A Bayesian Framework for Noise Covariance Estimation Using the Facet Model

Desikachari Nadadur, *Member, IEEE*, Robert Martin Haralick, *Fellow, IEEE*, and David Earl Gustafson

***Abstract*—In image processing literature, thus far, researchers have assumed the perturbation in the data to be *white* (or uncorrelated) having a covariance matrix $\sigma^2 \boldsymbol{I}$, i.e., assumption of equal variance for all the data samples and that no correlation exists between the data samples. However, there have been very few attempts to estimate noise characteristics under the assumption that there is a correlation between data samples. In this work, we propose a new and a novel approach for the simultaneous Bayesian estimation of the *unknown* colored or correlated noise (population) covariance matrix and the hyperparameters of the covariance model using the well-known *facet model*. We also estimate the facet model coefficients. We use the facet model because of its simple, yet elegant, mathematical formulation. We use the *generalized inverted Wishart* density as the prior model for the noise covariance matrix. We place a structure on the covariance matrix using the parameters of a *correlation filter*. These hyperparameters are estimated by a *new* extension of the expectation-maximization algorithm called the *generalized constrained expectation maximization* algorithm that we developed.**

***Index Terms*—Colored noise, constraints, correlation filter, expectation-maximization (EM) algorithm, generalized constrained expectation maximization (GCEM) algorithm, generalized inverted Wishart ($\mathcal{GIW}$), hypercovariance, hyperparameters, inverted Wishart ($\mathcal{IW}$), noise covariance matrix, nonlinear programming, white noise.**

I. Introduction

THE field of *image processing* and *computer vision*, in layman's terms, is a science of making the computers to perceive and understand the real world scenes. Commonly, one of the first and foremost steps involved in any *computer vision* task is the preprocessing of the digital images that are created by a sensor [for example, a camera, a medical ultrasound (US) scanning system, and a magnetic resonance imaging (MRI) system]. This involves noise cleaning and enhancement. Images, so enhanced, are passed through the subsequent steps that are to be carried out to realize a computer vision task. A typical computer vision system is illustrated in Fig. 1. An *ellipse* represents an algorithm, procedure or method acting upon a data object represented by a *rectangle*. A digital image formed by sensing the real world is preprocessed to clean up the noise and enhance the image content for subsequent processing. The enhanced image might then be processed to extract features that are of interest to us in a particular application. The features of interest could be gradient magnitude and direction, edge, ridge, and/or valley pixels, or some other information, such as histogram statistics, to name a few. This step is usually referred to as *feature extraction*. These two steps collectively are called the *low-level computer vision task*. Once the these low-level features are detected, the next step could be to group them together and label them into more reasonable entities. This process could involve, for example, connecting detected edge pixels to form longer edge segments, by using some criteria. This step is referred to as *grouping and labeling* or *transformation to higher level entities*. This step is sometimes called *mid-level computer vision* task. Finally, these features are input to the *high-level computer vision* task of *object recognition* to detect object(s) of interest. As one can clearly see, that output of one step forms the input to the next. Consequently, any errors occurring in one step get propagated to the subsequent step and the output of the high-level vision task may not produce desired results. In recent years, increased attention has been paid to the development of algorithms for performing mid- and high-level vision tasks. Researchers are concentrating less on the low-level task of *noise estimation and enhancement*. For example, recent works on boundary detection and three dimensional reconstruction of the organs of the body using deformable models and templates [1]–[6] concentrated on finding the object of interest without heed to the accuracy of the results under the chosen noise model. They perform Gaussian smoothing of images assuming white noise with an assumed variance (covariance matrix $\sigma^2 \boldsymbol{I}$). That is, an assumption of equal variance for all the data samples with no correlation among them. They did not even estimate

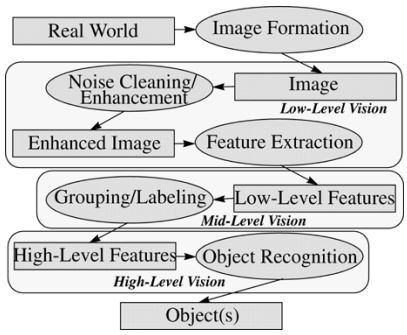


Fig. 1. Typical computer vision system.

Manuscript received May 6, 2002; revised September 21, 2004. This work was supported by Siemens Medical Solutions USA, Inc., Ultrasound Division. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick Bouthemy.

D. Nadadur is with the Advanced Imaging Applications, Developing Competency Department, Siemens Medical Solutions USA, Inc., Ultrasound Division, Issaquah, WA 98029 USA (e-mail: desikachari.nadadur@siemens.com; dnadadur@comcast.net).

R. M. Haralick is with the Department of Computer Science, Graduate Center, City University of New York, NY 10016 USA (e-mail: haralick@gc.cuny.edu).

