2}} \left|\mathbf{A}_{2,2}\right|^{\frac{n-L_2-1}{2}}} && (3.20) \\
= & \frac{\prod_{i=1}^{L_2} \Gamma\left(\frac{1}{2}(n+1-i)\right)}{\prod_{j=1+L_1}^{L} \Gamma\left(\frac{1}{2}(n+1-j)\right)} \cdot \frac{1}{\pi^{\frac{L_1 L_2}{2}} \left|\mathbf{A}_{1,1}\right|^{\frac{n-L_1-1}{2}} \left|\mathbf{A}_{2,2}\right|^{\frac{n-L_2-1}{2}}} && (3.21) \\
= & \frac{\prod_{i=1}^{L_2} \Gamma\left(\frac{1}{2}(n+1-i)\right)}{\prod_{j=1}^{L-L_1} \Gamma\left(\frac{1}{2}(n-L_1+1-j)\right)} \cdot \frac{1}{\pi^{\frac{L_1 L_2}{2}} \left|\mathbf{A}_{1,1}\right|^{\frac{n-L_1-1}{2}} \left|\mathbf{A}_{2,2}\right|^{\frac{n-L_2-1}{2}}} \\
& && (3.22) \\
= & \frac{\Gamma_{L_2}\left(\frac{n}{2}\right)}{\Gamma_{L_2}\left(\frac{1}{2}(n-L_1)\right)} \cdot \frac{1}{\pi^{\frac{L_1 L_2}{2}} \left|\mathbf{A}_{1,1}\right|^{\frac{n-L_1-1}{2}} \left|\mathbf{A}_{2,2}\right|^{\frac{n-L_2-1}{2}}}. && (3.23)
\end{aligned}$$

The above distribution is the inverted matrix variate t-distribution whose definition is as follows [23]:

Definition 1 *The random matrix $\mathbf{T} \in \mathbb{R}^{L \times M}$ is said to have an inverted matrix variate t-distribution with parameters $\mathbf{M} \in \mathbb{R}^{L \times M}$, $\mathbf{\Sigma} \in \mathbb{R}^{L \times L}$, $\mathbf{\Omega} \in \mathbb{R}^{M \times M}$ and n if its pdf is given by*

$$\begin{aligned}
p(\mathbf{T}) = & \frac{\Gamma_L\left(\frac{1}{2}(n+M+L-1)\right)}{\pi^{\frac{ML}{2}} \Gamma_L\left(\frac{1}{2}(n+L-1)\right)} \left|\mathbf{\Sigma}\right|^{-\frac{M}{2}} \left|\mathbf{\Omega}\right|^{-\frac{L}{2}} \\
& \left|\mathbf{I} - \mathbf{\Sigma}^{-1}(\mathbf{T}-\mathbf{M})\mathbf{\Omega}^{-1}(\mathbf{T}-\mathbf{M})^\top\right|^{\frac{n-2}{2}}, && (3.24)
\end{aligned}$$

where $\mathbf{\Omega} \succ 0$, $\mathbf{\Sigma} \succ 0$, $n > 0$ and $\mathbf{I} - \mathbf{\Sigma}^{-1}(\mathbf{T}-\mathbf{M})\mathbf{\Omega}^{-1}(\mathbf{T}-\mathbf{M})^\top \succ 0$. We denote this by $\mathbf{T} \sim \mathcal{IT}_{L,M}(n, \mathbf{M}, \mathbf{\Sigma}, \mathbf{\Omega})$.

For our case in (3.19), it is not difficult to obtain that

$$\mathbf{A}_{1,2}^\top | \mathbf{A}_{1,1}, \mathbf{A}_{2,2} \sim \mathcal{IT}_{L_2, L_1}\left(n - L + 1, \mathbf{O}, \mathbf{A}_{2,2}, \mathbf{A}_{1,1}\right). \qquad (3.25)$$

For sampling from the inverted matrix variate t-distribution, we use the following lemma [23]:

Lemma 3 *Let* $\mathbf{S} \sim \mathcal{W}_L(n+L-1, \mathbf{I}_L)$ *and* $\mathbf{X} \sim \mathcal{N}_{L,M}(0, \mathbf{I}_L \otimes \mathbf{I}_M)$ *be independently distributed. For* $\mathbf{M} \in \mathbb{R}^{L\times M}$, *define*

$$\mathbf{T} = \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\mathbf{S} + \mathbf{X}\mathbf{X}^\top\right)^{-\frac{1}{2}} \mathbf{X}\boldsymbol{\Omega}^{\frac{1}{2}} + \mathbf{M}, \tag{3.26}$$

where $\mathbf{S} + \mathbf{X}\mathbf{X}^\top = \left(\mathbf{S} + \mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}} \left(\left(\mathbf{S} + \mathbf{X}\mathbf{X}^\top\right)^{\frac{1}{2}}\right)^\top$ *and* $\boldsymbol{\Sigma}^{\frac{1}{2}}$ *and* $\boldsymbol{\Omega}^{\frac{1}{2}}$ *are the symmetric square roots of the positive definite matrices* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Omega}$, *respectively. Then,* $\mathbf{T} \sim \mathcal{IT}_{L,M}(n, \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$.

According to Lemma 3, the following theorem follows immediately.

Theorem 1 *Let the random matrices* $\mathbf{S} \sim \mathcal{W}_{L_2}(n-L_1, \mathbf{I}_{L_2})$ *and* $\mathbf{X} \sim \mathcal{N}_{L_2,L_1}(0, \mathbf{I}_{L_2} \otimes \mathbf{I}_{L_1})$. *If*

$$\mathbf{A}_{1,2}^\top = (\mathbf{A}_{2,2})^{\frac{1}{2}} \left(\mathbf{S} + \mathbf{X}\mathbf{X}^\top\right)^{-\frac{1}{2}} \mathbf{X} (\mathbf{A}_{1,1})^{\frac{1}{2}}, \tag{3.27}$$

then $\mathbf{A}_{1,2}^\top \sim p(\mathbf{A}_{1,2}|\mathbf{A}_{1,1}, \mathbf{A}_{2,2})$.

Remark. We can see that the hyperparameter $\boldsymbol{\Sigma}$ in the prior disappears in the conditional distribution as long as it is a block diagonal matrix. On the other hand, the degree of freedom n reflects the prior belief on the correlation between the two estimates. This can be used to exploit the available information for improved estimation.

3.3 Fusion for More Than Two Nodes

In this section, we consider the situation when we have three or more nodes. For multiple nodes, the conditional distribution of the off-diagonal submatrices is not inverted matrix variate t-distribution, and there is no way to directly sample from it. However we can proceed as follows.

Suppose we have N nodes, each measurement is an $M \times 1$ vector. The covariance matrix is

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,N} \\ \mathbf{A}_{1,2}^{\top} & \mathbf{A}_{2,2} & \cdots & \mathbf{A}_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{1,N}^{\top} & \mathbf{A}_{2,N}^{\top} & \cdots & \mathbf{A}_{N,N} \end{bmatrix}, \tag{3.28}$$

where $\mathbf{A}_{j,j} \in \mathbb{R}^{M\times M}$. We use $\mathbf{B}_j$ to denote

$$\mathbf{B}_j = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,j} \\ \mathbf{A}_{1,2}^{\top} & \mathbf{A}_{2,2} & \cdots & \mathbf{A}_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{1,j}^{\top} & \mathbf{A}_{2,j}^{\top} & \cdots & \mathbf{A}_{j,j} \end{bmatrix}. \tag{3.29}$$

The conditional distribution becomes

$$\begin{aligned} & p\left(\mathbf{A}_{1,2}, \mathbf{A}_{1,3}, \mathbf{A}_{2,3}, \cdots, \mathbf{A}_{N-1,N} \middle| \mathbf{A}_{1,1}, \cdots, \mathbf{A}_{N,N}\right) \\ = & \frac{p(\mathbf{A})}{p(\mathbf{A}_{1,1})p(\mathbf{A}_{2,2})\cdots p(\mathbf{A}_{N,N})} \end{aligned} \tag{3.30}$$

and there is no existing method for sampling from it. By repeatedly invoking Bayes chain rule, we can write the above density in this way:

$$\begin{aligned}
&p\left(\mathbf{A}_{1,2},\mathbf{A}_{1,3},\mathbf{A}_{2,3},\cdots,\mathbf{A}_{N-1,N}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)\\
=&p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)p\left(\mathbf{A}_{1,3},\mathbf{A}_{2,3},\cdots,\mathbf{A}_{N-1,N}|\mathbf{A}_{1,2},\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right) \qquad (3.31)\\
=&p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)p\left(\mathbf{A}_{1,3},\mathbf{A}_{2,3}|\mathbf{A}_{1,2},\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)\\
&\qquad p\left(\mathbf{A}_{1,4},\mathbf{A}_{2,4},\mathbf{A}_{3,4},\cdots,\mathbf{A}_{N-1,N}|\mathbf{A}_{1,3},\mathbf{A}_{2,3},\mathbf{A}_{1,2},\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right) \qquad (3.32)\\
=&p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)p\left(\mathbf{A}_{1,3},\mathbf{A}_{2,3}|\mathbf{A}_{1,2},\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)\\
&\qquad p\left(\mathbf{A}_{1,4},\mathbf{A}_{2,4},\mathbf{A}_{3,4},\cdots,\mathbf{A}_{N-1,N}|\mathbf{B}_{3},\mathbf{A}_{4,4},\cdots,\mathbf{A}_{N,N}\right) \qquad (3.33)\\
=&p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)\\
&p\left(\mathbf{A}_{1,3},\mathbf{A}_{2,3}|\mathbf{B}_{2},\mathbf{A}_{3,3},\cdots,\mathbf{A}_{N,N}\right)\\
&\cdots\\
&p\left(\mathbf{A}_{1,j},\mathbf{A}_{2,j},\cdots,\mathbf{A}_{j-1,j}|\mathbf{B}_{j-1},\mathbf{A}_{j,j},\cdots,\mathbf{A}_{N,N}\right)\\
&\cdots\\
&p\left(\mathbf{A}_{1,N},\cdots,\mathbf{A}_{N-1,N}|\mathbf{B}_{N-1},\mathbf{A}_{N,N}\right). \qquad (3.34)
\end{aligned}$$

Note that according to Lemma 1, we can write the conditional distribution according to

$$p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\cdots,\mathbf{A}_{N,N}\right)=p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right) \qquad (3.35)$$

or in general we have

$$p\left(\mathbf{A}_{1,j},\mathbf{A}_{2,j},\cdots,\mathbf{A}_{j-1,j}|\mathbf{B}_{j-1},\mathbf{A}_{j,j},\cdots,\mathbf{A}_{N,N}\right) \tag{3.36}$$

$$=p\left(\mathbf{A}_{1,j},\mathbf{A}_{2,j},\cdots,\mathbf{A}_{j-1,j}|\mathbf{B}_{j-1},\mathbf{A}_{j,j}\right). \tag{3.37}$$

Therefore, (3.34) becomes

$$p\left(\mathbf{A}_{1,2}|\mathbf{A}_{1,1},\mathbf{A}_{2,2}\right) \tag{3.38}$$

$$p\left(\mathbf{A}_{1,3},\mathbf{A}_{2,3}|\mathbf{B}_{2},\mathbf{A}_{3,3}\right) \tag{3.39}$$

$$\cdots$$

$$p\left(\mathbf{A}_{1,j},\mathbf{A}_{2,j},\cdots,\mathbf{A}_{j-1,j}|\mathbf{B}_{j-1},\mathbf{A}_{j,j}\right) \tag{3.40}$$

$$\cdots$$

$$p\left(\mathbf{A}_{1,N},\cdots,\mathbf{A}_{N-1,N}|\mathbf{B}_{N-1},\mathbf{A}_{N,N}\right). \tag{3.41}$$

Now things becomes easy for us since each factor in (3.38)-(3.41) is the inverted matrix variate t-distribution, which can be easily sampled from. Specifically, we can do it as follows.

According to Theorem 1, we can readily sample $\mathbf{A}_{1,2}$ according to (3.38). Then we sample $\mathbf{A}_{1,3},\mathbf{A}_{2,3}$ from (3.39). Let the random matrices $\mathbf{S}\sim\mathcal{W}_M\left(n-2M,\mathbf{I}_M\right)$ and $\mathbf{X}\sim\mathcal{N}_{M,2M}\left(0,\mathbf{I}_M\otimes\mathbf{I}_{2M}\right)$. Let also

$$\begin{bmatrix}\mathbf{A}_{1,3}^{\top}\\ \mathbf{A}_{2,3}^{\top}\end{bmatrix}=\mathbf{A}_{3,3}^{\frac{1}{2}}\left(\mathbf{S}+\mathbf{X}\mathbf{X}^{\top}\right)^{-\frac{1}{2}}\mathbf{X}\mathbf{B}_2^{\frac{1}{2}}, \tag{3.42}$$

where

$$\mathbf{B}_2 = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{1,2}^{\top} & \mathbf{A}_{2,2} \end{bmatrix}. \tag{3.43}$$

Then $[\mathbf{A}_{1,3}, \mathbf{A}_{2,3}] \sim p\left(\mathbf{A}_{1,3}, \mathbf{A}_{2,3} | \mathbf{B}_2, \mathbf{A}_{3,3}\right)$.

The process goes on for $k-1$ times. In the jth step, we sample $\mathbf{A}_{1,j+1}, \mathbf{A}_{2,j+1}, \cdots, \mathbf{A}_{j,j+1}$ from (3.40). Let the random matrices $\mathbf{S} \sim \mathcal{W}_M\left(n - jM, \mathbf{I}_M\right)$ and $\mathbf{X} \sim \mathcal{N}_{M,jM}\left(0, \mathbf{I}_M \otimes \mathbf{I}_{jM}\right)$. Let also

$$\begin{bmatrix} \mathbf{A}_{1,j+1}^{\top} \\ \mathbf{A}_{2,j+1}^{\top} \\ \vdots \\ \mathbf{A}_{j,j+1}^{\top} \end{bmatrix} = \mathbf{A}_{j+1,j+1}^{\frac{1}{2}} \left(\mathbf{S} + \mathbf{X}\mathbf{X}^{\top}\right)^{-\frac{1}{2}} \mathbf{X}\mathbf{B}_j^{\frac{1}{2}}, \tag{3.44}$$

then

$$\begin{bmatrix} \mathbf{A}_{1,j+1} \\ \mathbf{A}_{2,j+1} \\ \vdots \\ \mathbf{A}_{j,j+1} \end{bmatrix} \sim p\left(\mathbf{A}_{1,j+1}, \mathbf{A}_{2,j+1}, \cdots, \mathbf{A}_{j,j+1} | \mathbf{B}_j, \mathbf{A}_{j+1,j+1}\right). \tag{3.45}$$

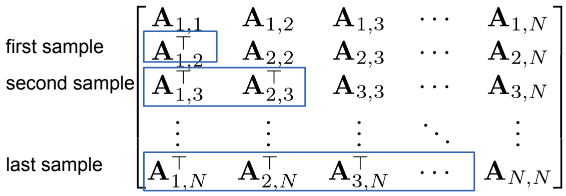
first sample
second sample
last sample
$\mathbf{A}_{1,1}$ $\mathbf{A}_{1,2}$ $\mathbf{A}_{1,3}$ $\cdots$ $\mathbf{A}_{1,N}$
$\mathbf{A}_{1,2}^{\top}$ $\mathbf{A}_{2,2}$ $\mathbf{A}_{2,3}$ $\cdots$ $\mathbf{A}_{2,N}$
$\mathbf{A}_{1,3}^{\top}$ $\mathbf{A}_{2,3}^{\top}$ $\mathbf{A}_{3,3}$ $\cdots$ $\mathbf{A}_{3,N}$
$\mathbf{A}_{1,N}^{\top}$ $\mathbf{A}_{2,N}^{\top}$ $\mathbf{A}_{3,N}^{\top}$ $\cdots$ $\mathbf{A}_{N,N}$

Figure 3.1: Illustration of the sampling of the off-diagonal block matrices.

Figure 3.1 shows the steps of the sampling algorithm. Before ending this section, we wish to emphasize that in situations with multiple nodes, fusing two nodes at a time using the method discussed in Section 3.2 will not work. By 'not work', we mean that fusing two nodes at a time and repeatedly doing this for multiple nodes is not equivalent to fusing multiple nodes at a time.

3.4 Numerical Experiments

In this section, we perform numerical experiment to test our algorithm. Suppose the true state $\mathbf{x}_0 = [0, 0]^\top$. We have N available measurements $\mathbf{x}_i$ for $i \in \{1, \cdots, N\}$. The measurements have normal distribution with covariance matrix $\mathbf{P}_x$. We generate $\mathbf{P}_x$ according to $\mathcal{W}(n, \sigma^2\mathbf{I})$ in each run, where $n = 3N$. Since $\mathbf{x}_0$ is assumed to be zero, the measurements are with zero mean. We carry out the experiment as follows. For each run, we first generate $\mathbf{P}_x$ according to the Wishart distribution and then sample from the corresponding normal distribution to get sample of $\mathbf{x}_i$. We suppose the diagonal blocks of $\mathbf{P}_x$ are known. Then the proposed method is used to calculate the weighting coefficients. We use 100 samples to estimate the integral ($K = 100$). Finally, we compare $\hat{\mathbf{x}}_0$ with $\mathbf{x}_0$, which is zero, to measure the performance. We compare the proposed method with the Covariance Intersection method and the optimal method. In the optimal method, we simply assume we know the entire covariance matrix. The result is averaged over 2000 instances of simulation.

Figure 3.2 shows the mean square error (MSE) performance for two nodes and Fig. 3.3 shows the performance for three nodes. The proposed method is roughly 10% better than the Covariance Intersection method

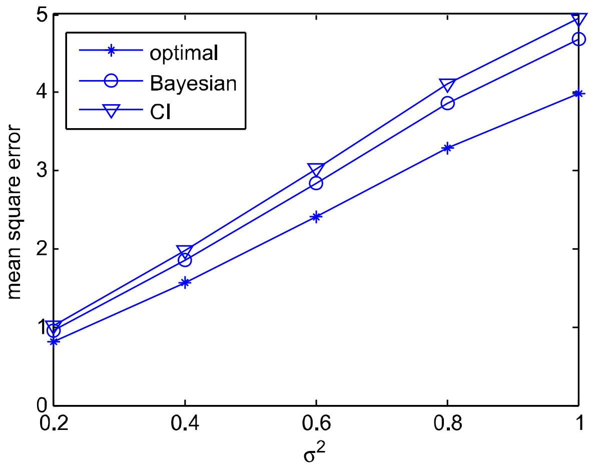

Figure 3.2: MSE of the three estimators. $(N = 2)$

in both situations. Figure. 3.4 shows the normalized mean square error performance, which is obtained by normalizing the MSE of the estimator using the MSE of the optimal estimator as a measure of scale. We see that in both situations, the proposed estimator outperforms the Covariance Intersection. However, with the number of nodes growing, the gap between the optimal estimator and the others becomes larger.