D. E. Gustafson is with the Developing Competency Department, Siemens Medical Solutions USA, Inc., Ultrasound Division, Issaquah, WA 98029 USA (e-mail: david.gustafson@siemens.com).

Digital Object Identifier 10.1109/TIP.2005.854480

this variance from data. Others went on to estimate the noise variance σ^2 by techniques such as least squares fitting of the observed image data. Use of white noise assumption does not reflect the true correlation between samples that are observed in the real world. Therefore, this modeling error carries on to the higher level computer vision tasks such as feature extraction, perceptual grouping, and final scene reconstruction, thereby producing suboptimal results. We believe that, by properly modeling the correlation of the samples, it is possible to reduce the modeling errors, which then propagate through the various steps of computer vision algorithms to produce optimal results. Our belief is that a computer vision system without a proper modeling of the noise and the estimation of parameters of these models at each stage will not produce optimal and stable results.

This paper concentrates on the above mentioned *low-level vision* tasks of *noise covariance estimation* and *feature extraction*.

A. Background and Motivation

As stated earlier, researchers have focused very little energy, in this very important field of low-level image processing to estimate noise characteristics under the assumption that there is correlation between data samples. Most of the recent works in the covariance matrix estimation could only be found in statistical literature, for example, non-Bayesian approaches by Stein [7], [8], Lin and Perlman [9], and emperical Bayes procedure by Haff [10], fully Bayesian approach by Dickey *et al.* [11], the path models of Wright [12], the LISREL models of Jöreskog [13] and the factor analysis models of Spearman [14], mixture models of Hoffbeck [15], Toeplitz models of Cadre [16], matrix logarithm models of Leonard and Hsu [17], and hierarchical nonconjugate prior models of Daniels [18]. Brown *et al.* [19] propose a new class of priors for the covariance matrix, called the *generalized inverted Wishart* ($\mathcal{GIW}$) prior, which is based on the Bartlett decomposition and apply it to the estimation of the covariance matrix in the environmental monitoring problem. We used this prior model as it naturally applies to the problem that we are trying to solve. Therefore, a lack of proper (general) framework for noise covariance matrix estimation in the field of image processing and computer vision, motivated us to undertake this research work.

B. Organization of the Paper

This paper is organized as follows. In Section II, we introduce the facet model and the additive Gaussian noise model that we will be using in the formulation of the problem of simultaneous estimation of the *noise covariance matrix* and the *facet model coefficients* in a Bayesian framework. We make use of the $\mathcal{GIW}$ density as the prior probability function for the unknown noise covariance matrix. We derive an objective function for the maximum *a posteriori* (MAP) estimation of the facet model coefficients and the noise covariance matrix. This section formally states the problem that needs to be solved. Section III discusses the correlation structure that we place on the hypercovariance matrix describing the $\mathcal{GIW}$ distribution. We design a *generalized constrained expectation maximization* (GCEM) algorithm for the estimation of the hyperparameters used to define this structure. In Section IV, we implement the *constrained maximization* (M-step) of the GCEM algorithm using *sequential unconstrained minimization technique* (SUMT). We convert the constrained problem into a sequence of unconstrained problems using *barrier functions*. Section V discusses the evaluation protocol used to assess the performance of the algorithm and validate the results produced. We do this via the statistical hypothesis testing. Section VI discusses briefly an application that we developed to take advantage of the noise covariance matrix estimation procedure. In Section VII, we summarize the results reported in this paper.

II. Problem Formulation

A. Introduction

In general, the solution to an image processing problem starts by specifying a model for the underlying true noise-free image, a noise or perturbation model that describes how the underlying true noise-free data might have been contaminated, and designing an algorithm to use these models to solve the problem at hand more accurately. By having such a statistical model, one can produce error estimates for the results produced so that the performance of the algorithm can be evaluated. In this section, we formulate the problem of noise covariance matrix estimation by using the *facet model* to describe the underlying noise free image and use an *additive* Gaussian noise model. The facet model was first introduced by Haralick [20]–[22]. This has been extensively used to extract features, such as edges, ridges, and corners of objects in images. See [23]–[30] for examples.

B. Facet Model

The principle of the facet model states that the image can be thought of as an underlying continuum or piecewise continuous gray level intensity surface. The observed digital image is a noisy, discretized sampling of a distorted version of this surface.

For the nth image neighborhood, we can write the facet model [22] representing the ideal noise free signal energy as

$$\boldsymbol{s}_n = \boldsymbol{B}\boldsymbol{\alpha}_n \tag{1}$$

where $\boldsymbol{s}_n$ represents the $n = 1, \ldots, N$ K-dimensional noiseless vectors from the signal space and $\boldsymbol{B}$ is a matrix whose columns represent the discrete orthonormal polynomial (DOP) basis of the space which is modeled to contain the signal energy and $\boldsymbol{\alpha}_n$ are $n = 1, \ldots, N$ M-dimensional vectors of coefficients of the facet model, and $K = (2R+1) \times (2R+1)$ where R is the half-width of the discrete support of the neighborhood. $\boldsymbol{B}$ is obtained as discussed in [22], [31], and [32].

C. Noise Model

Let $\boldsymbol{x}_n; n = 1, \ldots, N$ be the N K-dimensional independent samples of the noisy observed signal. Then, the facet model represents this noisy signal as

$$\boldsymbol{x}_n = \boldsymbol{B}\boldsymbol{\alpha}_n + \boldsymbol{\eta}_n \tag{2}$$

where $\boldsymbol{\eta}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ are the independent identically distributed Gaussian random variates. Arranging these N samples

together into columns of a matrix $\boldsymbol{X}$ and using the notation given in [33] for matrix variate normal distribution, we can write

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{\Lambda} + \boldsymbol{\eta} \tag{3}$$

where

$$\begin{aligned} \boldsymbol{X} \in \mathbb{R}^{K\times N} &= (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \\ \boldsymbol{\Lambda} \in \mathbb{R}^{M\times N} &= (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N) \\ \boldsymbol{\eta} \in \mathbb{R}^{K\times N} &= (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_N). \end{aligned} \tag{4}$$

In this notation, we write the matrix normal distribution as $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \otimes \boldsymbol{I}_N)$, where $\otimes$ is the *Kronecker product* of matrices. Hence, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{B}\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\eta} \otimes \boldsymbol{I}_N)$. This distribution denotes that the columns of $\boldsymbol{X}$ are independent and that the covariance matrix within each column is $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} \in \mathbb{R}^{K\times K}$. For some details on matrix normal distribution, see Appendix I [32]. Our goal is to estimate the polynomial coefficients $\boldsymbol{\alpha}_n$ and the common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

D. Estimation Problem

Given $\boldsymbol{X}$, we wish to estimate $\boldsymbol{\Lambda}$ and the common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. We formulate the problem solution in a Bayesian framework to estimate $\widehat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}$ of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ to *maximize* the posterior probability function

$$p(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}} | \boldsymbol{X}). \tag{5}$$

The values of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ that maximize $p(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}} | \boldsymbol{X})$ will also maximize the joint probability

$$p(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}, \boldsymbol{X}) = p(\boldsymbol{X} | \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) p(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}). \tag{6}$$

We model $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ as being independent *a priori* and we take the prior probability for each $\boldsymbol{\alpha}_n$ to be uniform and we denote the prior probability for $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ as $p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ (see Section II-H for a discussion of the prior probability function used). For detailed derivations of all the expressions in this section, consult [32].

Under these assumptions, we have

$$p(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}, \boldsymbol{X}) = p(\boldsymbol{X} | \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}). \tag{7}$$

We assume the probability function for the likelihood function as

$$p(\boldsymbol{X} | \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) = \frac{\exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{\alpha}_n)' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} (\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{\alpha}_n) \right\}}{(2\pi)^{\frac{NK}{2}} \det \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{\frac{N}{2}}}. \tag{8}$$

Using the stated assumptions and likelihood function and neglecting the constant terms, we write the maximization problem as a minimization of the negative logarithm of the joint probability function as

$$\begin{aligned} \epsilon^2(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) = &-\frac{2}{N} \log p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \log \det \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \\ &+ \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{\alpha}_n)' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} (\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{\alpha}_n). \end{aligned} \tag{9}$$

1) Estimation of the Facet Model Coefficients: This section describes the minimization of ϵ^2 with respect to $\boldsymbol{\alpha}_k$. Taking partial derivatives of ϵ^2 with respect to $\boldsymbol{\alpha}_k$, equating it to zero and solving the resulting equation, we get the estimate $\widehat{\boldsymbol{\alpha}}_k$ of $\boldsymbol{\alpha}_k$ as

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} \right)^{-1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{x}_k. \tag{10}$$

2) Estimation of the Common Noise Covariance Matrix: In this section, we carry out the minimization of ϵ^2 with respect to $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. Eliminating $\boldsymbol{\alpha}_k$ using $\widehat{\boldsymbol{\alpha}}_k$ from the objective function ϵ^2, we get

$$\begin{aligned} \epsilon^2(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) = &-\frac{2}{N} \log p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \log \det \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \\ &+ \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n' \left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} \left(\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} \right)^{-1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \right) \boldsymbol{x}_n. \end{aligned} \tag{11}$$

Notice that the operator $(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} (\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B})^{1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1})$ is one whose columns and rows are orthogonal to the columns of $\boldsymbol{B}$. That is, $(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} (\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B})^{-1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}) \boldsymbol{B} = 0$ and $\boldsymbol{B}' (\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} (\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B})^{-1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}) = 0$.