3.5 Discussion

In this chapter, we proposed a Bayesian approach to solve the data fusion problem in wireless sensor network when the cross-covariance between the estimates was not available. We first assumed that the prior of the covariance matrix was the Wishart distribution. Because we knew the covariance of each estimate, which was the diagonal block of the covariance matrix, we could obtain the conditional distribution of the off-diagonal

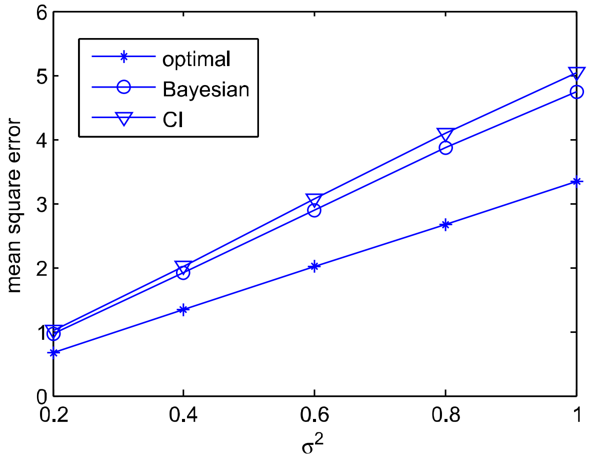

Figure 3.3: MSE of the three estimators. $(N = 3)$

block. For the case of two nodes, the conditional distribution of this block is the inverted matrix variate t-distribution. We also showed how to sample from this distribution. For the case of multiple nodes, the conditional distribution becomes much more complicated and there is no direct way to sample from it. We used the Bayes' chain rule to decompose the distribution into a product of several inverted matrix variate t-distribution so that we could still sample from it. As a result, we used the Monte Carlo method to compute the MMSE estimator. Numerical experiments showed that the performance of our method was better than that of the Covariance Intersection method. Another advantage of our algorithm is that under the Bayesian framework, we can modify the hyperparameter of the prior, the degree of freedom n, according to the available prior information, to make the algorithm perform better in some special cases.

The curious reader may wonder why we assumed the parameter $\mathbf{\Sigma}$ of the prior Wishart distribution $\mathcal{W}(n, \mathbf{\Sigma})$ to be a block diagonal matrix. The

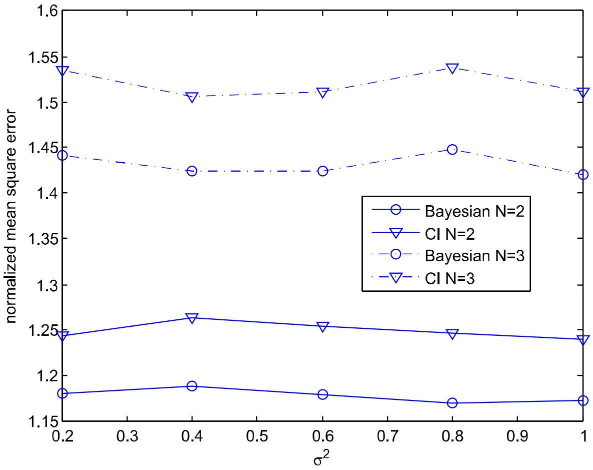

Figure 3.4: Normalized MSE of the two estimators.

reason is that by doing so, the diagonal blocks of the resulting covariance matrix are independent from each other. Otherwise, the joint distribution of the diagonal blocks are very complicated making the derivation of the conditional distribution of the off-diagonal blocks very difficult, if not impossible. We can see in the numerical experiment that the Wishart distribution with block diagonal parameter matrix $\mathbf{\Sigma}$ is still general enough to allow for good performance. However, if we can extend $\mathbf{\Sigma}$ to a general positive definite matrix, it would give us more freedom to manipulate the prior according to available information. This should be the direction of the future efforts.

The proposed method outperforms the Covariance Intersection when we compare the MSEs. However, we need to be cautious because the two methods use different criteria for obtaining the estimates. In some cases, it is possible that the Covariance Intersection works better than the proposed method. The purpose of our work is to provide an alternative to dealing with the difficult fusion issue in wireless sensor networks.

Chapter 4

Convex Optimization Approach

4.1 Introduction

In the previous chapter, a Bayesian method is applied to the covariance estimation problem. In this chapter, we introduce the convex optimization method to solve the problem. Our strategy is to estimate the cross-covariance first and then fuse the information from the various sources. We consider two cases, without and with priors: if we do not have any prior information about the covariance matrix, we can use the maximum-entropy (ME) principle as a criterion in the search for the optimal cross-covariance; if we have priors of the covariance matrices, we maximize the a posteriori distributions of these matrices. The problems in both cases can be formulated as convex optimization problems and therefore, they can readily be solved by some well-known methods.

4.2 Problem Model

The model considered in this chapter is a special case of the model in Chapter 2. Suppose that the true state $\mathbf{x}_0$ is a random vector with zero mean and covariance $\mathbf{C}_0$, and that $\mathbf{x}_i$ is the estimate of $\mathbf{x}_0$ corrupted by a zero mean noise with covariance $\mathbf{C}_i$, for $i \in \{1, \cdots, N\}$. The covariance matrix of $\mathbf{x}$, $\mathbf{P}_x$, which is defined in (2.6), becomes

$$\mathbf{P}_x = \begin{bmatrix} \mathbf{C}_1 + \mathbf{C}_0 & \mathbf{C}_0 & \cdots & \mathbf{C}_0 \\ \mathbf{C}_0 & \mathbf{C}_2 + \mathbf{C}_0 & \cdots & \mathbf{C}_0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_0 & \mathbf{C}_0 & \cdots & \mathbf{C}_N + \mathbf{C}_0 \end{bmatrix}. \tag{4.1}$$

In the fusion problem, we know the diagonal blocks of the covariance matrix, i.e., $\mathbf{P}_{i,i}$. Note $\mathbf{P}_{i,i} = \mathbf{C}_i + \mathbf{C}_0$. But we do not know $\mathbf{C}_i$ or $\mathbf{C}_0$. We wish to have an estimate of $\mathbf{C}_0$ so that we can determine the weighting coefficients for combining those $\mathbf{x}_i$s.

4.3 The Maximum Entropy approach

In this section, for the sake of simplicity we start with the Gaussian model and assume $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_0)$ and $\mathbf{x}_i | \mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_0, \mathbf{C}_i)$. Here we do not have the information about the priors of $\mathbf{C}_0$ and $\mathbf{C}_i$, and we propose to exploit the maximum-entropy principle. The rationale for using the ME principle is discussed thoroughly in [24, 25].

The entropy is basically a functional, i.e., it maps a function f to a real number. It is defined as

$$H(f) = -\int f(\mathbf{y}) \log f(\mathbf{y})\, \mathrm{d}\mathbf{y}. \tag{4.2}$$

Plugging (3.3) in (4.2), we obtain the entropy of the multivariate Gaussian distribution,

$$\begin{aligned} H(f) &= -\int_{\mathbf{y}\in\mathbb{R}^M} f(\mathbf{y}) \left(-\log\left((2\pi)^{M/2} |\mathbf{C}|^{1/2} \right) - \frac{1}{2}\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} \right) \mathrm{d}\mathbf{y} \quad (4.3) \\ &= \log\left((2\pi)^{M/2} |\mathbf{C}|^{1/2} \right) + \int_{\mathbf{y}\in\mathbb{R}^M} \frac{1}{2} f(\mathbf{y})\, \mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y}\mathrm{d}\mathbf{y}. \quad (4.4) \end{aligned}$$

Let

$$\begin{aligned} \mathbf{y} &= [y_1, \cdots, y_M]^\top \quad (4.5) \\ \mathbf{D} = \mathbf{C}^{-1} &= \begin{bmatrix} c_{1,1} & \cdots & c_{1,M} \\ \vdots & \ddots & \vdots \\ c_{1,M} & \cdots & c_{M,M} \end{bmatrix}^{-1} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,M} \\ \vdots & \ddots & \vdots \\ d_{1,M} & \cdots & d_{M,M} \end{bmatrix}. \quad (4.6) \end{aligned}$$

We have

$$\mathbf{y}^\top \mathbf{C}^{-1}\mathbf{y} = \sum_{i=1}^{M}\sum_{j=1}^{M} d_{i,j} y_i y_j. \tag{4.7}$$

Therefore the second term in (4.4) becomes

$$
\begin{aligned}
&\int_{\mathbf{y}\in\mathbb{R}^M} f(\mathbf{y})\,\mathbf{y}^\top\mathbf{C}^{-1}\mathbf{y}d\mathbf{y} && (4.8)\\
=&\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{y})\sum_{i=1}^{M}\sum_{j=1}^{M} d_{i,j}y_iy_j\ \mathrm{d}y_1\cdots dy_2 && (4.9)\\
=&\sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,j}c_{i,j} && (4.10)\\
=&\mathrm{tr}\left[\mathbf{DC}\right] && (4.11)\\
=&M. && (4.12)
\end{aligned}
$$

Thus the maximization of $H(f)$ reduces to the maximization of $\log(|\mathbf{C}|)$.

In order to estimate the cross-covariance, we try to maximize the entropy of the joint distribution of $\mathbf{x}_0$ and $\mathbf{x}_i$. Specifically, we try to maximize $H\left(p_{\mathbf{x}_0,\mathbf{x}_1,\cdots,\mathbf{x}_N}\right)$. We have

$$
\begin{aligned}
H\left(p_{\mathbf{x}_0,\mathbf{x}_1,\cdots,\mathbf{x}_N}\right) =&H\left(p_{\mathbf{x}_0}\right)+\sum_{i=1}^{N}H\left(p_{\mathbf{x}_i|\mathbf{x}_0}\right) && (4.13)\\
&\propto\log\left(|\mathbf{C}_0|\right)+\sum_{i=1}^{N}\log\left(|\mathbf{C}_i|\right) && (4.14)\\
&=\log\left(|\mathbf{C}_0|\right)+\sum_{i=1}^{N}\log\left(|\mathbf{P}_{i,i}-\mathbf{C}_0|\right), && (4.15)
\end{aligned}
$$

where the first equality can be found in [26]. Therefore, the entire optimization problem can be formulated as

$$
\begin{aligned}
\text{maximize}\quad & \log\left(|\mathbf{C}_0|\right)+\sum_{i=1}^{N}\log\left(|\mathbf{P}_{i,i}-\mathbf{C}_0|\right) && (4.16)\\
\text{subject to}\quad & \mathbf{P}_{i,i}-\mathbf{C}_0\succeq 0 \qquad i\in\{1,\cdots,N\} && (4.17)\\
& \mathbf{C}_0\succeq 0, && (4.18)
\end{aligned}
$$

where the variable is the symmetric matrix $\mathbf{C}_0$. The objective function (4.16) is a convex function on the positive semidefinite cone [27]; the constraints (4.17) and (4.18) are convex sets. Thus the optimization problem can be easily solved by some existing well-known methods, e.g., the interior point method [28].

Before ending this section, we wish to emphasize that the model does not have to be normal. In fact, it can be unknown as long as the first and the second moments are specified. Recall that the normal distribution has ME among all real-valued distributions with given mean and variance [29]. That is to say, even if the model is unknown, we still obtain the same solution if we employ the ME criterion.

4.4 The Maximum Posterior Approach

In this section, we consider the case where the priors of the unknown covariance matrices are available. Suppose the priors of $\mathbf{C}_0$ and $\mathbf{C}_i$ are $p_0\left(\mathbf{C}_0\right) = \mathcal{W}_M\left(\mathbf{C}_0|L_0, \Sigma_0\right)$ and $p_i\left(\mathbf{C}_i\right) = \mathcal{W}_M\left(\mathbf{C}_i|L_i, \Sigma_i\right)$, respectively. We use the maximum a posteriori (MAP) distribution as a criterion, and the optimal estimator can be written as

$$\mathbf{C}_0 = \max_{\mathbf{C}\succeq 0} \arg\, p_0\left(\mathbf{C}\right) \prod_{i=1}^{N} p_i\left(\mathbf{P}_{i,i} - \mathbf{C}\right). \tag{4.19}$$

If we substitute (3.1) into (4.19), we have

$$\mathbf{C}_0 = \max_{\mathbf{C}\succeq 0} \arg\, g\left(\mathbf{C}\right), \tag{4.20}$$

where $g(\mathbf{C})$ is defined as

$$\begin{aligned} g(\mathbf{C}) = & \frac{L_0 - M - 1}{2} \log|\mathbf{C}| - \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_0^{-1}\mathbf{C}\right] \\ & + \sum_{i=1}^{N} \frac{L_i - M - 1}{2} \log|\mathbf{P}_{i,i} - \mathbf{C}| \\ & - \sum_{i=1}^{N} \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}(\mathbf{P}_{i,i} - \mathbf{C})\right]. \end{aligned} \tag{4.21}$$

The optimization problem can be cast as

$$\begin{aligned} \text{maximize} \quad & g(\mathbf{C}) && (4.22) \\ \text{subject to} \quad & \mathbf{P}_{i,i} - \mathbf{C} \succeq 0 \qquad i \in \{1, \cdots, N\} && (4.23) \\ & \mathbf{C} \succeq 0, && (4.24) \end{aligned}$$

where the optimization variable is the symmetric matrix $\mathbf{C}$. We know $\log|\cdot|$ is a concave function and $\mathrm{tr}[\cdot]$ is a convex function over the positive semidefinite cone [27]. Therefore $g(\mathbf{C})$ is a concave function with respect to $\mathbf{C}$. Since the constraint also specifies a convex set, the problem is a convex optimization problem as well, which can be solved with no difficulty.

4.5 Discussion

We point out that (4.16) and (4.21) are both with the same $\log(|\cdot|)$ terms. In fact, (4.16) is a special case of (4.21), where the hyperparameters $\boldsymbol{\Sigma}_i$s are infinitely large and make the terms $\mathrm{tr}\left[\boldsymbol{\Sigma}_0^{-1}\mathbf{C}\right]$ and $\mathrm{tr}\left[\boldsymbol{\Sigma}_i^{-1}(\mathbf{P}_{i,i} - \mathbf{C})\right]$ vanish for finite $\mathbf{P}_{i,i}$ and $\mathbf{C}$.

To illustrate the connections further, we first associate an ellipsoid to each covariance matrix. The ellipsoid of $\mathbf{A}$ is defined as

$$\{\mathbf{x} \in \mathbb{R}^M | \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} = 1\}. \tag{4.25}$$

For $N = 2$, the ellipsoid becomes an ellipse. Basically the major and minor axes of the ellipse show how large the variances are in the directions of the axes. The angle between the x-axis and the major axis of the ellipse indicates how much the data from the two dimensions correlate with each other. Figure 4.1 shows the ellipses of $\mathbf{P}_{1,1}, \mathbf{P}_{2,2}$ and the estimated cross-covariances. We set $N = 2$ for simplicity and let

$$\mathbf{P}_{1,1} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 3 \end{bmatrix}, \quad \mathbf{P}_{2,2} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.26}$$

The priors are $\mathbf{C}_0 \sim \mathcal{W}_2(4, \sigma_0^2 \mathbf{I}_2)$, $\mathbf{C}_i \sim \mathcal{W}_2(4, \sigma^2 \mathbf{I}_2)$. The ellipses for different matrices are shown in Figure 4.1. We can see the ellipses of $\mathbf{C}_0$ are inside those of $\mathbf{C}_1$ and $\mathbf{C}_2$, which makes sense since any point in the feasible set shall make its associated ellipse in the intersection of those of the $\mathbf{P}_{i,i}$s. Also, for larger σ_0^2/σ^2 the solution ellipse becomes larger; for smaller σ_0^2/σ^2, the ellipse becomes smaller. When both σ_0^2 and σ^2 are large, in this case $\sigma_0^2 = \sigma^2 = 10$, the ellipse (green) is very close to the ME solution (red).

4.6 Simulation

We use the Gaussian model in the numerical experiment. Suppose that the variable to be estimated is $\mathbf{x}_0$ and that it has distribution

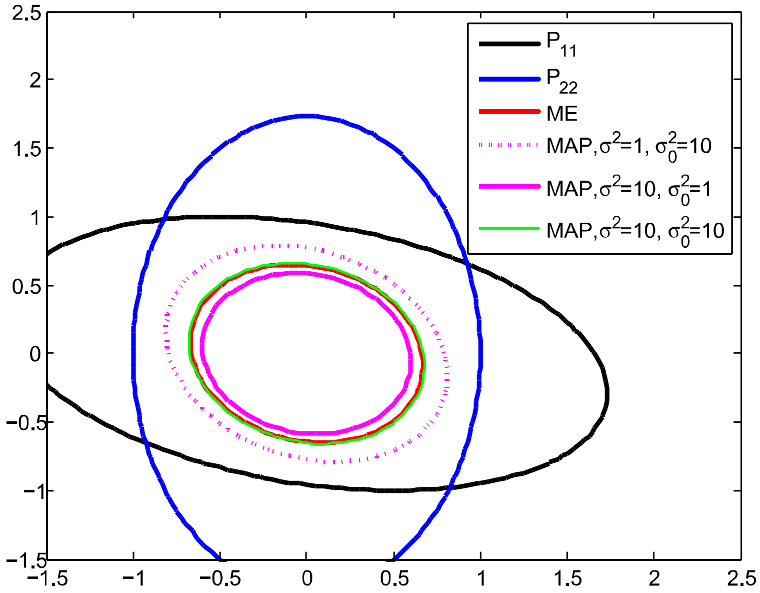

Figure 4.1: Illustration of $\mathbf{P}_{1,1}$, $\mathbf{P}_{2,2}$ and the estimated covariances.