This means that the only part of $\boldsymbol{x}_n$ that makes a difference in the quadratic term is that part that is in the complement space to the column space of $\boldsymbol{B}$. This part can be written as $(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}')\boldsymbol{x}_n$, since $\boldsymbol{B}\boldsymbol{B}'$ is the orthogonal projection operator to the space spanned by the columns of $\boldsymbol{B}$. Let $\boldsymbol{y}_n = (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}')\boldsymbol{x}_n$, $\boldsymbol{\Sigma}_{\boldsymbol{yy}} = (1/N) \sum_{n=1}^{N} \boldsymbol{y}_n \boldsymbol{y}_n'$ and removing terms independent of the summation index, n out of the summation, we get

$$\begin{aligned} \epsilon^2(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) = &-\frac{2}{N} \log p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}) + \log \det \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \\ &+ \operatorname{tr} \left(\left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} \left(\boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \boldsymbol{B} \right)^{-1} \boldsymbol{B}' \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \right) \boldsymbol{\Sigma}_{\boldsymbol{yy}} \right). \end{aligned} \tag{12}$$

E. When is This Approach an Improvement Over the White Noise Assumption?

Theorem 1 (Case of Uncorrelated Signal and Noise Spaces): When the noise covariance matrix, expressed in the coordinates of the signal space and its complement space is block diagonal with no coupling (i.e., no correlation) between the two spaces, then the same result is obtained for $\boldsymbol{\alpha}_k$ as in the white noise case.

Proof: Under the white noise assumption, we have $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \sigma^2 \boldsymbol{I}$. Therefore, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} = (1/\sigma^2)\boldsymbol{I}$. Now, the DOP basis coefficients $\boldsymbol{\alpha}_k$ become

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{B}' \frac{1}{\sigma^2} \boldsymbol{I} \boldsymbol{B} \right)^{-1} \boldsymbol{B}' \frac{1}{\sigma^2} \boldsymbol{I} \boldsymbol{x}_k = \boldsymbol{B}' \boldsymbol{x}_k. \tag{13}$$

Let $\boldsymbol{C}$ be a matrix whose columns form an orthonormal basis for the space complement to that spanned by the columns of $\boldsymbol{B}$. In this case, the matrix $(\boldsymbol{B}\ \boldsymbol{C})$ is an orthonormal basis with inverse $(\boldsymbol{B}\ \boldsymbol{C})'$. Now

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = (\boldsymbol{B} \quad \boldsymbol{C})(\boldsymbol{B} \quad \boldsymbol{C})' \boldsymbol{\Sigma}_{\boldsymbol{\eta}} (\boldsymbol{B} \quad \boldsymbol{C})(\boldsymbol{B} \quad \boldsymbol{C})'. \tag{14}$$

The matrix $(\boldsymbol{B}\ \boldsymbol{C})' \boldsymbol{\Sigma}_{\boldsymbol{\eta}} (\boldsymbol{B}\ \boldsymbol{C})$ is the noise covariance matrix expressed in the coordinate system of the basis formed by the columns of $(\boldsymbol{B}\ \boldsymbol{C})$.

Let us examine the case when this matrix is block diagonal, i.e.,

$$(\boldsymbol{B} \quad \boldsymbol{C})'\boldsymbol{\Sigma_\eta}(\boldsymbol{B} \quad \boldsymbol{C}) = \begin{pmatrix} \boldsymbol{\Sigma_{BB}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_{CC}} \end{pmatrix}. \tag{15}$$

Using $\boldsymbol{\Sigma_\eta^{-1}}$ computed using the above equation, $\widehat{\boldsymbol{\alpha}}_k$ becomes

$$\widehat{\boldsymbol{\alpha}}_k = \boldsymbol{B}'\boldsymbol{x}_k. \tag{16}$$

■

We understand the above result in the following manner: Because of the relationship $\boldsymbol{x}_k = \boldsymbol{B}\boldsymbol{\alpha}_k + \boldsymbol{\eta}_k$, when the least squares estimate $\widehat{\boldsymbol{\alpha}}_k$ of $\boldsymbol{\alpha}_k$ is produced, all the energy of the noise in the space spanned by the columns of $\boldsymbol{B}$ goes into the computed $\widehat{\boldsymbol{\alpha}}_k$. The noise in the space (spanned by the columns of $\boldsymbol{C}$) complement to the space spanned by the columns of $\boldsymbol{B}$ will only influence the estimate $\widehat{\boldsymbol{\alpha}}_k$ of $\boldsymbol{\alpha}_k$ to the extent that there is coupling between the signal space and the signal complement space (noise space). It is also clear that we only observe the signal from the space spanned by the columns of $\boldsymbol{B}$. We have no *direct* observations from the space spanned by the columns of $\boldsymbol{C}$, but only through the observations from the space spanned by the columns of $\boldsymbol{B}$.