$\mathcal{N}(\mathbf{m}_0, \mathbf{C}_0)$. We let $\mathbf{m}_0 = \mathbf{0}$ for the sake of simplicity. The estimates $\mathbf{x}_i$ have the conditional distributions $\mathcal{N}(\mathbf{x}_0, \mathbf{C}_i)$ for $i \in \{1, \cdots, N\}$. The noise of the measurements is assumed to be independent of each other. We can consider $\mathbf{x}_i$ to be measurements as well as estimates, since we shall let $\hat{\mathbf{x}}_i = \mathbf{x}_i$ if we make estimation only based on $\mathbf{x}_i$. If we concatenate N estimates into one vector $\mathbf{x}$ as before, the distribution of the vector conditioned on $\mathbf{x}_0$ is

$$\mathbf{x}|\,\mathbf{x}_0 \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_0\\ \vdots\\ \mathbf{x}_0\end{bmatrix}, \begin{bmatrix}\mathbf{C}_1 & \cdots & \mathbf{O}\\ \vdots & \ddots & \vdots\\ \mathbf{O} & \cdots & \mathbf{C}_N\end{bmatrix}\right). \tag{4.27}$$

The marginal distribution of $\mathbf{x}$ becomes

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{P}_x), \tag{4.28}$$

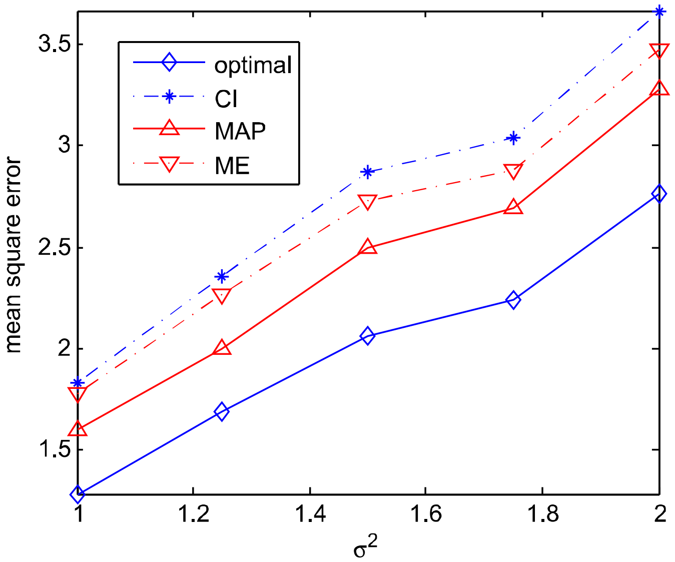

Figure 4.2: Performance comparison in the case of two estimates.

where $\mathbf{P}_x$ is defined in (4.1). The diagonal blocks $\mathbf{C}_i + \mathbf{C}_0$ are known exactly. On the other hand, neither $\mathbf{C}_0$ nor $\mathbf{C}_i$ is known.

To generate the data for our numerical experiment, we first draw $\mathbf{C}_0$ from its prior $\mathcal{W}_2\left(2, \mathbf{\Lambda}_1\right)$ and $\mathbf{C}_1, \cdots, \mathbf{C}_N$ from $\mathcal{W}_2\left(2, \sigma^2\mathbf{\Lambda}_2\right)$ independently, where

$$\mathbf{\Lambda}_1 = \begin{bmatrix} 4 & -1 \\ -1 & 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{\Lambda}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}. \tag{4.29}$$

Then we generate the true value $\mathbf{x}_0$ by sampling from $\mathcal{N}\left(\mathbf{0}, \mathbf{C}_0\right)$. Similarly, we generate the measurements $\mathbf{x}_i$ from $\mathcal{N}\left(\mathbf{x}_0, \mathbf{C}_i\right)$. We set $N = 2, 3$. Now we have all the data we need for testing and comparing the estimators. For comparison, we use three other estimators, the optimal estimator (2.20) with all the information (including $\mathbf{C}_0$), and the fast Covariance Intersection method (2.4) from [19]. For each configuration, we ran 2000 tests. In the legend, we use *optimal, ME, MAP*, and *CI* to indicate

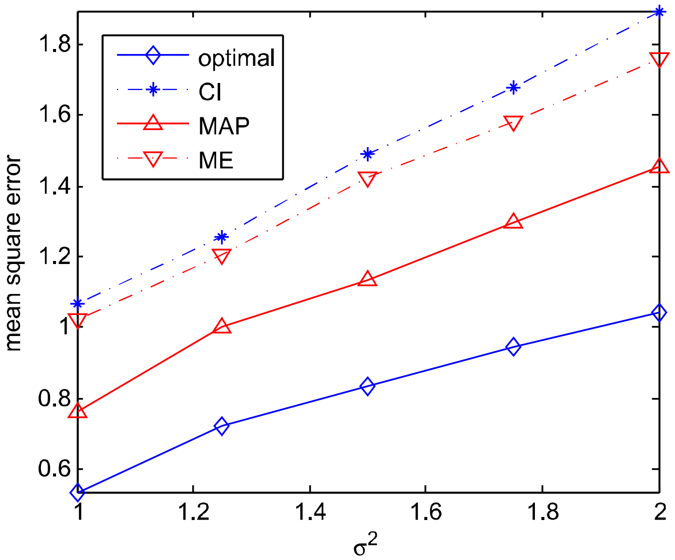

Figure 4.3: Performance comparison in the case of three estimates.

the optimal method, the ME method, the MAP method, and the fast Covariance Intersection method, respectively.

Figure 4.2 and Fig. 4.3 show the mean square error performance for $N = 2$ and $N = 3$, respectively. We can see that the proposed methods are better than the CI method in both situations. Meanwhile, as the hyperparameters $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_i$ are much different, the MAP estimator outperforms the ME estimator thanks to its priors.

4.7 Conclusion

In this chapter, we proposed convex optimization techniques to solve the fusion of correlated estimates with unknown correlations. Specifically, given the diagonal block of the error covariance matrix, we cast the problem of estimating cross-covariance as convex optimization problem which could be readily solved by well-known methods. Two cases were considered: for the non-Bayesian case, we employed the maximum entropy criterion in the search for optimal cross-covariance; for the Bayesian case, we assumed that the priors of the unknown covariance matrices were the Wishart distribution. We then maximized the posterior probability of the cross-covariance. As soon as the cross-covariance was obtained, the weighting coefficients could be determined and the distributed estimates could be combined by a simple calculation. With numerical experiments we demonstrated the performance of our methods.

Chapter 5

Sequential Estimation in Networks

5.1 Introduction

In the previous three chapters, the data fusion problem with unknown correlation is considered. Now we turn our focus on the sequential learning problem, where the parameter of interest is static and each node obtains a local observation at each time slot. The objective is to estimate the parameter in a distributed way. This problem has been studied in [30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. However, they all use search-like methods which is far from being optimal. In [35], the authors compare the mean-square performance of two main strategies for distributed estimation: consensus strategies and diffusion strategies. They claim that the diffusion leads to faster convergence and lower mean-square deviation than consensus. Note that when the parameter of interest is dynamic, it becomes sequential filtering problem, which is another popular category of distributed estimation problems. See [5, 3, 40] for example.

In this chapter, we propose an efficient estimation algorithm for the case where the noises are correlated. This algorithm is neither diffusion nor consensus. But it is closer to the latter. We use a doubly stochastic matrix to combine the information from different nodes at each iteration. We prove that our algorithm approaches the optimal centralized least square estimator asymptotically.

5.2 Introduction and formulation

The problem is mathematically formulated as follows. The network is represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges, respectively. Two nodes exchange information only if there is an edge between them. There are N nodes in the network, namely $N = |\mathcal{V}|$. Let $\boldsymbol{\theta} \in \mathbb{R}^{L\times 1}$ be a static parameter vector of interest. At time instant t, the observation at node i is modeled as

$$\mathbf{y}_{i,t} = \mathbf{H}_{i,t}\boldsymbol{\theta} + \mathbf{w}_{i,t}, \tag{5.1}$$

where $\mathbf{w}_{i,t}, \mathbf{y}_{i,t} \in \mathbb{R}^{M\times 1}$, $\mathbf{H}_{i,t} \in \mathbb{R}^{M\times L}$; $\mathbf{H}_{i,t}$ is the observation matrix; $\mathbf{y}_{i,t}$ is the observation; and $\mathbf{w}_{i,t}$ is a Gaussian noise vector. The mean and covariance of the noise are

$$\mathbb{E}[\mathbf{w}_{i,t}] = \mathbf{0}_M, \tag{5.2}$$

$$\mathbb{E}[\mathbf{w}_{i,t}\mathbf{w}_{j,s}^\top] = \delta_{t,s}\boldsymbol{\Sigma}_{i,j}, \tag{5.3}$$

where $\delta_{t,s}$ is the Kronecker delta function; $\mathbf{0}_M$ is an M-size column vector with all of its elements being zero. Let

$$\mathbf{H}_t = \begin{bmatrix} \mathbf{H}_{1,t} \\ \vdots \\ \mathbf{H}_{N,t} \end{bmatrix}, \tag{5.4}$$

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_{1,t} \\ \vdots \\ \mathbf{y}_{N,t} \end{bmatrix}, \tag{5.5}$$

$$\mathbf{w}_t = \begin{bmatrix} \mathbf{w}_{1,t} \\ \vdots \\ \mathbf{w}_{N,t} \end{bmatrix}, \tag{5.6}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{N,1} & \cdots & \boldsymbol{\Sigma}_{N,N} \end{bmatrix}. \tag{5.7}$$

Thus, $\mathbf{H}_t \in \mathbb{R}^{NM \times L}$, $\mathbf{y}_t \in \mathbb{R}^{NM \times 1}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{NM \times NM}$. The matrix $\boldsymbol{\Sigma}$ is assumed to be strictly positive definite. Then the entire model can be expressed as

$$\mathbf{y}_t = \mathbf{H}_t \boldsymbol{\theta} + \mathbf{w}_t, \tag{5.8}$$

where $\mathbf{w}_t$ is zero mean white Gaussian noise with covariance matrix $\mathbf{\Sigma}$. From all the observations received at t, the least squares estimator is [41]

$$\hat{\boldsymbol{\theta}}_t = \left(\mathbf{H}_t^\top \mathbf{\Sigma}^{-1} \mathbf{H}_t\right)^{-1} \mathbf{H}_t^\top \mathbf{\Sigma}^{-1} \mathbf{y}_t. \tag{5.9}$$

We assume that $\mathbf{w}_{i,t}$ satisfies the Markov property with respect to the graph $\mathcal{G}$, i.e., the noises of any pair of nonadjacent nodes are conditionally independent given the remaining noise values,

$$p\left(\mathbf{w}_{i,t}, \mathbf{w}_{j,t} | \mathbf{w}_{\mathcal{V}\setminus i,j}\right) = p\left(\mathbf{w}_{i,t} | \mathbf{w}_{\mathcal{V}\setminus i,j}\right) p\left(\mathbf{w}_{j,t} | \mathbf{w}_{\mathcal{V}\setminus i,j}\right)$$
$$\text{for all } \ \{i,j\} \notin \mathcal{E}, \text{and for all } \ t \in \mathbb{N}. \tag{5.10}$$

In the sequel we use $\mathbf{K} = \mathbf{\Sigma}^{-1}$ and refer to it as a precision matrix. Since $\mathbf{w}_{i,t}$ is Gaussian, we have [42]

$$\mathbf{K}_{i,j} = \mathbf{O} \quad \text{for all } \ \{i,j\} \notin \mathcal{E}, \tag{5.11}$$

where $\mathbf{K}_{i,j}$ is the (i,j)th block of $\mathbf{K}$ and $\mathbf{O}$ is a matrix with zero elements and of the same size as $\mathbf{K}$. Given the observations from the beginning to time instant t, the least squares estimate $\tilde{\boldsymbol{\theta}}_t$ can be expressed as

$$\tilde{\boldsymbol{\theta}}_t = \left(\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{H}_s\right)^{-1} \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{y}_s, \tag{5.12}$$

where $\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{y}_s$ represents the sufficient statistics of the model. Our objective is to calculate $\tilde{\boldsymbol{\theta}}_t$ in a distributed way. To make the problem well-defined, we need to make the following mild assumptions:

1. The vector sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ is bounded,

2. The matrix sequence $\{\mathbf{H}_t\}_{t\in\mathbb{N}}$ is bounded, full rank and does not converge to a rank deficit matrix.

There is work in the literature that is related to what is addressed here, notably the running consensus [43, 44, 45, 46, 47, 48] and the distributed diffusion [30, 31, 32, 36]. The running consensus is in fact a special case of the proposed method. If we let $L = M = 1$, $H_{i,t} = 1$ and the noises be i.i.d., the proposed method reduces to the running consensus. We point out that the introduction of the multidimensional time-varying observation matrices makes the problem much complicated and the proof of efficiency nontrivial. In the model used with the distributed diffusion method, the observation matrix is replaced by a vector, and therefore the observation is a scalar. However, by concatenation of the measurements from different time slots, the models become basically the same. More important differences are that here the noises between neighboring nodes are correlated and that we use a doubly stochastic matrix to assign the mixing weighting coefficients. We need such assignment to allow the distributed estimator achieve the global optimum.

5.3 The proposed distributed estimation algorithm

In this section, we describe how the distributed estimator works. We assume node i has access to $\mathbf{H}_{j,t}$ for $j \in \mathcal{N}_i$ through communication at time instant t. $\mathcal{N}_i$ stands for the neighbors of node i. Let $\mathbf{Q} \in \mathbb{R}^{N\times N}$ be a irreducible and aperiodic doubly stochastic matrix, which satisfies

$$\mathbf{Q}\mathbf{1}_N = \mathbf{1}_N, \quad \mathbf{1}_N^\top \mathbf{Q} = \mathbf{1}_N^\top. \tag{5.13}$$

Denote by $Q_{i,j}$ and $Q_{i,j}^t$ the (i,j)th entry of $\mathbf{Q}$ and $\mathbf{Q}^t$, respectively. We note that $Q_{i,j} = 0$ if nodes i and j are not connected. Such $\mathbf{Q}$ can be constructed by letting $\mathbf{Q} = \mathbf{I}_N - \epsilon \boldsymbol{\Xi}$, where $\boldsymbol{\Xi}$ is the Laplacian matrix of the graph $\mathcal{G}$; ϵ is a coefficient satisfying $\epsilon < 1/\max_i(\deg(i))$, with $\deg(i)$ denoting the degree of node i. Note that

$$\lim_{t\to\infty} Q_{i,j}^t = \frac{1}{N} \quad \text{for} \;\; i,j \in \{1,\cdots,N\}. \tag{5.14}$$

This is the principle we use behind the averaging of the sufficient statistics.

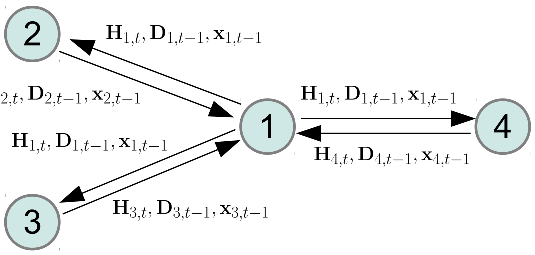
2
$\mathbf{H}_{1,t}, \mathbf{D}_{1,t-1}, \mathbf{x}_{1,t-1}$
$\mathbf{H}_{2,t}, \mathbf{D}_{2,t-1}, \mathbf{x}_{2,t-1}$
$\mathbf{H}_{1,t}, \mathbf{D}_{1,t-1}, \mathbf{x}_{1,t-1}$
1
4
$\mathbf{H}_{1,t}, \mathbf{D}_{1,t-1}, \mathbf{x}_{1,t-1}$
$\mathbf{H}_{4,t}, \mathbf{D}_{4,t-1}, \mathbf{x}_{4,t-1}$
3
$\mathbf{H}_{3,t}, \mathbf{D}_{3,t-1}, \mathbf{x}_{3,t-1}$

Figure 5.1: Information exchange at time instant t.

In the distributed algorithm, each node keeps two variables, the matrix $\mathbf{D}_i \in \mathbb{R}^{L\times L}$ and the vector $\mathbf{x}_i \in \mathbb{R}^{L\times 1}$, which approximate $\sum_{s=1}^t \mathbf{H}_s^\top \mathbf{K}\mathbf{H}_s$ and $\sum_{s=1}^t \mathbf{H}_s^\top \mathbf{K}\mathbf{y}_s$, respectively. The method is based on the following formulas:

$$
\begin{aligned}
\mathbf{D}_{i,t} &= \sum_{j\in\mathcal{N}_i} Q_{i,j}\mathbf{D}_{j,t-1} + \sum_{j\in\mathcal{N}_i} \mathbf{H}_{j,t}^{\top}\mathbf{K}_{j,i}\mathbf{H}_{i,t} && (5.15)\\
&= \sum_{s=1}^{t}\sum_{j=1}^{N} Q_{i,j}^{t-s} \sum_{k\in\mathcal{N}_j} \mathbf{H}_{k,s}^{\top}\mathbf{K}_{k,j}\mathbf{H}_{j,s}, && (5.16)\\
\mathbf{x}_{i,t} &= \sum_{j\in\mathcal{N}_i} Q_{i,j}\mathbf{x}_{j,t-1} + \sum_{j\in\mathcal{N}_i} \mathbf{H}_{j,t}^{\top}\mathbf{K}_{j,i}\mathbf{y}_{i,t} && (5.17)\\
&= \sum_{s=1}^{t}\sum_{j=1}^{N} Q_{i,j}^{t-s} \sum_{k\in\mathcal{N}_j} \mathbf{H}_{k,s}^{\top}\mathbf{K}_{k,j}\mathbf{y}_{j,s}, && (5.18)\\
\tilde{\boldsymbol{\theta}}_{i,t} &= \mathbf{D}_{i,t}^{-1}\mathbf{x}_{i,t}, && (5.19)
\end{aligned}
$$

and where $\mathbf{Q}^0$ is defined to be the identity matrix. The information a node transmits to its neighbors includes $\mathbf{H}_{i,t}, \mathbf{D}_{i,t}$ and $\mathbf{x}_{i,t}$ (see Fig. 5.1). We note that the centralized estimate is given by

$$\tilde{\boldsymbol{\theta}}_t = \left(\sum_{s=1}^{t} \mathbf{H}_s^{\top}\mathbf{K}\mathbf{H}_s\right)^{-1} \sum_{s=1}^{t} \mathbf{H}_s^{\top}\mathbf{K}\mathbf{y}_s. \tag{5.20}$$

The factors $\mathbf{D}_{i,t}$ and $\mathbf{x}_{i,t}$ in (5.19) are approximations of $\sum_{s=1}^{t} \mathbf{H}_s^{\top}\mathbf{K}\mathbf{H}_s$ and $\sum_{s=1}^{t} \mathbf{H}_s^{\top}\mathbf{K}\mathbf{y}_s$, respectively. Let $\mathbf{Q}_{(i)}^t \in \mathbb{R}^{NM\times NM}$ be defined by

$$\mathbf{Q}_{(i)}^t = \begin{bmatrix} Q_{i,1}^t\mathbf{I}_M & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & Q_{i,2}^t\mathbf{I}_M & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & Q_{i,N}^t\mathbf{I}_M \end{bmatrix}. \tag{5.21}$$

Then $\mathbf{D}_{i,t}$ and $\mathbf{x}_{i,t}$ can be expressed as

$$\begin{aligned} \mathbf{D}_{i,t} &= \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{Q}_{(i)}^{t-s} \mathbf{H}_s, && (5.22) \\ \mathbf{x}_{i,t} &= \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{Q}_{(i)}^{t-s} \mathbf{y}_s. && (5.23) \end{aligned}$$

Thus, (5.19) can also be written as

$$\tilde{\boldsymbol{\theta}}_{i,t} = \left(\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{Q}_{(i)}^{t-s} \mathbf{H}_s \right)^{-1} \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{Q}_{(i)}^{t-s} \mathbf{y}_s. \tag{5.24}$$

It has been proved that this estimator is unbiased [49]. The main theorem in this section is that the proposed method has the same asymptotic performance as the centralized one, which can be formally expressed as

$$\lim_{t \to \infty} \frac{\operatorname{tr}\left[\operatorname{Cov}\left[\tilde{\boldsymbol{\theta}}_{i,t}\right]\right]}{\operatorname{tr}\left[\operatorname{Cov}\left[\tilde{\boldsymbol{\theta}}_{t}\right]\right]} = 1. \tag{5.25}$$

We prove the main theorem in the appendix.

5.4 Simulation

In this section, we test the algorithms in a network with 20 nodes. The topology is shown in Fig. 5.2. We let $\boldsymbol{\theta} = [1, -1]$. All the entries in $\mathbf{H}$ are i.i.d. Gaussian variables, and $\epsilon = 0.1$. $\mathbf{K}$ is generated by summing a group of Wishart random matrices that correspond to the cliques in the graph. Figure 5.3 shows the mean square error performance of an implementation for the setting. We can see that the performance of the proposed method approaches the centralized estimator after 300 iterations.

Figure 5.4 shows the averaged results of 500 runs. The performance of the proposed estimator converges quickly to the centralized estimator.

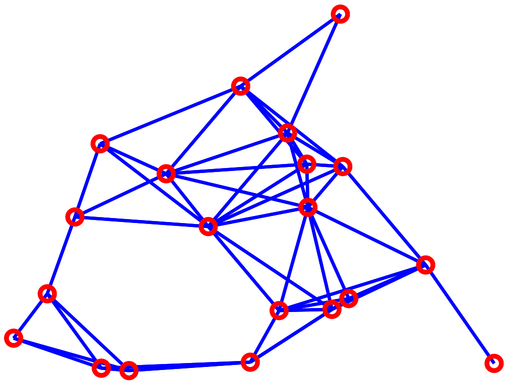

Figure 5.2: Topology of the network.

5.5 Conclusion

In this chapter, we proposed a distributed sequential algorithm for the case that the noises are correlated. We assumed that the noises had the conditional independence property, i.e., given the noise values of the neighbors of a node, the noise at the node was independent of other noises in the network. We showed that the proposed algorithm was asymptotically equivalent to the centralized algorithm, regardless of the actual values of observation, as long as they were bounded. Since the centralized estimator (5.12) is an efficient estimator [41], the proposed estimator therefore asymptotically approaches the Cramér-Rao bound. The simulations justify the statement.

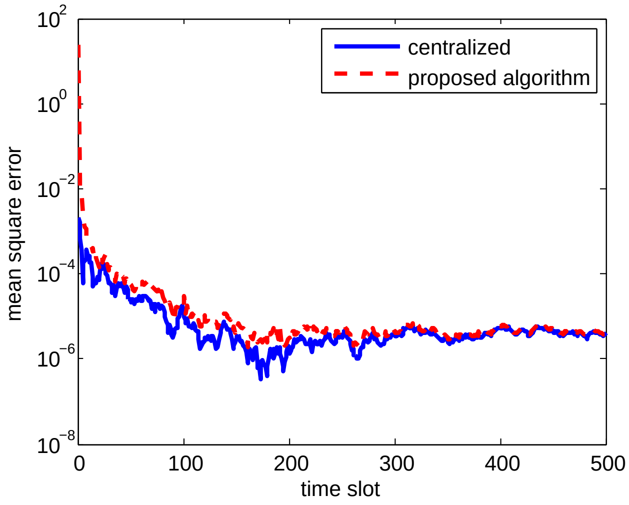

Figure 5.3: Performance of a sample run.

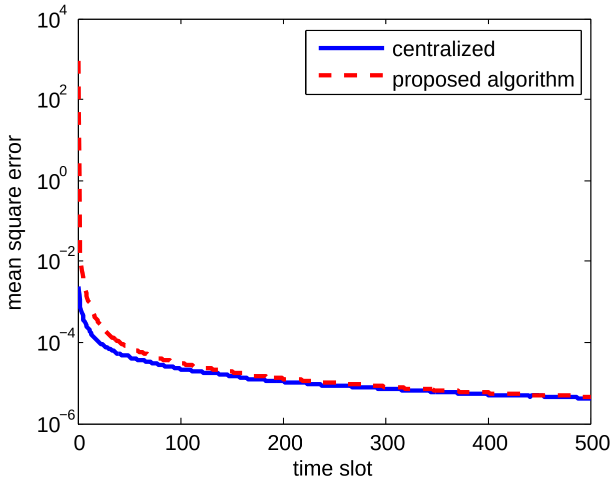

Figure 5.4: Averaged performance over 500 instances.

Chapter 6

Belief Consensus

6.1 Introduction

In this chapter, we study the belief consensus problem. In consensus estimation, a node in the network performs global estimation tasks through iterative information exchange with its neighbors and update its own state based on the received information. Most consensus problems that have been studied fall into several basic categories. The original consensus problems study how a group of nodes, each with a piece of belief, reach agreement by local information exchange. One of the earliest work on consensus problems is [50], where consensus for discrete distribution is studied. In [51] and [52], convex optimization techniques are used to accelerate the convergence. In [53], consensus problems in dynamic network with time-delays are investigated. In [54], asymmetric interaction mechanism with time-varying weights are introduced to increase the convergence rate of the consensus. In [55], the authors try to find the mixing matrix that leads to the highest convergence rate. In [56], efforts are put in the search for structures that accelerate the convergence. An optimal control scheme for achieving consensus is proposed in [57]. In [58], the authors investigate a broadcasting-based gossiping algorithm to compute the average of the initial measurements of the nodes. In [59], the authors examine the problem of designing weights when the network is subject to random link failures and switching topology. In [60], consensus is used to solve the distributed total least squares problem.

In this chapter, we consider the consensus of continuous densities in the Bayesian framework. Ideally, we would like the nodes to reach consensus at the density equal to the product of all the belief densities of the nodes because if we assume a noninformative prior, the product of all the initial densities is just the unnormalized Bayesian posterior. We show that the Bayesian posterior is not achievable without the knowledge of the network size. To approximate the Bayesian posterior, we employ the weighted product of the densities. We use the χ^2 information metric as the criterion to choose the weighting coefficients in the fusion of the continuous densities. The method is general for all probability distributions. We then confine ourselves to Gaussian cases and show that the χ^2 information function is convex with respect to the weighting coefficients. Very few works consider the consensus of continuous densities. In [61], the authors cast the problem in a Bayesian framework and adopt an information-theoretic approach to data fusion by using the Kullback-Leibler average of the density functions. Here, not only the criteria, but also the way we formulate the problems are different.

The chapter is organized as follows. The problem is formulated in Section 6.2. In Section 6.3, we introduce the proposed method. In

Section 6.4, we give some insight into the proposed method. We present experimental results in Section 6.5, and conclude in Section 6.6 . Proofs are left to the appendix.

6.2 Problem Formulation

Consider a network with N nodes. Each node has an initial belief about an unknown parameter of interest $\mathbf{x}$. In traditional consensus problems, the initial belief is usually a point estimate and we fuse them using the criteria like maximum likelihood or minimum mean square error. In this chapter, we assume that each node has a belief in the form of a continuous density instead of a point estimate. In such cases, those criteria used in point estimation are no longer applicable. Therefore, we propose the use of χ^2 information as the criterion and formulate the problem as follows.

Let the belief of node i be $p_i(\mathbf{x})$. Ideally, with all the beliefs of N nodes, the best belief we can have about $\mathbf{x}$ is the Bayesian posterior

$$p_c(\mathbf{x}) = \frac{\prod_i p_i(\mathbf{x})}{\int_{\mathbf{x}} \prod_i p_i(\mathbf{x})\, d\mathbf{x}} \tag{6.1}$$

with the assumption of a noninformative prior. Taking logarithm of both sides, we have

$$\log p_c(\mathbf{x}) = \sum_i \log p_i(\mathbf{x}) - \log \int_{\mathbf{x}} \prod_i p_i(\mathbf{x})\, d\mathbf{x}. \tag{6.2}$$

Note that the second term is a normalizing constant and its value depends on the first term. If we know the value of N, we can first try to achieve consensus at the average, i.e., $\frac{1}{N}\log p_c(\mathbf{x})$, and then multiply the average by N. To make the consensus converge at the average, we can simply use a doubly stochastic matrix where each row corresponds to the weighting coefficients a node assigns to its neighbors. Without knowing N, this approach is not possible.

In this work, we use the weighted product of the densities to approximate the Bayesian posterior. The weighted product of the densities is expressed as

$$p_d(\mathbf{x}) = \frac{\prod_i p_i^{w_i}(\mathbf{x})}{\int_{\mathbf{x}} \prod_i p_i^{w_i}(\mathbf{x})\, d\mathbf{x}} \tag{6.3}$$

where $\sum_i w_i = 1$. The symbol w_i is the exponent of $p_i(\mathbf{x})$. This form is also referred to as generalized fusion in [62]. The discussion so far is general for any probability density functions. However, for general continuous densities, the computation of (6.3) is not tractable, unless the function is parametric. Hereafter, we confine ourselves to the Gaussian cases. We assume the initial belief of each node is a Gaussian density with mean $\mathbf{m}_i$ and covariance $\mathbf{C}_i$, denoted by $\mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i)$. We note that given N Gaussian densities $\mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i)$ for $i \in \{1, \cdots N\}$, the product $\prod_{i=1}^N \mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i)$ is still a Gaussian. Denote by $\mathcal{N}(\mathbf{x}|\mathbf{m}_c, \mathbf{C}_c)$ the Bayesian posterior, i.e.,

$$\mathcal{N}(\mathbf{x}|\mathbf{m}_c, \mathbf{C}_c) = \frac{1}{Z_c}\prod_{i=1}^{N} \mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i), \tag{6.4}$$

where Z_c is the normalizing coefficient. Then the mean $\mathbf{m}_c$ and the covariance matrix $\mathbf{C}_c$ can be easily derived [63]:

$$\mathbf{C}_c = \left(\sum_{i=1}^{N} \mathbf{C}_i^{-1}\right)^{-1}, \tag{6.5}$$

$$\mathbf{m}_c = \mathbf{C}_c \left(\sum_{i=1}^{N} \mathbf{C}_i^{-1}\mathbf{m}_i\right). \tag{6.6}$$

Likewise, the product (6.3) is also a Gaussian density function. Let $\mathcal{N}\left(\mathbf{x}|\mathbf{m}_d, \mathbf{C}_d\right)$ denote the weighted product of the densities,

$$\mathcal{N}\left(\mathbf{x}|\mathbf{m}_d, \mathbf{C}_d\right) = \frac{1}{Z_d} \prod_{i=1}^{N} \mathcal{N}^{w_i}\left(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i\right) \tag{6.7}$$

with

$$\sum_{i=1}^{N} w_i = 1,\, w_i \geq 0, \tag{6.8}$$

we will have

$$\mathbf{C}_d = \left(\sum_{i=1}^{N} w_i \mathbf{C}_i^{-1}\right)^{-1} \tag{6.9}$$

$$\mathbf{m}_d = \left(\sum_{i=1}^{N} w_i \mathbf{C}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{N} w_i \mathbf{C}_i^{-1}\mathbf{m}_i\right). \tag{6.10}$$

The inverse of the covariance matrix is called the precision matrix, denoted by $\mathbf{K}$. Thus we have

$$\mathbf{K}_c = \sum_{i=1}^{N} \mathbf{K}_i, \tag{6.11}$$

$$\mathbf{m}_c = \left(\sum_{i=1}^{N} \mathbf{K}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{K}_i \mathbf{m}_i\right) \tag{6.12}$$

and

$$\mathbf{K}_d = \sum_{i=1}^{N} w_i \mathbf{K}_i \tag{6.13}$$

$$\mathbf{m}_d = \left(\sum_{i=1}^{N} w_i \mathbf{K}_i \right)^{-1} \sum_{i=1}^{N} w_i \mathbf{K}_i \mathbf{m}_i. \tag{6.14}$$

6.3 Consensus optimization

In the previous section, we have introduced the equations (6.13) and (6.14) for belief fusion in the Gaussian cases. However, the weighting coefficients, w_i, are yet to be chosen. In this section, we discuss the methods for determining the values of the weighting coefficients. First, we revisit the famous Covariance Intersection algorithm. Then we discuss how we use the χ^2 information as a criterion to determine the weighting coefficients.

6.3.1 Covariance Intersection

We have discussed the Covariance Intersection (CI) in Section 2.1. The objective of the Covariance Intersection was to obtain a consistent estimate of the covariance matrix when multiple random variables were linearly combined without knowing the correlation. CI selects the value of w_i such that the determinant or trace of $\mathbf{C}_d$ is minimized. In [62], it has been pointed out that the criterion used in CI is equivalent to minimizing the Shannon information of the fused function with the assumption that the fusion functions are Gaussian. The optimization problem can be expressed as

$$\text{minimize} \ -\int_{\mathbf{x}} p_d(\mathbf{x}) \log(p_d(\mathbf{x}))\, d\mathbf{x} \tag{6.15}$$

$$\text{subject to} \ \sum_{i=1}^{N} w_i = 1 \tag{6.16}$$

$$w_i \geq 0, \tag{6.17}$$

where w_i for $i = 1, \cdots, N$ are the variables.

6.3.2 χ^2 information metric

To approximate the Bayesian posterior (6.1) by the weighted product of the densities (6.3), we must first have a criterion to measure the difference between two densities. Here we adopt the χ^2 information as the metric. We seek the weighting coefficients w_i such that the χ^2 information [64] between the Bayesian posterior and the weighted product of the densities is minimized. The χ^2 information between density p and q is defined as

$$\chi^2(p||q) = \int \frac{p^2(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} - 1. \tag{6.18}$$

Note that like the Kullback–Leibler divergence, it is not a symmetric measure of the difference between two densities. However, in our case, we put $p_c(\mathbf{x})$ first and use $\chi^2(p_c(\mathbf{x})|p_d(\mathbf{x}))$ instead of $\chi^2(p_d(\mathbf{x})|p_c(\mathbf{x}))$. The reason is that for Gaussian $p_c(\mathbf{x})$ and $p_d(\mathbf{x})$, $p_c^2(\mathbf{x})/p_d(\mathbf{x})$ will still be Gaussian as long as the covariance matrix $\left(2\mathbf{C}_c^{-1} - \mathbf{C}_d^{-1}\right)^{-1}$ is positive definite. According to the definition of $\mathbf{C}_d$ and $\mathbf{C}_c$ in (6.13) and (6.11), respectively, it is easy to check that $\left(2\mathbf{C}_c^{-1} - \mathbf{C}_d^{-1}\right)^{-1}$ is positive definite but $\left(2\mathbf{C}_d^{-1} - \mathbf{C}_c^{-1}\right)^{-1}$ is not. In other words, $p_c^2(\mathbf{x})/p_d(\mathbf{x})$ is Gaussian but

$p_d^2(\mathbf{x})/p_c(\mathbf{x})$ is not.

Formally, we try to minimize

$$\chi^2\left(p_c(\mathbf{x})\,||p_d(\mathbf{x})\right) \tag{6.19}$$

with respect to w_i under the constraint $\sum_{i=1}^{N} w_i = 1$. The χ^2 information can be derived as

$$\chi^2\left(\mathcal{N}(\mathbf{m}_c, \mathbf{C}_c)\,||\mathcal{N}(\mathbf{m}_d, \mathbf{C}_d)\right) = T_2 \exp(T_1) - 1, \tag{6.20}$$

where

$$T_1 = -\frac{1}{2}(\mathbf{m}_c - \mathbf{m}_d)^\top \left(\frac{\mathbf{C}_c}{2} - \mathbf{C}_d\right)^{-1}(\mathbf{m}_c - \mathbf{m}_d) \tag{6.21}$$

$$T_2 = \frac{\sqrt{|\mathbf{C}_d|}}{|\mathbf{C}_c|}\sqrt{\left|\left(2\mathbf{C}_c^{-1} - \mathbf{C}_d^{-1}\right)^{-1}\right|}. \tag{6.22}$$

The optimization problem becomes

$$\begin{aligned} \text{minimize} \quad & \chi^2\left(\mathcal{N}(\mathbf{m}_c, \mathbf{C}_c)\,||\mathcal{N}(\mathbf{m}_d, \mathbf{C}_d)\right) && (6.23)\\ \text{subject to} \quad & \sum_{i=1}^{N} w_i = 1 && (6.24)\\ & w_i \geq 0, && (6.25) \end{aligned}$$

where w_i are the variables. In the appendix, we show that the objective function is convex.

6.3.3 Application in the network

Our assumption is that the total number of nodes in the network is unknown and each node is only able to communicate with its neighbors. This means the optimization of weighting coefficients can only be performed locally. We specify some additional notations: $w_{j,i}$ means the weighting coefficient node j assign to the belief from node i; let $p_{j,c}(\mathbf{x})$ be the local Bayesian posterior

$$p_{j,c}(\mathbf{x}) = \frac{\prod_{i \in \mathcal{N}_j} p_i(\mathbf{x})}{\int_{\mathbf{x}} \prod_{i \in \mathcal{N}_j} p_i(\mathbf{x})\, d\mathbf{x}}, \tag{6.26}$$

and $p_{j,d}(\mathbf{x})$ the local weighted product of the densities

$$p_{j,d}(\mathbf{x}) = \frac{\prod_{i \in \mathcal{N}_j} p_i^{w_{j,i}}(\mathbf{x})}{\int_{\mathbf{x}} \prod_{i \in \mathcal{N}_j} p_i^{w_{j,i}}(\mathbf{x})\, d\mathbf{x}}. \tag{6.27}$$

At each iteration node j optimizes the following problem

$$\text{minimize } \chi^2 \left(p_{j,c}(\mathbf{x}) || p_{j,d}(\mathbf{x})\right) \tag{6.28}$$

$$\text{subject to } \sum_{i \in \mathcal{N}_j} w_{j,i} = 1 \tag{6.29}$$

$$w_{j,i} > 0. \tag{6.30}$$

6.4 Analysis

In this section, we provide further insight into the objective function of the optimization problem. We consider two extreme cases and compare them with the case of uniform weighting coefficients. By uniform we mean the weighting coefficients are all equal to $\frac{1}{N}$. We consider the first extreme case where $\mathbf{C}_i = \mathbf{C}_0$ for all i. Then we have

$$\mathbf{C}_c^{-1} = N\mathbf{C}_0^{-1} \tag{6.31}$$

$$\mathbf{C}_d^{-1} = \mathbf{C}_0^{-1} \tag{6.32}$$

$$\mathbf{m}_c = \frac{1}{N}\sum_i \mathbf{m}_i \tag{6.33}$$

$$\mathbf{m}_d = \sum_i w_i \mathbf{m}_i. \tag{6.34}$$

We can see that in this case T_2 does not depend on w_i. To maximize the χ^2 information, we only need to look at T_1,

$$T_1 = -\frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \left(\frac{\mathbf{C}_c}{2} - \mathbf{C}_d\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right) \tag{6.35}$$

$$= \frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \left(\mathbf{C}_0 - \frac{\mathbf{C}_0}{2N}\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right) \tag{6.36}$$

$$= \frac{N}{2N-1}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \mathbf{C}_0^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right). \tag{6.37}$$

Since $\mathbf{C}_0$ is positive definite and T_1 is a quadratic form, T_1 achieves minimum when $\mathbf{m}_c = \mathbf{m}_d$. This happens when $w_i = \frac{1}{N}$ for all i, which reduces to the uniform approach.