F. General Case

In this section, we discuss the general case when there may be correlation between the space spanned by the columns of $\boldsymbol{B}$ and that spanned by the columns of $\boldsymbol{C}$.

Theorem 2 (Case of Correlated Signal and Noise Spaces): Let $\boldsymbol{\Sigma_{BB}}$ be the component of the noise covariance matrix in the space spanned by the columns of $\boldsymbol{B}$, $\boldsymbol{\Sigma}_{\mathrm{CC}}$ be the component of the noise covariance matrix in the space spanned by the columns of $\boldsymbol{C}$, and $\boldsymbol{\Sigma}_{\mathrm{BC}}$ and $\boldsymbol{\Sigma}_{\mathrm{CB}}$ be the cross-covariance matrices of the spaces spanned by $\boldsymbol{B}$ and $\boldsymbol{C}$, such that $\boldsymbol{\Sigma}_{\mathrm{BC}} = \boldsymbol{\Sigma}'_{\mathrm{CB}}$. Then

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{B}' - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C}'\right)\boldsymbol{x}_k. \tag{17}$$

Proof: Consider the equation

$$\widehat{\boldsymbol{\alpha}}_k = \left(\bar{\boldsymbol{B}}'\boldsymbol{\Sigma_\eta^{-1}}\bar{\boldsymbol{B}}\right)^{-1}\bar{\boldsymbol{B}}'\boldsymbol{\Sigma_\eta^{-1}}\boldsymbol{x}_k \tag{18}$$

where the noise covariance matrix $\boldsymbol{\Sigma_\eta}$ is represented in the basis of $(\bar{\boldsymbol{B}}\ \boldsymbol{C})$. Let $\boldsymbol{\Sigma_\eta}$ represented in the basis of $(\bar{\boldsymbol{B}}\ \boldsymbol{C})$ be partitioned as

$$\begin{aligned}\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}} \in \mathbb{R}^{K\times K} &= (\boldsymbol{B} \quad \boldsymbol{C})'\boldsymbol{\Sigma_\eta}(\boldsymbol{B} \quad \boldsymbol{C}) \\ &= \begin{pmatrix} \boldsymbol{\Sigma_{BB}} \in \mathbb{R}^{M\times M} & \boldsymbol{\Sigma_{BC}} \in \mathbb{R}^{M\times Q} \\ \boldsymbol{\Sigma_{CB}} \in \mathbb{R}^{Q\times M} & \boldsymbol{\Sigma_{CC}} \in \mathbb{R}^{Q\times Q} \end{pmatrix}. \end{aligned} \tag{19}$$

Using $\boldsymbol{\Sigma_\eta^{-1}}$ computed using the above partitioned structure, we get the estimate $\widehat{\boldsymbol{\alpha}}_k$ of $\boldsymbol{\alpha}_k$ as

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{B}' - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C}'\right)\boldsymbol{x}_k. \tag{20}$$

■

Let

$$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{C}'\boldsymbol{x}_n\boldsymbol{x}_n'\boldsymbol{C}. \tag{21}$$

Using these results, the objective function in (9) becomes

$$\epsilon^2(\boldsymbol{\Sigma_\eta}) = -\frac{2}{N}\log p(\boldsymbol{\Sigma_\eta}) + \log\det\boldsymbol{\Sigma_\eta} + \mathrm{tr}\left(\boldsymbol{\Sigma_{CC}^{-1}}\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}}\right). \tag{22}$$

From the above equations, we can write $\log\det\boldsymbol{\Sigma_\eta}$ as

$$\log\det\boldsymbol{\Sigma_\eta} = \log\det\boldsymbol{\Sigma_{CC}} + \log\det\left(\boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}\right) \tag{23}$$

where $\det(\bar{\boldsymbol{B}}\ \boldsymbol{C})(\bar{\boldsymbol{B}}\ \boldsymbol{C})' = 1$. We can replace $p(\boldsymbol{\Sigma_\eta})$ by $p(\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}})$ since $\det(\boldsymbol{B}\ \boldsymbol{C})'(\boldsymbol{B}\ \boldsymbol{C})$, the determinant of the *Jacobian* of the transformation is unity. Therefore, the objective function can be written as

$$\begin{aligned}\epsilon^2\left(\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}}\right) &= -\frac{2}{N}\log p\left(\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}}\right) + \log\det\boldsymbol{\Sigma_{CC}} \\ &+ \log\det\left(\boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}\right) + \mathrm{tr}\left(\boldsymbol{\Sigma_{CC}^{-1}}\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}}\right). \end{aligned} \tag{24}$$

In Section II-H, we discuss the probability density function used to describe our *a priori* knowledge about $\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}}$.