Next we consider the second extreme case when $\mathbf{m}_i = \mathbf{m}_0$ for all i. For the sake of simplicity, we only consider the scalar case. In the scalar case, we assume the mean and variance of node i are m_0 and σ_i^2, respectively. The centralized variances and means become

$$\sigma_c^2 = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^{-1} \tag{6.38}$$

$$m_c = \sigma_c^2 \left(\sum_{i=1}^{N} \frac{m_0}{\sigma_i^2}\right) = m_0 \tag{6.39}$$

and the decentralized ones are

$$\sigma_d^2 = \left(\sum_{i=1}^{N} \frac{w_i}{\sigma_i^2} \right)^{-1} \tag{6.40}$$

$$m_d = \sigma_d^2 \left(m_0 \sum_{i=1}^{N} \frac{w_i}{\sigma_i^2} \right) = m_0. \tag{6.41}$$

Because $m_c = m_d$, T_1 becomes constant and we only need to look at T_2.

$$T_2 = \frac{\sqrt{\sigma_d^2}}{\sigma_c^2} \sqrt{\left(\frac{2}{\sigma_c^2} - \frac{1}{\sigma_d^2} \right)^{-1}} \tag{6.42}$$

$$= \frac{1}{\sigma_c^2 \sqrt{\frac{1}{\sigma_d^2} \left(\frac{2}{\sigma_c^2} - \frac{1}{\sigma_d^2} \right)}} \tag{6.43}$$

$$= \frac{1}{\sigma_c^2 \sqrt{\sum_{i=1}^{N} \frac{w_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{2 - w_i}{\sigma_i^2}}}. \tag{6.44}$$

In order to minimize T_2 with respect to w_i, we only need to investigate the part in the square root. Suppose

$$f\left(w_i\right) = \sum_{i=1}^{N} \frac{w_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{2 - w_i}{\sigma_i^2}. \tag{6.45}$$

Our problem becomes

$$\text{maximize} f\left(w_i\right) \tag{6.46}$$

$$\text{subject to} \sum_{i=1}^{N} w_i = 1 \tag{6.47}$$

$$w_i \geq 0. \tag{6.48}$$

Because we intend to analyze the problem instead of solving it, we ignore the second constraint. The reason will be clear later. To find the minimum value of $f\left(w_i\right)$, we use the method of Lagrange multipliers. The Lagrangian becomes

$$\mathcal{L}\left(w_i,\lambda\right) = f\left(w_i\right) + \lambda\left(\sum_i w_i - 1\right). \tag{6.49}$$

Take derivative of $\mathcal{L}\left(w_i,\lambda\right)$ with respect to w_i, and we have

$$\frac{\partial\mathcal{L}\left(w_i,\lambda\right)}{\partial w_j} = \frac{1}{\sigma_j^2}\left(\sum_{i=1}^{N}\frac{2-w_i}{\sigma_i^2}\right) - \frac{1}{\sigma_j^2}\left(\sum_{i=1}^{N}\frac{w_i}{\sigma_i^2}\right) - \lambda \tag{6.50}$$

for all j. In order that the derivative be equal to 0 for all j, we must have

$$\sum_{i=1}^{N}\frac{2-w_i}{\sigma_i^2} = \sum_{i=1}^{N}\frac{w_i}{\sigma_i^2}. \tag{6.51}$$

Substitute (6.51) into $f\left(w_i\right) = 0$. We have

$$\left(\sum_{i=1}^{N} w_i\frac{1}{\sigma_i^2}\right)^2 = 0 \tag{6.52}$$

which leads to

$$\sum_{i=1}^{N} w_i\frac{1}{\sigma_i^2} = 0. \tag{6.53}$$

To make (6.53) equal to zero, at least one w_i has to be zero. But recall that w_i should all be nonnegative. This tells us that there is no extreme point in the simplex defined by (6.47) and (6.48). Moreover, we can conclude that the optimal value must lie on the boundary of the simplex, that is at least one w_i is zero. This makes the problem reduce to $N-1$ dimensions. With the same reasoning, we can reduce the problem to two dimensions, and it is easy to see that the w_i with the smaller σ_i^2 wins. We can conclude that the solution to the optimization problem in (6.46) is $w_j = 1$ with the assumption that σ_j^2 is the smallest variance and $w_i = 0$ for $i \neq j$. Although this analysis is for scalar cases only, it provides intuitive insight for fusion with close mean values.

In summary, when the covariances from different nodes are equal, the proposed method reduces to the method of uniform weighting coefficients. For scalar cases, when all the nodes share the same mean value, the optimal value of the weighting coefficients lies on a vertex of the simplex, i.e., only the w_j with the smallest σ_j^2 is equal to one, the other coefficients are zero. For vector cases, the solution becomes complicated but it more or less follows a similar pattern, which works like the minimum operator.

6.5 Numerical Experiment

In this section, we provide numerical experiments to show the performance of the proposed methods. The experiment is carried out as follows. In a network, each node has an initial belief. At each iteration, a node collects the beliefs from its neighbors and fuses the beliefs according to (6.11)-(6.14). The weighting coefficients w_i used in the fusion are determined by solving the optimization problem expressed in (6.23). In other words, we use local optimal value to approximate the global Bayesian posterior. As we have mentioned before, the reason is that we do not know the number of nodes in the network. The initial belief of each node is generated as follows. The mean values are generated according to a zero-mean multivariate normal distribution with covariance being an identity matrix. The covariance matrices are generated according to a Wishart distribution with degree 4 and identity scale matrix. We average the performance over 500 instances of simulation for each experiment. We compare the proposed method with the Covariance Intersection and the uniform weighting method. In the uniform weighting method, each node assigns equal weighting coefficients to its neighbors and itself. To evaluate the performance, we examine the normalized average χ^2 information between the global Bayesian posterior and the local weighted product of the densities. We denote the average χ^2 information between the belief of each node and the global Bayesian posterior at time t by

$$\frac{1}{N}\sum_{j=1}^{N}\chi_t^2\left(p_c\left(\mathbf{x}\right)||p_{j,d}\left(\mathbf{x}\right)\right). \tag{6.54}$$

Then we *normalize* (6.54) by the χ^2 information at the beginning. Therefore, the normalized average χ^2 information can be expressed as

$$\frac{\sum_{j=1}^{N}\chi_t^2\left(p_c\left(\mathbf{x}\right)||p_{j,d}\left(\mathbf{x}\right)\right)}{\sum_{j=1}^{N}\chi_0^2\left(p_c\left(\mathbf{x}\right)||p_{j,d}\left(\mathbf{x}\right)\right)}.$$

Besides, we also look at the normalized MSE of the mean and covariance values. Suppose $p_c\left(\mathbf{x}\right)$ is $\mathcal{N}\left(\mathbf{m}_c, \mathbf{C}_c\right)$ and the local belief $p_{j,d}\left(\mathbf{x}\right)$ at time t is $\mathcal{N}\left(\mathbf{m}_{j,d}\left(t\right), \mathbf{C}_{j,d}\left(t\right)\right)$. Then the normalized MSE of the mean and covariance values can be expressed as

$$\frac{\sum_{j=1}^{N}\left\|\mathbf{m}_{j,d}\left(t\right)-\mathbf{m}_c\right\|^2}{\sum_{j=1}^{N}\left\|\mathbf{m}_{j,d}\left(0\right)-\mathbf{m}_c\right\|^2}$$

and

$$\frac{\sum_{j=1}^{N}\left\|\mathbf{C}_{j,d}\left(t\right)-\mathbf{C}_c\right\|^2}{\sum_{j=1}^{N}\left\|\mathbf{C}_{j,d}\left(0\right)-\mathbf{C}_c\right\|^2},$$

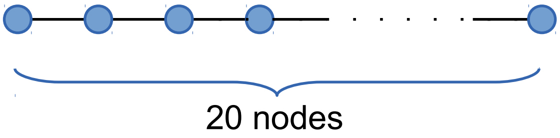

Figure 6.1: Topology of the first experiment.

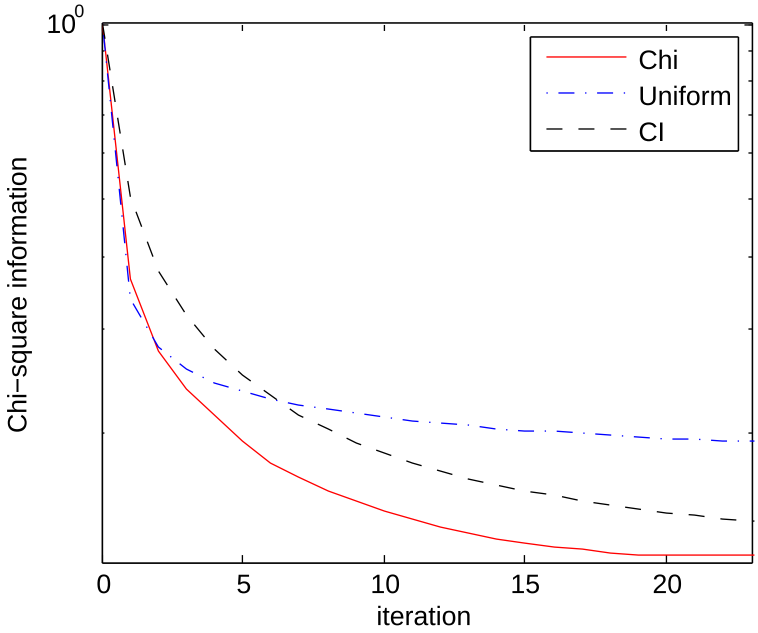

Figure 6.2: Comparison of the normalized χ^2 information.

respectively.

In the first experiment, the topology is a line as shown in Fig. 6.1. Each node is only able to talk to its immediate neighbors. Figure 6.2 shows the average *normalized* χ^2 information between the belief of each node and the global Bayesian posterior. Figure 6.3 shows the MSE of the mean values versus the iterations. Figure 6.4 shows the MSE of the covariance matrices versus the iterations.

In the second experiment, we change the topology to be a 5 by 5 grid

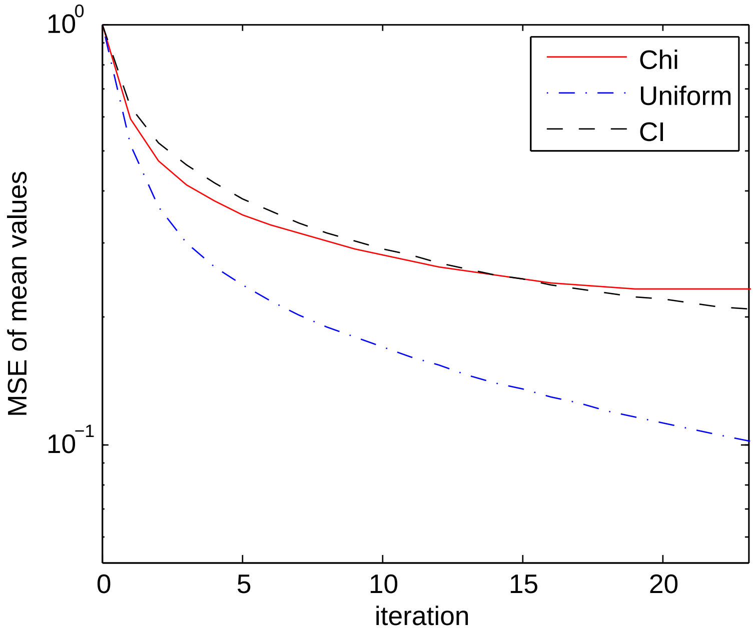

Figure 6.3: Comparison of the normalized MSE of the mean values.

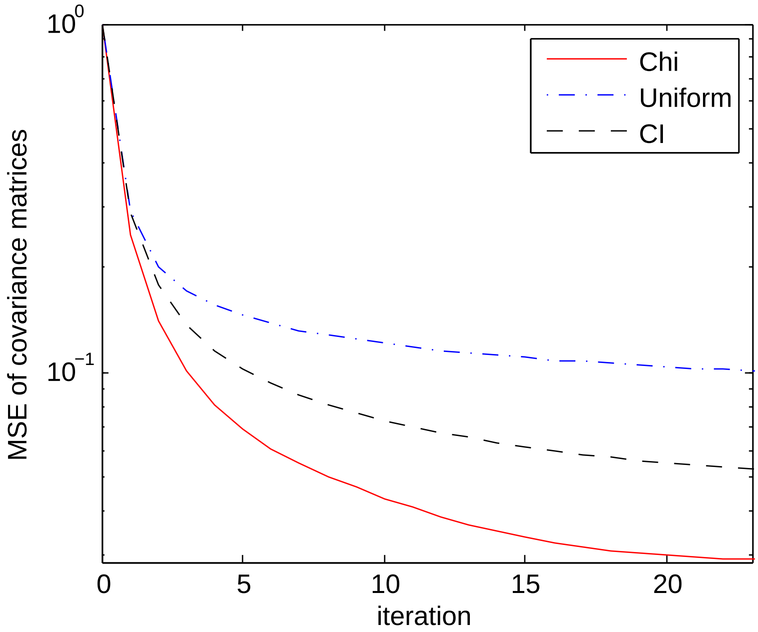

Figure 6.4: Comparison of the normalized MSE of the covariance matrices.

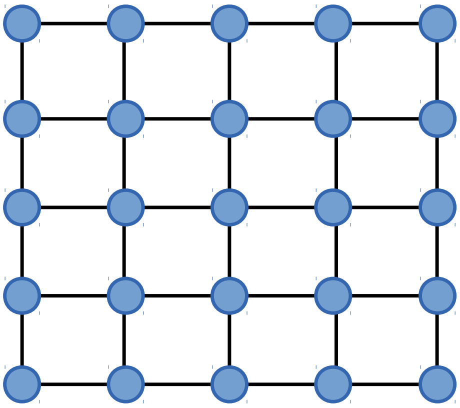

Figure 6.5: The topology of the network in the second experiment.

as shown in Fig. 6.5. Each node is only able to talk to its immediate right, left, upper and lower neighbors. Figure 6.6 shows the χ^2 information versus the iterations. Figure 6.7 shows the MSE of the mean values versus the iterations. Figure 6.8 shows the MSE of the covariance matrices versus the iterations.

In both experiments, we can see that the proposed method outperforms the others in the χ^2 information and the MSE of covariance matrices. But the uniform weighting method beats the other methods in MSE of the mean values. This is not surprising because in the fusion process, the fused mean value is a linear combination of the mean values from different nodes. Uniform strategy will not differ too much as the point defined by the weighting coefficients is near the center in the solution domain. On the other hand, the fusion of the covariance matrix is not linear. The proposed method puts more weight on the covariance matrix in the optimization.

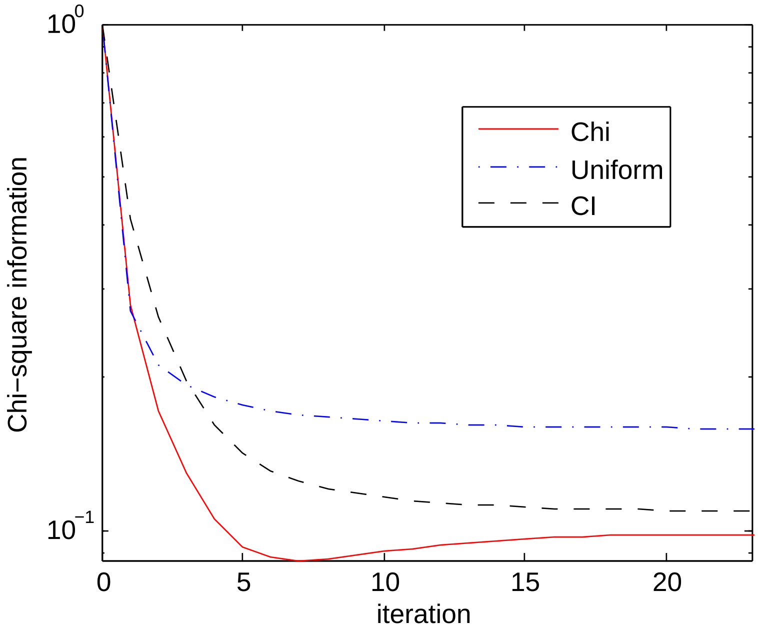

Figure 6.6: Comparison of the normalized χ^2 information.

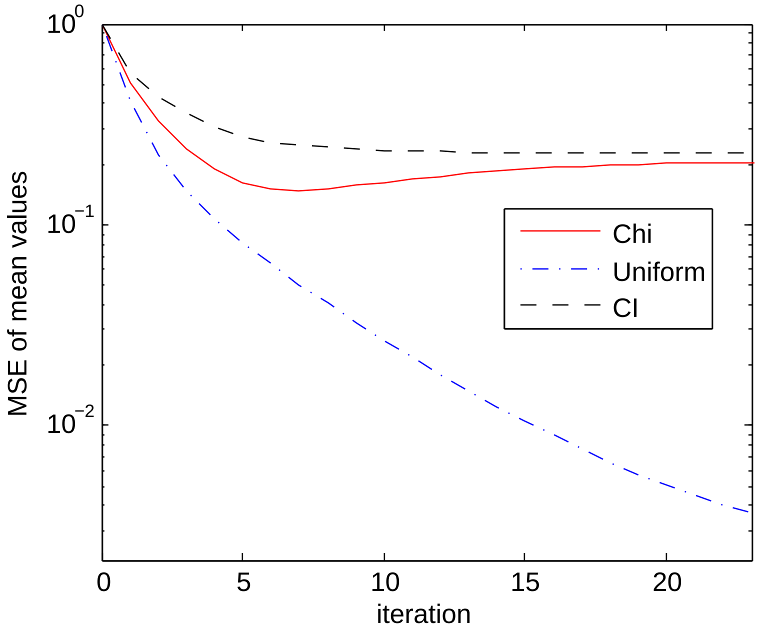

Figure 6.7: Comparison of the normalized MSE of the mean values.

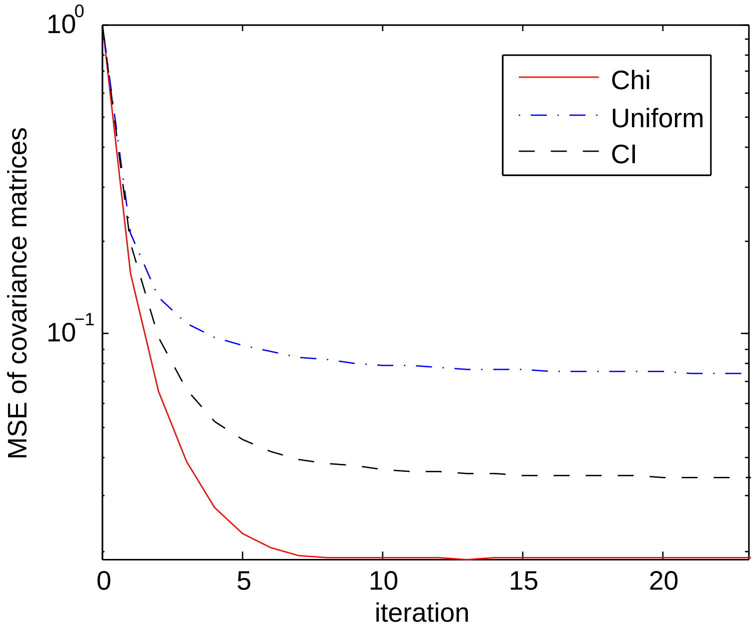

Figure 6.8: Comparison of the normalized MSE of the covariance matrices.

6.6 Conclusion

In this chapter, we proposed a new approach for belief consensus. Unlike traditional consensus, where nodes reach a consensus of point estimate, we considered the consensus of probability densities. Ideally the consensus algorithm should converge at the Bayesian posterior probability density given all the information available over the network. In the case where the nodes do not know the size of the network, this is not achievable. We proposed the use of the weighted product of the belief densities to approximate the Bayesian posterior. We confined ourselves in cases where the beliefs were Gaussian densities. We adopted the χ^2 information metric as the criterion for belief consensus. The criterion was used for choosing values of the weighting coefficients in the fusion. We proved that the optimization of the weighting coefficients was a convex problem under the χ^2 information metric. We studied the performance in the numerical experiments. It was shown that the proposed method outperforms others in the comparison of the χ^2 information.

Chapter 7

Conclusion

Distributed estimation and fusion is a popular topic in WSNs and has been applied in various areas including monitoring, surveillance and target tracking. The key challenge is to design algorithms that allow nodes in the network to cooperate with each other in an efficient way so that the information obtained by every single node can be spread all over the network. This becomes even more challenging when the measurements are correlated.

In this dissertation, we proposed several methods to handle the cases where correlation was present. In Chapter 2, we introduced the problem of fusing multiple correlated estimates with unknown correlation. In Chapter 3, the Bayesian approach was proposed in which we assumed the entire matrix was a Wishart random matrix. Since the values of the diagonal blocks were known, we could obtain the conditional distribution of those off-diagonal blocks in the covariance matrix. The conditional distribution was shown to be inverted matrix-variate t-distribution. Then Monte Carlo method was used to marginalize with respect to the off-diagonal blocks to obtain the MMSE estimate. In Chapter 4, we studied the same problem with slightly different assumptions. We assumed the covariance matrix had a special structure. Besides, instead of marginalizing with respect to the off-diagonal blocks, we sought the optimal values under two criteria by using convex optimization techniques.

We considered the problem of distributed estimation in the networks in Chapter 5. The correlation was assumed known and had the Markov property. We proposed an efficient algorithm that asymptotically achieves the same performance as the centralized method.

In Chapter 6, we considered the belief consensus problem. We assumed each node had an initial belief represented as a probability density. The objective was that the network reaches consensus at the density of the Bayesian posterior, i.e., the product of all the initial densities. However, with the assumption that the number of the nodes was unknown, it was not possible to achieve the Bayesian posterior through consensus. An approximation of the Bayesian posterior by the weighted product of the densities was proposed. We adopted the χ^2 information as the criterion to measure the distance between the Bayesian posterior and the weighted product. Besides, we proved that the optimization problem of minimizing the χ^2 information with respect to the weighting coefficients was convex.

There are many directions to continue this work. In Chapter 5, we assumed that the correlation of noises was known. Without this assumption, the nodes can still perform the estimation by simply ignoring the correlation. It would be interesting to quantify how much information is lost in this process of learning. Alternatively, the nodes can jointly estimate the parameter and the correlation. How well can the joint estimation improve the aggregation of information is worth of investigation. In Chapter 6, the χ^2 information metric was used. There are other criteria to measure the distance between two densities, for example, the Kullback-Leibler divergence. It would be interesting to quantitatively compare the performance of the nodes of these two criteria. Finally, notice that we used the weighted product, which was also called the general fusion, for the belief consensus. This was in fact a compromise we made to avoid the shrinkage of the densities in the consensus. Consequently we could not achieve the Bayesian posterior but had to approximate it. Whether the Bayesian posterior can be achieved in a network where the nodes know nothing about the topology is still an open question. The pursuing of this objective definitely requires the learning of the topology through belief exchange. This is perhaps the most interesting and challenging direction to extend this work.

Bibliography

[1] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer New York, 2006.

[2] J. Hu, L. Xie, and C. Zhang, "Diffusion Kalman filtering based on covariance intersection," *IEEE Transactions on Signal Processing*, vol. 60, pp. 891–902, Feb. 2012.

[3] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2069–2084, 2010.

[4] R. Olfati-Saber, "Distributed Kalman filtering for sensor networks," in *46th IEEE Conference on Decision and Control, 2007*, pp. 5492–5498, IEEE, 2007.

[5] R. Olfati-Saber, "Distributed Kalman filter with embedded consensus filters," in *44th IEEE Conference on Decision and Control, 2005*, pp. 8179–8184, IEEE, 2005.

[6] D. L. Hall and J. Llinas, *Handbook of Multisensor Data Fusion.* CRC Press, 2001.

[7] Y. Bar-Shalom and L. Campo, “The effect of the common process noise on the two-sensor fused-track covariance,” *IEEE Transactions on Aerospace and Electronic Systems*, no. 6, pp. 803–805, 1986.

[8] Y. Bar-Shalom, “On the track-to-track correlation problem,” *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 571–572, 1981.

[9] X. R. Li, Y. Zhu, J. Wang, and C. Han, “Optimal linear estimation fusion. I. Unified fusion rules,” *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2192–2208, 2003.

[10] Y. M. Zhu and X. R. Li, “Best linear unbiased estimation fusion,” in *Proceedings of the 1999 International Conference on Information Fusion*, pp. 1054–1061, 1999.

[11] X. R. Li and J. Wang, “Unified optimal linear estimation fusion. Part II: Discussions and examples,” in *Proceedings of the 2000 International Conference on Information Fusion*, 2000.

[12] X. R. Li and K. S. Zhang, “Optimal linear estimation fusion. Part IV: Optimality and efficiency of distributed fusion,” in *Proceedings of the 2001 International Conference on Information Fusion*, 2001.

[13] C. Y. Chong, S. Mori, and K. C. Chang, “Information fusion in distributed sensor networks,” in *American Control Conference, 1985*, pp. 830–835, IEEE, 1985.

[14] K. C. Chang, R. K. Saha, and Y. Bar-Shalom, “On optimal track-to-track fusion,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 4, pp. 1271–1276, 1997.

[15] O. E. Drummond, "Hybrid sensor fusion algorithm architecture and tracklets," in *Proceedings of SPIE*, vol. 3163, p. 485, 1997.

[16] A. Willsky, M. Bello, D. Castanon, B. Levy, and G. Verghese, "Combining and updating of local estimates and regional maps along sets of one-dimensional tracks," *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 799–813, 1982.

[17] M. D. Miller, O. E. Drummond, and A. J. Perrella, "Tracklets and covariance truncation options for theater missile tracking," in *Proceedings of the 1998 International Conference on Multisource-Multisensor Data Fusion (FUSION)*, 1998.

[18] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *American Control Conference, 1997*, vol. 4, pp. 2369–2373, IEEE, 1997.

[19] W. Niehsen, "Information fusion based on fast covariance intersection filtering," in *Proceedings of the Fifth International Conference on Information Fusion, 2002*, vol. 2, pp. 901–904, IEEE, 2002.

[20] D. Franken and A. Hupper, "Improved fast covariance intersection for distributed data fusion," in *8th International Conference on Information Fusion*, vol. 1, p. 7, IEEE, 2005.

[21] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics, 2003.

[22] A. T. James, "Distributions of matrix variates and latent roots derived from normal samples," *The Annals of Mathematical Statistics*, pp. 475–501, 1964.

[23] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*, vol. 104. Chapman & Hall/CRC, 2000.

[24] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[25] E. T. Jaynes, *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[26] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill Science/Engineering/Math, 2001.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[28] Y. Nesterov and A. S. Nemirovskii, *Interior-point polynomial algorithms in convex programming*, vol. 13. SIAM, 1994.

[29] S. Y. Park and A. K. Bera, "Maximum entropy autoregressive conditional heteroskedasticity model," *Journal of Econometrics*, vol. 150, no. 2, pp. 219–230, 2009.

[30] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, 2008.

[31] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4795–4810, 2010.

[32] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, “Diffusion recursive least-squares for distributed estimation over adaptive networks,” *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1865–1877, 2008.

[33] S. Kar and J. M. F. Moura, “Convergence rate analysis of distributed gossip linear parameter estimation: Fundamental limits and tradeoffs,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 674–690, 2011.

[34] S. Kar, J. M. F. Moura, and H. V. Poor, “Distributed linear parameter estimation: Asymptotically efficient adaptive strategies,” *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2200–2229, 2013.

[35] S. Tu and A. H. Sayed, “Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks,” *IEEE Transactions on Signal Processing*, vol. 60, pp. 6217–6234, Dec. 2012.

[36] A. H. Sayed, “Diffusion adaptation over networks,” *arXiv preprint arXiv:1205.4220*, 2012.

[37] I. D. Schizas, G. B. Giannakis, S. I. Roumeliotis, and A. Ribeiro, “Consensus in ad hoc WSNs with noisy links. Part II: Distributed estimation and smoothing of random signals,” *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1650–1666, 2008.

[38] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, “Consensus in ad hoc WSNs with noisy links. Part I: Distributed estimation of deterministic signals,” *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[39] S. Kar, J. M. F. Moura, and K. Ramanan, “Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication,” *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

[40] O. Hlinka, F. Hlawatsch, and P. M. Djurić, “Distributed particle filtering in agent networks: A survey, classification, and comparison,” *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 61–81, 2013.

[41] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory.* Prentice Hall, 1993.

[42] A. Wiesel and A. O. Hero, “Distributed covariance estimation in Gaussian graphical models,” *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 211–220, 2012.

[43] P. Braca, S. Marano, V. Matta, and P. Willett, “Asymptotic optimality of running consensus in testing binary hypotheses,” *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 814–825, 2010.

[44] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura, “Distributed detection via gaussian running consensus: Large deviations asymptotic analysis,” *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4381–4396, 2011.

[45] P. Braca, S. Marano, and V. Matta, “Enforcing consensus while monitoring the environment in wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3375–3380, 2008.

[46] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Asymptotic performance of adaptive distributed detection over networks," *arXiv preprint arXiv:1401.5742*, 2014.

[47] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4306–4320, 2012.

[48] D. Bajovic, D. Jakovetic, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-Gaussian observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5987–6002, 2012.

[49] Z. Weng and P. M. Djurić, "Distributed estimation in the presence of correlated noises," in *Proceedings of the 21th European Signal Processing Conference (EUSIPCO)*, pp. 2352–2356, September 2013.

[50] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[51] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[52] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[53] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[54] S. Sardellitti, M. Giona, and S. Barbarossa, “Fast distributed average consensus algorithms based on advection-diffusion processes,” *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 826–842, 2010.

[55] S. Boyd, P. Diaconis, and L. Xiao, “Fastest mixing Markov chain on a graph,” *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.

[56] J. Delvenne, R. Carli, and S. Zampieri, “Optimal strategies in the average consensus problem,” *Systems & Control Letters*, vol. 58, no. 10, pp. 759–765, 2009.

[57] S. Ghosh and J. Lee, “Optimal distributed consensus on unknown undirected graphs,” in *IEEE 51st Annual Conference on Decision and Control (CDC), 2012*, pp. 2244–2249, IEEE, 2012.

[58] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, “Broadcast gossip algorithms for consensus,” *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2748–2761, 2009.

[59] D. Jakovetic, J. Xavier, and J. M. F. Moura, “Weight optimization for consensus algorithms with correlated switching topology,” *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3788–3801, 2010.

[60] A. Bertrand and M. Moonen, “Consensus-based distributed total least squares estimation in ad hoc wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2320–2330, 2011.

[61] G. Battistelli and L. Chisci, “Kullback-Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability,” *Automatica*, vol. 50, no. 3, pp. 707–718, 2014.

[62] M. B. Hurley, "An information theoretic justification for Covariance Intersection and its generalization," in *Proceedings of the Fifth International Conference on Information Fusion, 2002*, vol. 1, pp. 505–511, IEEE, 2002.

[63] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook.* Technical University of Denmark, 2006.

[64] M. Vemula, M. F. Bugallo, and P. M. Djurić, "Performance comparison of Gaussian-based filters using information measures," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1020–1023, 2007.

[65] E. Carlen, "Trace inequalities and quantum entropy: an introductory course," *Entropy and the quantum: Arizona School of Analysis with Applications*, March 2009.

[66] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, 2012.

Appendices

The proof of the main theorem:

In this section, we prove that the distributed estimator, (5.24), is efficient. Denote by $g\left(\mathbf{A},\mathbf{B}\right)$ the matrix-variate function

$$g\left(\mathbf{A},\mathbf{B}\right) = \text{tr}\left[\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{A}^{\top}\right)^{-1}\right]. \tag{7.1}$$

We define the following notation:

$$\mathbf{F}_t = \sum_{s=1}^{t}\mathbf{H}_s^{\top}\mathbf{K}\mathbf{Q}_{(i)}^{t-s}\mathbf{H}_s \tag{7.2}$$

$$\mathbf{G}_t = \sum_{s=1}^{t}\mathbf{H}_s^{\top}\mathbf{K}\mathbf{Q}_{(i)}^{t-s}\mathbf{K}^{-1}\mathbf{Q}_{(i)}^{t-s}\mathbf{K}\mathbf{H}_s \tag{7.3}$$

$$\mathbf{X}_{t,s} = \mathbf{Q}_{(i)}^{t-s} - \frac{1}{N}\mathbf{I}. \tag{7.4}$$

Then the covariance of the distributed estimator can be expressed as

$$\begin{aligned}\text{Cov}\left[\tilde{\boldsymbol{\theta}}_{i,t}\right] &= \mathbb{E}\left[\left(\tilde{\boldsymbol{\theta}}_{i,t}-\boldsymbol{\theta}\right)\left(\tilde{\boldsymbol{\theta}}_{i,t}-\boldsymbol{\theta}\right)^{\top}\right] \\ &= \mathbf{F}_t^{-1}\mathbf{G}_t\left(\mathbf{F}_t^{\top}\right)^{-1}. \end{aligned} \tag{7.5}$$

The covariance of the centralized estimate (5.20) is

$$\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}_t\right] = \left(\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \mathbf{H}_s\right)^{-1}. \tag{7.6}$$

Formally, we need to prove

$$\lim_{t\to\infty} \frac{\mathrm{tr}\left[\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}_{i,t}\right]\right]}{\mathrm{tr}\left[\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}_t\right]\right]} = 1. \tag{7.7}$$

The first step of the proof is to decompose both $\mathbf{G}_t$ and $\mathbf{F}_t$ into two parts using the identity $\mathbf{Q}_{(i)}^{t-s} = \frac{1}{N}\mathbf{I} + \mathbf{Q}_{(i)}^{t-s} - \frac{1}{N}\mathbf{I}$. We have

$$\begin{aligned}\mathbf{G}_t =& \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K} \left(\frac{1}{N}\mathbf{I} + \mathbf{X}_{t,s}\right) \mathbf{K}^{-1} \left(\frac{1}{N}\mathbf{I} + \mathbf{X}_{t,s}\right) \mathbf{K}\mathbf{H}_s & (7.8)\\ =& \frac{1}{N^2}\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K}\mathbf{H}_s + \frac{1}{N}\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s & \\ &+ \frac{1}{N}\sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{H}_s + \sum_{s=1}^{t} \mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{K}^{-1}\mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s & (7.9)\end{aligned}$$

and

$$\mathbf{F}_t = \frac{1}{N}\sum_{q=1}^{t} \mathbf{H}_q^\top \mathbf{K}\mathbf{H}_q + \sum_{q=1}^{t} \mathbf{H}_q^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{H}_q. \tag{7.10}$$

Then (7.7) becomes

$$\begin{aligned}\mathrm{tr}\left[\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}_{i,t}\right]\right] \Big/ \mathrm{tr}\left[\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}_t\right]\right] =& \frac{g\left(\mathbf{A}_1 + \mathbf{A}_2, \frac{1}{N}\mathbf{A}_1 + \mathbf{B}_1 + \mathbf{B}_2\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]} & (7.11)\\ =& \frac{g\left(\mathbf{A}_1 + \mathbf{A}_2, \frac{1}{N}\mathbf{A}_1\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]} & (7.12)\\ &+ \frac{g\left(\mathbf{A}_1 + \mathbf{A}_2, \mathbf{B}_1 + \mathbf{B}_2\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]}, & (7.13)\end{aligned}$$

where

$$\mathbf{A}_1 = \frac{1}{N}\sum_{s=1}^{t}\mathbf{H}_s^\top \mathbf{K}\mathbf{H}_s, \tag{7.14}$$

$$\mathbf{A}_2 = \sum_{s=1}^{t}\mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{H}_s, \tag{7.15}$$

$$\mathbf{B}_1 = \frac{1}{N}\sum_{s=1}^{t}\mathbf{H}_s^\top \mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s + \frac{1}{N}\sum_{s=1}^{t}\mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{H}_s, \tag{7.16}$$

$$\mathbf{B}_2 = \sum_{s=1}^{t}\mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{K}^{-1}\mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s. \tag{7.17}$$

We shall prove (7.12) approaches 1 and (7.13) approaches 0 as t grows. The main principle we use is the squeeze theorem in calculus. Basically, we look for the upper and lower bounds for both terms and prove they have the limit we want. Before we start, we prepare three lemmas for use later:

Lemma 4 $\sum_{s=1}^{t}|\mathbf{X}_{t,s}|$ *is bounded as t approaches infinity.*

Proof: Since $\mathbf{Q}$ is symmetric, it can be decomposed as $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ where $\mathbf{U}$ is an orthonormal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues being the diagonal entries. Therefore $\mathbf{Q}^t = \mathbf{U}\mathbf{\Lambda}^t\mathbf{U}^\top$. Denote by $u_{i,j}$ the (i,j)th entry of $\mathbf{U}$, and by λ_l the lth smallest eigenvalue of $\mathbf{Q}$. Then we have $Q_{i,j}^{t-s} = \sum_{l=1}^{N} u_{j,l}u_{i,l}\lambda_l^{t-s}$. Because $\mathbf{Q}$ is an irreducible and aperiodic doubly stochastic matrix, we know that one of the eigenvalues is 1 with the corresponding eigenvector being $(1/\sqrt{N})\mathbf{1}_N$, and the rest of the eigenvalues being strictly less than 1 in magnitude. Let $\lambda_N = 1$; we have

$u_{i,N}u_{j,N}\lambda_N^{t-s} = \frac{1}{N}$ and

$$\begin{aligned}
\sum_{s=1}^{t}\left|Q_{i,j}^{t-s} - \frac{1}{N}\right| &= \sum_{s=1}^{t}\left|\sum_{l=1}^{N-1} u_{j,l}u_{i,l}\lambda_l^{t-s}\right| & (7.18)\\
&\le \sum_{s=1}^{t}\sum_{l=1}^{N-1}\left|u_{j,l}u_{i,l}\right|\left|\lambda_l^{t-s}\right| & (7.19)\\
&= \sum_{l=1}^{N-1}\left|u_{j,l}u_{i,l}\right|\frac{1-\left|\lambda_l\right|^t}{1-\left|\lambda_l\right|}. & (7.20)
\end{aligned}$$

As the absolute values of $\{\lambda\}_{l=1}^{N-1}$ are all strictly smaller than 1, (7.20) is bounded. □

Lemma 5 $\mathbf{A}_2$ *is bounded as* t *grows.*

Proof: Let $h_{i,j}(s)$ be the (i,j)th entry of $\mathbf{H}_s$, $k_{i,j}$, that of $\mathbf{K}$, $a_{i,j}$, that of $\mathbf{A}_2$. Let $x_i(t,s)$ be the ith diagonal entry of $\mathbf{X}_{t,s}$. Then we have

$$\begin{aligned}
a_{i,j} &= \sum_{s=1}^{t} x_1(t,s)\, h_{1,j}(s) \sum_{l=1}^{MN} h_{l,i}(s)\, k_{l,1} + \cdots\\
&\quad + \sum_{s=1}^{t} x_{MN}(t,s)\, h_{MN,j}(s) \sum_{l=1}^{MN} h_{l,i}(s)\, k_{l,MN}. \qquad (7.21)
\end{aligned}$$

We show that each term in the above expression is bounded. Consider the first term: Because H_s is bounded, we can always find a constant c such that $-c \le h_{1,j}(s)\sum_{l=1}^{MN} h_{l,i}(s)\, k_{l,1} \le c$ for all s. Therefore, the first term (say tm_1) can be upper-bounded as

$$tm_1 \le \sum_{s=1}^{t}\left|x_1(t,s)\right|\left|h_{1,j}(s)\sum_{l=1}^{MN} h_{l,i}(s)\, k_{l,1}\right| \le c\sum_{s=1}^{t}\left|x_1(t,s)\right| \qquad (7.22)$$

and lower-bounded by $-c\sum_{s=1}^{t}\left|x_1(t,s)\right|$. According to Lemma 4, both the upper and lower bounds are bounded. Therefore, $a_{i,j}$ and A_2 are bounded.