G. Noise Model Revisited

The random vectors in the $(\boldsymbol{B}\ \boldsymbol{C})$ space can be written as

$$\begin{aligned}\boldsymbol{y}_n \in \mathbb{R}^{K\times 1} = (\boldsymbol{B} \quad \boldsymbol{C})'\boldsymbol{x}_n &= \begin{pmatrix} \boldsymbol{y}_{\boldsymbol{B}_n} \in \mathbb{R}^{M\times 1} \\ \boldsymbol{y}_{\boldsymbol{C}_n} \in \mathbb{R}^{Q\times 1} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\alpha}_n + \boldsymbol{B}'\boldsymbol{\eta}_n \\ \boldsymbol{0} + \boldsymbol{C}'\boldsymbol{\eta}_n \end{pmatrix}. \end{aligned} \tag{25}$$

In matrix form, data $\boldsymbol{X}$ in the $(\boldsymbol{B}\ \boldsymbol{C})$ space be represented by $\boldsymbol{Y}$ and given by

$$\boldsymbol{Y} \in \mathbb{R}^{K\times N} = (\boldsymbol{B} \quad \boldsymbol{C})'\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Y_B} \in \mathbb{R}^{M\times N} \\ \boldsymbol{Y_C} \in \mathbb{R}^{Q\times N} \end{pmatrix}. \tag{26}$$

Under this updated noise model, the estimate $\widehat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ is

$$\widehat{\boldsymbol{\Lambda}} = \boldsymbol{Y_B} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C}. \tag{27}$$

H. Choice of Prior Probability Function for the Noise Covariance Matrix

Brown *et al.* [19] proposed what they call the $\mathcal{GIW}$ distribution, which is based on the partitioned structure of the covariance matrix, and derived using the 1-1 Bartlett [34] matrix decomposition. For some details on the inverted Wishart ($\mathcal{IW}$) and $\mathcal{GIW}$ distributions, see Appendices II and III.

Let $\boldsymbol{\Sigma_\eta} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi_\eta})$. Then, $\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi}^*_{\boldsymbol{\eta}})$. If $\boldsymbol{\Sigma}^*_{\boldsymbol{\eta}} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi}^*_{\boldsymbol{\eta}})$, then this exactly corresponds to the $\mathcal{GIW}$ distribution with parameter δ and partitioned $\boldsymbol{\Psi}^*_{\boldsymbol{\eta}}$ as given below.

The data matrix $\boldsymbol{Y}$ in (26) represents the partitioning of data conformable with that of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*}$. Also, let the matrix $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*}$ be partitioned, conformably with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*}$, as

$$\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*} = (\boldsymbol{B} \quad \boldsymbol{C})'\boldsymbol{\Psi}_{\boldsymbol{\eta}}(\boldsymbol{B} \quad \boldsymbol{C}) = \begin{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{BB}} & \boldsymbol{\Psi}_{\boldsymbol{BC}} \\ \boldsymbol{\Psi}_{\boldsymbol{CB}} & \boldsymbol{\Psi}_{\boldsymbol{CC}} \end{pmatrix}. \tag{28}$$

The $\mathcal{GIW}$ distribution for our problem is then described as

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*} \in \mathbb{R}^{K\times K} \sim \mathcal{GIW}\left(\delta, \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*}\right) \tag{29}$$

implies that

1) $\boldsymbol{\Sigma}_{\boldsymbol{CC}} \in \mathbb{R}^{Q\times Q} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi}_{\boldsymbol{CC}})$, is distributed independently of $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$;
2) $\boldsymbol{\zeta} \in \mathbb{R}^{M\times M} \sim \mathcal{IW}(\delta+Q, \boldsymbol{\zeta}_0)$;
3) $\boldsymbol{\tau} \in \mathbb{R}^{Q\times M} | \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\tau}_0, \boldsymbol{\Psi}_{\boldsymbol{CC}}^{-1} \otimes \boldsymbol{\zeta})$;

where $\boldsymbol{\tau} = \boldsymbol{\Sigma}_{\boldsymbol{CC}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{CB}}$; $\boldsymbol{\zeta} = \boldsymbol{\Sigma}_{\boldsymbol{BB}} - \boldsymbol{\Sigma}_{\boldsymbol{BC}}\boldsymbol{\Sigma}_{\boldsymbol{CC}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{CB}} = \boldsymbol{\Sigma}_{\boldsymbol{BB}} - \boldsymbol{\tau}'\boldsymbol{\Sigma}_{\boldsymbol{CC}}\boldsymbol{\tau}$; $\boldsymbol{\tau}_0 = \boldsymbol{\Psi}_{\boldsymbol{CC}}^{-1}\boldsymbol{\Psi}_{\boldsymbol{CB}}$; $\boldsymbol{\zeta}_0 = \boldsymbol{\Psi}_{\boldsymbol{BB}} - \boldsymbol{\Psi}_{\boldsymbol{BC}}\boldsymbol{\Psi}_{\boldsymbol{CC}}^{-1}\boldsymbol{\Psi}_{\boldsymbol{CB}}$; $K = M + Q$.