□

Lemma 6 *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous, if $f(x)$ is monotone increasing (decreasing), so is $tr[f(\mathbf{A})]$ on $\mathbb{S}$ [65]. Here increasing means that if $\mathbf{A}_t \succ \mathbf{A}_s$, then $tr[f(\mathbf{A}_t)] > tr[f(\mathbf{A}_s)]$ ($tr[f(\mathbf{A}_t)] < tr[f(\mathbf{A}_s)]$).*

In the next two subsections, we prove (7.12) approaches 1 and (7.13) approaches 0 as t grows.

The limit of the first term, (7.12)

Define

$$f_1(x) = \frac{g\left(x\mathbf{I} + \mathbf{A}_2, \frac{1}{N}x\mathbf{I}\right)}{\frac{1}{N}\mathrm{tr}\left[(x\mathbf{I})^{-1}\right]}. \tag{7.23}$$

We first show $f_1(x)$ is monotone. The function $f_1(x)$ can be simplified to

$$f_1(x) = \frac{1}{L}\mathrm{tr}\left[\left(\mathbf{I} + \frac{\mathbf{A}_2}{x}\right)^{-1}\left(\mathbf{I} + \frac{\mathbf{A}_2^\top}{x}\right)^{-1}\right]. \tag{7.24}$$

According to the definition of matrix inversion, each entry in $\left(\mathbf{I} + \frac{\mathbf{A}_2}{x}\right)^{-1}$is a rational function of x. Also addition, multiplication and division of rational functions are still rational functions. Therefore $f_1(x)$ is a rational function. So is the derivative of $f_1(x)$, $f_1'(x)$. Since the orders of the polynomials are bounded, there exist a finite number of zeros and poles. Thus, there must exist a constant x_L such that $f_1'(x)$ becomes positive or negative when $x > x_L$. This says $f_1(x)$ becomes monotonic when $x > x_L$. Whether it is decreasing or increasing depends on the value of $\mathbf{A}_2$. So is the constant x_L. Without loss of generality, we assume $f_1(x)$ is increasing for $x > x_L$. We define $\mathbf{H}_d$ and $\mathbf{H}_D$ to be full-rank matrices such that

$$\mathbf{H}_d^\top \mathbf{K}\mathbf{H}_d \preceq \mathbf{H}_s^\top \mathbf{K}\mathbf{H}_s \preceq \mathbf{H}_D^\top \mathbf{K}\mathbf{H}_D \tag{7.25}$$

for all s. Then by Lemma 6 we have

$$\begin{aligned}
&\frac{g\left(\frac{t}{N}\mathbf{H}_d^\top \mathbf{K}\mathbf{H}_d + \mathbf{A}_2, \frac{t}{N^2}\mathbf{H}_d^\top \mathbf{K}\mathbf{H}_d\right)}{\frac{1}{N}\mathrm{tr}\left[\left(\frac{t}{N}\mathbf{H}_d^\top \mathbf{K}\mathbf{H}_d\right)^{-1}\right]} && (7.26)\\
\leq&\frac{g\left(\mathbf{A}_1 + \mathbf{A}_2, \frac{1}{N}\mathbf{A}_1\right)}{\frac{1}{N}\mathrm{tr}\left[\left(\mathbf{A}_1\right)^{-1}\right]} && (7.27)\\
\leq&\frac{g\left(\frac{t}{N}\mathbf{H}_D^\top \mathbf{K}\mathbf{H}_D + \mathbf{A}_2, \frac{t}{N^2}\mathbf{H}_D^\top \mathbf{K}\mathbf{H}_D\right)}{\frac{1}{N}\mathrm{tr}\left[\left(\frac{t}{N}\mathbf{H}_D^\top \mathbf{K}\mathbf{H}_D\right)^{-1}\right]} && (7.28)
\end{aligned}$$

for sufficiently large t. Because the limits of (7.26) and (7.28) are 1, according to the squeeze theorem, (7.12) also goes to 1 as t approaches infinity. The same result follows if $f_1(x)$ is decreasing.

The limit of the second term, (7.13)

In this subsection, we still use the squeeze theorem and follow a similar strategy. We need to show that the limit of (7.13) is zero. First, we replace $\mathbf{B}_1 + \mathbf{B}_2$ with $x\mathbf{I}$ and write the denominator as a function of x:

$$f_2(x) = g\left(\mathbf{A}_1 + \mathbf{A}_2, x\mathbf{I}\right). \tag{7.29}$$

Obviously $f_2(\mathbf{A})$ is increasing on $\mathbb{S}$. Define the function

$$\begin{aligned}
\mathbf{Z}_{s,t} =&\frac{1}{N}\mathbf{H}_s^\top \mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s + \frac{1}{N}\mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{H}_s\\
&+ \mathbf{H}_s^\top \mathbf{K}\mathbf{X}_{t,s}\mathbf{K}^{-1}\mathbf{X}_{t,s}\mathbf{K}\mathbf{H}_s. && (7.30)
\end{aligned}$$

Then we let $\mathbf{H}_B$ and $\mathbf{H}_b$ be matrices such that

$$\mathbf{Z}_{b,t} \preceq \mathbf{Z}_{s,t} \preceq \mathbf{Z}_{B,t} \tag{7.31}$$

and

$$\mathbf{Z}_{b,t} \preceq \mathbf{O} \preceq \mathbf{Z}_{B,t} \tag{7.32}$$

for all s and t. Note that the last expression implies that $\mathbf{Z}_{b,t}$ is negative semidefinite and $\mathbf{Z}_{B,t}$ is positive semidefinite. By (5.13) and the definition of $\mathbf{X}_{t,s}$ in (7.4), we can see that $\mathbf{X}_{t,s}$ is a diagonal matrix with both positive and negative entries. This makes it possible to find such $\mathbf{H}_B$ and $\mathbf{H}_b$ that satisfy (7.31) and (7.32). Due to the monotone property of f_2, we have

$$\begin{aligned} &\frac{g\left(\mathbf{A}_1+\mathbf{A}_2, \sum_{s=1}^{t} \mathbf{Z}_{b,t}\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]} && (7.33)\\ \leq &\frac{g\left(\mathbf{A}_1+\mathbf{A}_2, \mathbf{B}_1+\mathbf{B}_2\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]} && (7.34)\\ \leq &\frac{g\left(\mathbf{A}_1+\mathbf{A}_2, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]}. && (7.35)\end{aligned}$$

Next we prove that the limits of both the left side (7.33) and the right side (7.35) are zero as t approaches infinity. To find the limit of (7.35), we define the function $f_3(x)$ as

$$f_3(x) = g\left(x, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right). \tag{7.36}$$

Note that it is decreasing on $\mathbb{R}^{+}$. According to Lemma 6, $f_3(\mathbf{A}) = g\left(\mathbf{A}, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right)$ is decreasing on $\mathbb{S}^{+}$. Let $\mathbf{H}_a$ be a matrix such that

$$\mathbf{O} \preceq \mathbf{H}_a^\top \mathbf{K} \mathbf{H}_a \preceq \mathbf{H}_s^\top \mathbf{K} \mathbf{H}_s \tag{7.37}$$

for all s. Also let $\mathbf{H}_A$ be a matrix such that

$$\mathbf{H}_s^\top \mathbf{K} \mathbf{H}_s \preceq \mathbf{H}_A^\top \mathbf{K} \mathbf{H}_A \tag{7.38}$$

for all s. By enlarging the numerator and reducing the denominator of (7.35), we have

$$\frac{g\left(\mathbf{A}_1 + \mathbf{A}_2, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right)}{\frac{1}{N}\mathrm{tr}\left[\mathbf{A}_1^{-1}\right]} \tag{7.39}$$

$$\leq \frac{g\left(\frac{t}{N}\mathbf{H}_a^\top \mathbf{K} \mathbf{H}_a + \mathbf{A}_2, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right)}{\frac{1}{N}\mathrm{tr}\left[\left(\frac{t}{N}\mathbf{H}_A^\top \mathbf{K} \mathbf{H}_A\right)^{-1}\right]} \tag{7.40}$$

$$= \frac{\frac{1}{t^2} g\left(\frac{1}{N}\mathbf{H}_a^\top \mathbf{K} \mathbf{H}_a + \mathbf{A}_2, \sum_{s=1}^{t} \mathbf{Z}_{B,t}\right)}{\frac{1}{t}\mathrm{tr}\left[\left(\mathbf{H}_A^\top \mathbf{K} \mathbf{H}_A\right)^{-1}\right]}. \tag{7.41}$$

Since $\sum_{s=1}^{t} \mathbf{Z}_{B,t}$ and $\mathbf{A}_2$ are bounded, according to Lemma 6 we can see that the limit of (7.35) is zero. Similarly we can show that (7.33) is lower-bounded by an expression the limit of which is also zero. This completes the proof of the main theorem.

Derivation of (6.20)

The multivariate normal distribution of a N-dimensional random vector $\mathbf{x}$ with mean $\mathbf{m}$ and precision matrix $\mathbf{K}$ is defined as

$$\begin{aligned}\mathcal{N}\left(\mathbf{x}|\mathbf{m},\mathbf{K}^{-1}\right) =&\frac{1}{\sqrt{(2\pi)^{N}\left|\mathbf{K}^{-1}\right|}}\\ &\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}\right)^{\top}\mathbf{K}\left(\mathbf{x}-\mathbf{m}\right)\right).\end{aligned} \tag{7.42}$$

Let $p\left(\mathbf{x}\right)$ be $\mathcal{N}\left(\mathbf{x}|\mathbf{m}_c,\mathbf{K}_c^{-1}\right)$ and $q\left(\mathbf{x}\right)$ be $\mathcal{N}\left(\mathbf{x}|\mathbf{m}_d,\mathbf{K}_d^{-1}\right)$. Then we have

$$\frac{p^2\left(\mathbf{x}\right)}{q\left(\mathbf{x}\right)} =\frac{\sqrt{(2\pi)^{K}\left|\mathbf{K}_d^{-1}\right|}}{(2\pi)^{K}\left|\mathbf{K}_c^{-1}\right|}\exp\left(Z_1\right) \tag{7.43}$$

where

$$\begin{aligned}Z_1 =&-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}_c\right)^{\top}2\mathbf{K}_c\left(\mathbf{x}-\mathbf{m}_c\right)\\ &+\frac{1}{2}\left(\mathbf{x}-\mathbf{m}_d\right)^{\top}\mathbf{K}_d\left(\mathbf{x}-\mathbf{m}_d\right).\end{aligned} \tag{7.44}$$

We further rearrange the terms in Z_1 as follows:

$$\begin{aligned}Z_1 =&-\frac{1}{2}\left(\mathbf{x}^{\top}2\mathbf{K}_c\mathbf{x}-\mathbf{x}^{\top}\mathbf{K}_d\mathbf{x}\right)\\ &+\frac{1}{2}\left(2\mathbf{x}^{\top}2\mathbf{K}_c\mathbf{m}_c-2\mathbf{x}^{\top}\mathbf{K}_d\mathbf{m}_d\right)\\ &-\frac{1}{2}\left(\mathbf{m}_c^{\top}2\mathbf{K}_c\mathbf{m}_c-\mathbf{m}_d^{\top}\mathbf{K}_d\mathbf{m}_d\right) \qquad (7.45)\\ =&Z_2+Z_3 \qquad (7.46)\end{aligned}$$

where

$$\begin{aligned}
Z_2 =& -\frac{1}{2}\left(\mathbf{x}-\left(2\mathbf{K}_c-\mathbf{K}_d\right)^{-1}\left(2\mathbf{K}_c\mathbf{m}_c-\mathbf{K}_d\mathbf{m}_d\right)\right)^{\top} \\
& \times \left(2\mathbf{K}_c-\mathbf{K}_d\right) \\
& \times \left(\mathbf{x}-\left(2\mathbf{K}_c-\mathbf{K}_d\right)^{-1}\left(2\mathbf{K}_c\mathbf{m}_c-\mathbf{K}_d\mathbf{m}_d\right)\right) \qquad (7.47)
\end{aligned}$$

and

$$\begin{aligned}
Z_3 =& \frac{1}{2}\left(2\mathbf{K}_c\mathbf{m}_c-\mathbf{K}_d\mathbf{m}_d\right)^{\top} \\
& \times \left(2\mathbf{K}_c-\mathbf{K}_d\right)^{-1} \\
& \times \left(2\mathbf{K}_c\mathbf{m}_c-\mathbf{K}_d\mathbf{m}_d\right) \\
& -\frac{1}{2}\left(\mathbf{m}_c^{\top}2\mathbf{K}_c\mathbf{m}_c-\mathbf{m}_d^{\top}\mathbf{K}_d\mathbf{m}_d\right) \qquad (7.48) \\
=& -\frac{1}{2}\left(\mathbf{m}_c-\mathbf{m}_d\right)^{\top}\left(\frac{1}{2}\mathbf{K}_c^{-1}-\mathbf{K}_d^{-1}\right)^{-1}\left(\mathbf{m}_c-\mathbf{m}_d\right). \qquad (7.49)
\end{aligned}$$

According to a property of multivariate normal distribution, as long as $2\mathbf{K}_c - \mathbf{K}_d$ is positive definite, we have

$$\int \exp\left(Z_2\right)d\mathbf{x} = \sqrt{\left(2\pi\right)^{K}} \cdot \sqrt{\left|\left(2\mathbf{K}_c-\mathbf{K}_d\right)^{-1}\right|}. \qquad (7.50)$$

Therefore, $\int \frac{p^2(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ becomes

$$\begin{aligned}
&\int \frac{p^2\left(\mathbf{x}\right)}{q\left(\mathbf{x}\right)} d\mathbf{x} \\
=&\frac{\sqrt{\left(2\pi\right)^K \left|\mathbf{K}_d^{-1}\right|}}{\left(2\pi\right)^K \left|\mathbf{K}_c^{-1}\right|} \cdot \sqrt{\left(2\pi\right)^K} \cdot \sqrt{\left|\left(2\mathbf{K}_c - \mathbf{K}_d\right)^{-1}\right|} \\
&\cdot \exp\left(-\frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \left(\frac{1}{2}\mathbf{K}_c^{-1} - \mathbf{K}_d^{-1}\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right)\right) \qquad (7.51) \\
=&\frac{\sqrt{\left|\mathbf{K}_d^{-1}\right|}}{\left|\mathbf{K}_c^{-1}\right|} \cdot \sqrt{\left|\left(2\mathbf{K}_c - \mathbf{K}_d\right)^{-1}\right|} \\
&\cdot \exp\left(-\frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \left(\frac{1}{2}\mathbf{K}_c^{-1} - \mathbf{K}_d^{-1}\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right)\right). \qquad (7.52)
\end{aligned}$$

Proof of the convexity of (6.23).