I. Updated Objective Function

In the following, we make the dependence of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*}$ on δ and $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*}$ explicit in notation. From the preceding analysis, we write $p(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*})$ as

$$\begin{aligned} p\left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*}\right) &= p\left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*} | \delta, \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*}\right) \\ &= p(\boldsymbol{\Sigma}_{\boldsymbol{CC}} | \delta, \boldsymbol{\Psi}_{\boldsymbol{CC}}) p(\boldsymbol{\tau} | \boldsymbol{\tau}_0, \boldsymbol{\Psi}_{\boldsymbol{CC}}, \boldsymbol{\zeta}) p(\boldsymbol{\zeta} | \delta, \boldsymbol{\zeta}_0). \end{aligned} \tag{30}$$

Using the appropriate density functions for matrix normal and $\mathcal{IW}$ distributions (given in Appendices I and II), we have the objective function to be minimized as

$$\begin{aligned} &\epsilon^2(\boldsymbol{\Sigma}_{\boldsymbol{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta}, \mathcal{H}) \\ &\quad = \frac{\delta + 2Q + N}{2} \log\det \boldsymbol{\Sigma}_{\boldsymbol{CC}} \\ &\qquad + \frac{\delta + 2(Q+M) + N}{2} \log\det \boldsymbol{\zeta} \\ &\qquad + \frac{1}{2}\mathrm{tr}\left([\boldsymbol{\Psi}_{\boldsymbol{CC}} + N\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}}]\boldsymbol{\Sigma}_{\boldsymbol{CC}}^{-1}\right) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\zeta}_0\boldsymbol{\zeta}^{-1}) \\ &\qquad + \frac{1}{2}\mathrm{tr}\left([\boldsymbol{\Psi}_{\boldsymbol{CC}}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)]\left[(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\boldsymbol{\zeta}^{-1}\right]'\right) \end{aligned} \tag{31}$$

where $\mathcal{H}$ represents the hyperparameters to be elaborated later. Therefore, our goal is to estimate $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$, $\boldsymbol{\zeta}$ and $\boldsymbol{\tau}$, by minimizing the objective function given in (31) given the hyperparameters, $\mathcal{H}$. Once $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$, $\boldsymbol{\zeta}$ and $\boldsymbol{\tau}$ are estimated, we can then construct $\boldsymbol{\Sigma}_{\eta}^{*}$ and $\boldsymbol{\Sigma}_{\eta}$ by performing an inverse of the Bartlett decomposition.

1) Estimating $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$: To estimate $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$, minimize the following function:

$$\begin{aligned} \epsilon^2(\boldsymbol{\Sigma}_{\boldsymbol{CC}}) = \frac{\delta + 2Q + N}{2} \log\det \boldsymbol{\Sigma}_{\boldsymbol{CC}} \\ + \frac{1}{2}\mathrm{tr}\left([\boldsymbol{\Psi}_{\boldsymbol{CC}} + N\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}}]\boldsymbol{\Sigma}_{\boldsymbol{CC}}^{-1}\right). \end{aligned} \tag{32}$$

Using Theorem 6, we get the estimate as

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{CC}} = \frac{1}{\delta + 2Q + N}(\boldsymbol{\Psi}_{\boldsymbol{CC}} + N\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{CC}}). \tag{33}$$

2) Estimating $\boldsymbol{\tau}$: By minimizing the following function:

$$\epsilon^2(\tau) = \frac{1}{2}\mathrm{tr}\left([\boldsymbol{\Psi}_{\boldsymbol{CC}}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)]\left[(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\boldsymbol{\zeta}^{-1}\right]'\right) \tag{34}$$

the estimate $\widehat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ is obtained as

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{\tau}_0. \tag{35}$$

3) Estimating $\boldsymbol{\zeta}$: An estimate of $\boldsymbol{\zeta}$ is obtained by minimizing the following function:

$$\begin{aligned} \epsilon^2(\zeta) = \frac{\delta + 2(Q+M) + N}{2} \log\det \boldsymbol{\zeta} \\ + \frac{1}{2}\mathrm{tr}\left([\boldsymbol{\zeta}_0 + (\boldsymbol{\tau} - \boldsymbol{\tau}_0)'\boldsymbol{\Psi}_{\boldsymbol{CC}}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)]\,\boldsymbol{\zeta}^{-1}\right). \end{aligned} \tag{36}$$

Using Theorem 6, and $\boldsymbol{\tau} = \widehat{\boldsymbol{\tau}}$, we get

$$\widehat{\boldsymbol{\zeta}} = \frac{1}{\delta + 2(Q+M) + N}\boldsymbol{\zeta}_0. \tag{37}$$