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if the domain of f is a convex set and for all $\mathbf{x}$, $\mathbf{y}$ in the domain, and θ with $0 \le \theta \le 1$, we have

$$f\left(\theta\mathbf{x} + \left(1-\theta\right)\mathbf{y}\right) \le \theta f\left(\mathbf{x}\right) + \left(1-\theta\right) f\left(\mathbf{y}\right). \qquad (7.53)$$

Also a function is convex if and only if it is convex when restricted to any line that intersects its domain. This property is useful for us to check the convexity of a function. In our case, the variables are w_i for $i \in \{1, \cdots, N\}$ and the domain is a simplex defined by

$$\begin{aligned}
\sum_{i=1}^{N} w_i =& 1 \qquad (7.54) \\
w_i \ge& 0. \qquad (7.55)
\end{aligned}$$

Those lines can be expressed as

$$x_1 = w_1 + tv_1 \tag{7.56}$$

$$\vdots$$

$$x_N = w_N + tv_N \tag{7.57}$$

where $\sum_{i=1}^{N} v_i = 0$ and all the w_is lie in the simplex defined in (7.54) and (7.55); t is a real number. To check the convexity of the function on this line, we examine the second derivative of the function with respect to t. Let

$$\mathbf{m}_w = \sum_i w_i \mathbf{K}_i \left(\mathbf{m}_c - \mathbf{m}_i\right) \tag{7.58}$$

$$\mathbf{m}_v = \sum_i v_i \mathbf{K}_i \left(\mathbf{m}_c - \mathbf{m}_i\right) \tag{7.59}$$

$$\mathbf{m}_{w+vt} = \sum_i \left(w_i + tv_i\right) \mathbf{K}_i \left(\mathbf{m}_c - \mathbf{m}_i\right) \tag{7.60}$$

$$\mathbf{K}_w = \sum_i w_i \mathbf{K}_i \tag{7.61}$$

$$\mathbf{K}_v = \sum_i v_i \mathbf{K}_i \tag{7.62}$$

$$\mathbf{K}_{w+vt} = \sum_i \left(w_i + tv_i\right) \mathbf{K}_i. \tag{7.63}$$

and define $f\left(t\right)$ and $h\left(t\right)$ as

$$f\left(t\right) = \frac{1}{\sqrt{\mathbf{K}_{w+tv}}} \cdot \sqrt{\left|\left(2\mathbf{K}_c - \mathbf{K}_{w+tv}\right)^{-1}\right|} \tag{7.64}$$

$$h\left(t\right) = \exp\left(-\frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_{w+tv}\right)^{\top} \left(\frac{1}{2}\mathbf{K}_c^{-1} - \mathbf{K}_{w+tv}^{-1}\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_{w+tv}\right)\right). \tag{7.65}$$

We replace w_i with $w_i + tv_i$, then (6.23) becomes

$$\chi^2 \left(\mathcal{N}_d || \mathcal{N}_c\right) = |\mathbf{K}_c| f\left(t\right) \exp h\left(t\right) - 1. \tag{7.66}$$

Since $|\mathbf{K}_c|$ is a constant factor we omit it in the following derivation for the sake of simplicity. We show that the second order derivative of $f\left(t\right) \exp h\left(t\right)$ is nonnegative. We denote by f' and f'' the first order and the second order derivatives of $f\left(t\right)$, respectively. The first order derivative of $f\left(t\right) \exp h\left(t\right)$ can be derived as

$$\frac{\partial f\left(t\right) \exp h\left(t\right)}{\partial t} = f' \exp h + f h' \exp h. \tag{7.67}$$

and then the second derivative can be derived as

$$\begin{aligned}
\frac{\partial^2 \left(f\left(t\right) \exp h\left(t\right)\right)}{\partial t^2} =& f'' \exp h + f'h' \exp h \\
& + f'h' \exp h + f\left(h'\right)^2 \exp h + f h'' \exp h && (7.68)\\
=& \left(f'' + f'h' + f'h' + f\left(h'\right)^2 + f h''\right) \exp h && (7.69)\\
=& \left(f'' + 2f'h' + f\left(h'\right)^2 + f h''\right) \exp h && (7.70)\\
=& \left(f'' + f\left(h' + \frac{f'}{f}\right)^2 - \frac{\left(f'\right)^2}{f} + f h''\right) \exp h. && (7.71)
\end{aligned}$$

We are going to prove that (7.71) is nonnegative when $t = 0$. We note that the second term in the parentheses is in square form and therefore nonnegative. The second factor $\exp h\left(t\right)$ is always positive. In order to prove (7.71) is nonnegative, it suffices to show that $f'' - \left(f'\right)^2 / f$ and h'' are nonnegative.

The proof that $f'' - (f')^2/f$ is nonnegative

First, we define

$$\mathbf{A} = \mathbf{C}_d^{-1}\left(2\mathbf{C}_c^{-1} - \mathbf{C}_d^{-1}\right). \tag{7.72}$$

$$= \mathbf{K}_d\left(2\mathbf{K}_c - \mathbf{K}_d\right). \tag{7.73}$$

According to (7.72), we have

$$f(t) = \frac{1}{\sqrt{|\mathbf{A}|}}. \tag{7.74}$$

We first obtain the first and the second derivatives of $\mathbf{A}$:

$$\frac{\partial \mathbf{A}}{\partial t} = \sum_i v_i \mathbf{K}_i \left(2\mathbf{K}_c - \sum_i (w_i + tv_i)\mathbf{K}_i\right) - \sum_i (w_i + tv_i)\mathbf{K}_i \left(\sum_i v_i \mathbf{K}_i\right) \tag{7.75}$$

$$\frac{\partial^2 \mathbf{A}}{(\partial t)^2} = \sum_i v_i \mathbf{K}_i \left(-\sum_i v_i \mathbf{K}_i\right) - \sum_i v_i \mathbf{K}_i \left(\sum_i v_i \mathbf{K}_i\right) \tag{7.76}$$

$$= -2\left(\sum_i v_i \mathbf{K}_i\right)^2 \tag{7.77}$$

$$= -2\mathbf{K}_v^2. \tag{7.78}$$

Set $t = 0$, $\dfrac{\partial \mathbf{A}}{\partial t}$ becomes

$$\begin{aligned}\frac{\partial \mathbf{A}}{\partial t} =&\sum_i v_i \mathbf{K}_i \left(2\mathbf{K}_c - \sum_i w_i \mathbf{K}_i\right) \\ &- \sum_i w_i \mathbf{K}_i \left(\sum_i v_i \mathbf{K}_i\right) && (7.79) \\ =&\mathbf{K}_v \left(2\mathbf{K}_c - \mathbf{K}_d\right) - \mathbf{K}_d \mathbf{K}_v. && (7.80)\end{aligned}$$

The first order and the second order derivatives of f become

$$\begin{aligned}\frac{\partial f}{\partial t} =& -\frac{1}{2} |\mathbf{A}|^{-\frac{3}{2}} \frac{\partial |\mathbf{A}|}{\partial t} && (7.81) \\ \frac{\partial^2 f}{(\partial t)^2} =& -\frac{1}{2} |\mathbf{A}|^{-\frac{3}{2}} \frac{\partial^2 |\mathbf{A}|}{(\partial t)^2} + \frac{3}{4} |\mathbf{A}|^{-\frac{5}{2}} \left(\frac{\partial |\mathbf{A}|}{\partial t}\right)^2 && (7.82)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial |\mathbf{A}|}{\partial t} =& |\mathbf{A}| \operatorname{tr}\left[\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial t}\right] && (7.83) \\ \frac{\partial^2 |\mathbf{A}|}{(\partial t)^2} =& |\mathbf{A}| \operatorname{tr}\left[\mathbf{A}^{-1} \frac{\partial^2 \mathbf{A}}{(\partial t)^2}\right] \\ &+ |\mathbf{A}| \operatorname{tr}\left[\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial t}\right] \operatorname{tr}\left[\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial t}\right] \\ &- |\mathbf{A}| \operatorname{tr}\left[\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial t}\right)\left(A^{-1} \frac{\partial \mathbf{A}}{\partial t}\right)\right]. && (7.84)\end{aligned}$$

Using (7.80) and (7.78), we have

$$\frac{\partial |\mathbf{A}|}{\partial t} = |\mathbf{A}| \operatorname{tr}\left[\mathbf{A}^{-1}\left(\mathbf{K}_v \left(2\mathbf{K}_c - \mathbf{K}_d\right) - \mathbf{K}_d \mathbf{K}_v\right)\right] \qquad (7.85)$$

and

$$\begin{aligned}\frac{\partial^2 |\mathbf{A}|}{(\partial t)^2} =& |\mathbf{A}| \operatorname{tr}\left[-\mathbf{A}^{-1} 2\mathbf{K}_v^2\right] \\ &+ |\mathbf{A}| \operatorname{tr}\left[\mathbf{A}^{-1}\left(\mathbf{K}_v\left(2\mathbf{K}_0 - \mathbf{K}_d\right) - \mathbf{K}_d\mathbf{K}_v\right)\right]^2 \\ &- |\mathbf{A}| \operatorname{tr}\left[\left(\mathbf{A}^{-1}\left(\mathbf{K}_v\left(2\mathbf{K}_c - \mathbf{K}_d\right) - \mathbf{K}_d\mathbf{K}_v\right)\right)^2\right].\end{aligned} \tag{7.86}$$

Therefore

$$\begin{aligned}f'' - \frac{(f')^2}{f} =& -\frac{1}{2}|\mathbf{A}|^{-\frac{3}{2}} \frac{\partial^2 |\mathbf{A}|}{(\partial t)^2} + \frac{3}{4}|\mathbf{A}|^{-\frac{5}{2}}\left(\frac{\partial |\mathbf{A}|}{\partial t}\right)^2 \\ &- \frac{\left(\frac{1}{2}|\mathbf{A}|^{-\frac{3}{2}}\frac{\partial |\mathbf{A}|}{\partial t}\right)^2}{|\mathbf{A}|^{-\frac{1}{2}}} \qquad (7.87) \\ =& -\frac{1}{2}|\mathbf{A}|^{-\frac{3}{2}} \frac{\partial^2 |\mathbf{A}|}{(\partial t)^2} + \frac{1}{2}|\mathbf{A}|^{-\frac{5}{2}}\left(\frac{\partial |\mathbf{A}|}{\partial t}\right)^2 \qquad (7.88) \\ =& \frac{1}{2}|\mathbf{A}|^{-\frac{1}{2}} \operatorname{tr}\left[\mathbf{A}^{-1} 2\mathbf{K}_v^2\right. \\ &\left.+ \left(\mathbf{A}^{-1}\left(2\mathbf{K}_c\mathbf{K}_v - \mathbf{K}_v\mathbf{K}_w - \mathbf{K}_w\mathbf{K}_v\right)\right)^2\right]. \qquad (7.89)\end{aligned}$$

We consider the trace and define $g(\cdot)$ as:

$$\begin{aligned}g\left(2\mathbf{K}_c - \mathbf{K}_d\right) =& \operatorname{tr}\left[2\left(\mathbf{K}_d\left(2\mathbf{K}_c - \mathbf{K}_d\right)\right)^{-1}\mathbf{K}_v\mathbf{K}_v\right] \\ &+ \operatorname{tr}\left[\left(\mathbf{K}_d\left(2\mathbf{K}_c - \mathbf{K}_d\right)\right)^{-2}\right. \\ &\left.\left(2\mathbf{K}_c\mathbf{K}_v - \mathbf{K}_v\mathbf{K}_d - \mathbf{K}_d\mathbf{K}_v\right)^2\right].\end{aligned} \tag{7.90}$$

Replace $2\mathbf{K}_c - \mathbf{K}_d$ with $x\mathbf{I}$, we have

$$
\begin{aligned}
g\left(x\right) =& \mathrm{tr}\left[2x^{-1}\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_v\right] \\
&+ \mathrm{tr}\left[x^{-2}\mathbf{K}_d^{-1}\left(x\mathbf{K}_v - \mathbf{K}_d\mathbf{K}_v\right)\mathbf{K}_d^{-1}\left(x\mathbf{K}_v - \mathbf{K}_d\mathbf{K}_v\right)\right] && (7.91) \\
=& \mathrm{tr}\left[2x^{-1}\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_v + \mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_d^{-1}\mathbf{K}_v\right] \\
&+ \mathrm{tr}\left[-x^{-1}\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_v\right. \\
&\quad \left. -x^{-1}\mathbf{K}_v\mathbf{K}_d^{-1}\mathbf{K}_v + x^{-2}\mathbf{K}_v\mathbf{K}_v\right] && (7.92) \\
=& \mathrm{tr}\left[\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_d^{-1}\mathbf{K}_v + x^{-2}\mathbf{K}_v\mathbf{K}_v\right]. && (7.93)
\end{aligned}
$$

It is easy to see that $g\left(x\right)$ is decreasing with x. According to Lemma 6, if $x\mathbf{I} \succeq 2\mathbf{K}_c - \mathbf{K}_d$, then $g\left(x\right) \leq g\left(2\mathbf{K}_c - \mathbf{K}_d\right)$. We let x goes to infinity

$$
\lim_{x\to\infty} g\left(x\right) = \mathrm{tr}\left[\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_d^{-1}\mathbf{K}_v\right]. \tag{7.94}
$$

We then show that $\mathrm{tr}\left[\mathbf{K}_d^{-1}\mathbf{K}_v\mathbf{K}_d^{-1}\mathbf{K}_v\right]$ is positive. We note that $\mathbf{K}_d$ is a positive definite matrix and $\mathbf{K}_v$ is a symmetric matrix. According to [66, Theorem 7.6.3], $\mathbf{K}_d^{-1}\mathbf{K}_v$ is a diagonalizable matrix, all of whose eigenvalues are real. Let

$$
\mathbf{K}_d^{-1}\mathbf{K}_v = \mathbf{P}^{-1}\mathbf{D}\mathbf{P} \tag{7.95}
$$

where $\mathbf{P}$ is an invertible matrix and $\mathbf{D}$ is a diagonal matrix whose entries are real. Therefore

$$
\left(\mathbf{K}_d^{-1}\mathbf{K}_v\right)^2 = \mathbf{P}^{-1}\mathbf{D}^2\mathbf{P}. \tag{7.96}
$$

We have

$$\operatorname{tr}\left[\left(\mathbf{K}_d^{-1}\mathbf{K}_v\right)^2\right] = \operatorname{tr}\left[\mathbf{D}^2\right]. \tag{7.97}$$

Thus we have shown that $g(x)$ is a decreasing function and it approaches a nonnegative number as x grows. Therefore, by Lemma (6) the trace in (7.89) is also positive. We complete the proof of the nonnegativeness of $f'' - (f')^2/f$.

The proof that h'' is nonnegative

In order to show that h'' is nonnegative, we show that h'' can be rearranged into quadratic form. First, we rewrite T_1 as

$$T_1 = -\frac{1}{2}\left(\mathbf{m}_c - \mathbf{m}_d\right)^\top \left(\frac{1}{2}\mathbf{K}_c^{-1} - \mathbf{K}_d^{-1}\right)^{-1} \left(\mathbf{m}_c - \mathbf{m}_d\right) \tag{7.98}$$

$$= -\frac{1}{2}\left(\sum_i w_i \mathbf{K}_i \mathbf{m}_c - \sum_i w_i \mathbf{K}_i \mathbf{m}_i\right)^\top \left(\frac{1}{2}\mathbf{K}_d\mathbf{K}_c^{-1}\mathbf{K}_d - \mathbf{K}_d\right)^{-1} \left(\sum_i w_i \mathbf{K}_i \mathbf{m}_c - \sum_i w_i \mathbf{K}_i \mathbf{m}_i\right) \tag{7.99}$$

$$= -\frac{1}{2}\left(\sum_i w_i \mathbf{K}_i \left(\mathbf{m}_c - \mathbf{m}_i\right)\right)^\top \left(\frac{1}{2}\mathbf{K}_d\mathbf{K}_c^{-1}\mathbf{K}_d - \mathbf{K}_d\right)^{-1} \left(\sum_i w_i \mathbf{K}_i \left(\mathbf{m}_c - \mathbf{m}_i\right)\right) \tag{7.100}$$

$$= -\frac{1}{2}\mathbf{m}_w^\top \left(\frac{1}{2}\mathbf{K}_d\mathbf{K}_c^{-1}\mathbf{K}_d - \mathbf{K}_d\right)^{-1} \mathbf{m}_w. \tag{7.101}$$

Replace w_i with $w_i + tv_i$, we obtain the expression of $h\left(t\right)$:

$$h\left(t\right) = -\frac{1}{2}\mathbf{m}_{w+tv}^{\top}\left(\frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} - \mathbf{K}_{w+tv}\right)^{-1}\mathbf{m}_{w+tv}. \tag{7.102}$$

The first order derivative is

$$\begin{aligned}\frac{\partial h}{\partial t} =& -\mathbf{m}_{v}^{\top}\left(\frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} - \mathbf{K}_{w+tv}\right)^{-1}\mathbf{m}_{w+tv}\\ &+ \frac{1}{2}\mathbf{m}_{w+tv}^{\top}\left(\frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} - \mathbf{K}_{w+tv}\right)^{-1}\\ &\left(\frac{1}{2}\mathbf{K}_{v}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} + \frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{v} - \mathbf{K}_{v}\right)\\ &\left(\frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} - \mathbf{K}_{w+tv}\right)^{-1}\mathbf{m}_{w+tv}. \end{aligned} \tag{7.103}$$

Let

$$\mathcal{F}\left(\mathbf{K}_{w+tv}\right) = \mathbf{K}_{w+tv}\frac{1}{2}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} - \mathbf{K}_{w+tv} \tag{7.104}$$

$$\mathcal{G}\left(\mathbf{K}_{w+tv}\right) = \frac{1}{2}\mathbf{K}_{v}\mathbf{K}_c^{-1}\mathbf{K}_{w+tv} + \frac{1}{2}\mathbf{K}_{w+tv}\mathbf{K}_c^{-1}\mathbf{K}_{v} - \mathbf{K}_{v}. \tag{7.105}$$

The second order derivative becomes

$$
\begin{aligned}
\frac{\partial^2 h}{\partial t^2} =& - \mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&+ \mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&+ \frac{1}{2}\mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&- \frac{1}{2}\mathbf{m}_{w+tv}^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \\
&\mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&+ \frac{1}{2}\mathbf{m}_{w+tv}^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \left(\frac{1}{2}\mathbf{K}_v\mathbf{K}_c^{-1}\mathbf{K}_v + \frac{1}{2}\mathbf{K}_v\mathbf{K}_c^{-1}\mathbf{K}_v\right) \\
&\left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&- \frac{1}{2}\mathbf{m}_{w+tv}^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \\
&\mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_{w+tv} \\
&+ \frac{1}{2}\mathbf{m}_{w+tv}^\top \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_{w+tv}\right) \left(\mathcal{F}\left(\mathbf{K}_{w+tv}\right)\right)^{-1} \mathbf{m}_v. \qquad (7.106)
\end{aligned}
$$

Let $t = 0$ and rearrange the terms, we have

$$
\begin{aligned}
\left.\frac{\partial^2 h}{\partial t^2}\right|_{t=0} =& - \mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_v \\
&+ 2\mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_w \\
&- \mathbf{m}_w^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \\
&\mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_w \\
&+ \frac{1}{2}\mathbf{m}_w^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \left(\mathbf{K}_v\mathbf{K}_c^{-1}\mathbf{K}_v\right) \\
&\left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_w. \qquad (7.107)
\end{aligned}
$$

Since $\mathbf{K}_v \mathbf{K}_c^{-1} \mathbf{K}_v$ is a positive definite matrix, the last term is positive. Thus we have

$$\begin{aligned}
\left.\frac{\partial^2 h}{\partial t^2}\right|_{t=0} \geq & - \mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_v \\
& + 2\mathbf{m}_v^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_w \\
& - \mathbf{m}_w^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \\
& \mathcal{G}\left(\mathbf{K}_w\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathbf{m}_w \qquad (7.108) \\
= & - \left(\mathbf{m}_v^\top - \mathbf{m}_w^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right)\right) \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \\
& \left(\mathbf{m}_v^\top - \mathbf{m}_w^\top \left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1} \mathcal{G}\left(\mathbf{K}_w\right)\right)^\top . \qquad (7.109)
\end{aligned}$$

This is a quadratic form. Because $\left(\mathcal{F}\left(\mathbf{K}_w\right)\right)^{-1}$ is a negative definite matrix, (7.109) is positive. Therefore we have twenty nodes that work on the

$$\frac{\partial^2 h}{\partial t^2} \geq 0. \qquad (7.110)$$

This completes the proof. □