The above expressions are derived assuming that the hyperparameters are known. However, we do not know these in advance, and, hence, we need to employ an estimation procedure. Also, the optimization becomes tricky if we do not know how the space (signal space) spanned by the columns of $\boldsymbol{B}$ is related to the space (noise space) spanned by the columns of $\boldsymbol{C}$. This is because there exist several choices for the matrices $\boldsymbol{\Sigma}_{\boldsymbol{BB}}$, $\boldsymbol{\Sigma}_{\boldsymbol{BC}} = \boldsymbol{\Sigma}_{\boldsymbol{CB}}'$ and $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$, so that one can obtain a covariance matrix that satisfies the positive-definiteness constraint. However, this may not be the optimal solution. Therefore, it is best to have a model for the correlation between the spaces spanned by the columns of $\boldsymbol{B}$ and $\boldsymbol{C}$. This problem is best solved if we assume that the observations $\boldsymbol{y}_n$ were obtained by passing *white noise* though a filter h defined by a functional form with free parameters. Then, the entries in the covariance matrix become a function of these free parameters. It is better to place this structure on the hypercovariance matrix, $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*}$ rather than on $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{*}$. This is because if a structure is placed at the primary level, then any deviation from the assumed structure during the estimation procedure will result in serious errors [11], while placing the structure on a secondary level induces more flexibility and reduces the total number of parameters that need to be estimated. This is discussed in the next section along with the hyperparameter estimation.

III. Correlation Structure and Hyperparameter Estimation

A. Introduction

In this section, we discuss an approach to estimating the *hypercovariance matrix* and other *hyperparameters* by specifying a correlation structure on the hypercovariance matrix $\boldsymbol{\Psi}_{\eta}^{*}$ (this is the hypercovariance matrix $\boldsymbol{\Psi}_{\eta}$ expressed in $(\boldsymbol{B}\ \boldsymbol{C})$ space). By looking at the noise model in (25), it is clear that the observations $\boldsymbol{X}$ provide only samples from the signal space plus the noise proportional to the correlation between the spaces spanned by the columns of the matrices $\boldsymbol{B}$ and $\boldsymbol{C}$. In other words, we do not have direct observations of $\boldsymbol{Y}_{\boldsymbol{C}}$, but only through $\boldsymbol{X}$. Therefore, we need to, somehow, estimate the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{CC}}$

in the space spanned by the columns of the matrix $\boldsymbol{C}$ (noise space), from the observations $\boldsymbol{X}$. This estimation procedure becomes tricky if we do not have a model for the *correlation* between the space spanned by the columns of $\boldsymbol{B}$ and that spanned by the columns of $\boldsymbol{C}$, as there exist several choices for the matrices $\boldsymbol{\Sigma_{BB}}$, $\boldsymbol{\Sigma_{BC}} = \boldsymbol{\Sigma'_{CB}}$, and $\boldsymbol{\Sigma_{CC}}$ so that one can satisfy the *positive definiteness constraint* on the covariance matrix, $\boldsymbol{\Sigma}^*_\eta$. This may not be an optimal choice.

B. Correlation Structure

Our problem is best solved if we assume that the observations $\boldsymbol{Y}$ were obtained by passing *white noise* through a filter h defined by a functional form with free parameters. Associated with any covariance matrix $\boldsymbol{\Sigma}^*_\eta$ is a filter h. If *white noise* is put through the filter h, the resulting colored noise signal will have covariance matrix directly computable in closed form depending on the values of the components of h. Now, suppose that we consider a class of parametric filters, where the number of parameters of the filter is substantially less than the dimension of η. In this case, $\boldsymbol{\Sigma}^*_\eta$ is a matrix each of whose entries is a function of the parameters of the filter h. As discussed earlier, any requisite structure is placed on the hypercovariance matrix $\boldsymbol{\Psi}^*_\eta$, instead of on the matrix $\boldsymbol{\Sigma}^*_\eta$, as it offers flexibility in estimating $\boldsymbol{\Sigma}^*_\eta$, by allowing us to waver from the assumed structure secondary level when estimating the primary level variables.

To model the correlation, we assume our observations to be coming from a Sth order *moving average process* (MA), indexed by t, and defined by

$$y(t) = \sum_{i=0}^{S} a_i \omega(t - i) \tag{38}$$

where $\omega(t)$ is the *white noise* process with $\boldsymbol{E}[\omega(t)] = 0$ and $\boldsymbol{V}[\omega(t)] = \sigma^2_\omega$.

In the current problem at hand, $t = 1, \ldots, K$ and set $S = K - 1$ so that we make all the samples $y(t)$ correlated with one another. We use ρ to denote the correlation term and γ to denote the covariance term.

The condition on the covariances implies that the covariance (and, hence, the correlation) matrix of the vector $(y_1, y_2, \ldots, y_K)$ has the form

$$\boldsymbol{\Psi}^*_\eta = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{(K-1)} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{(K-2)} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{(K-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{(K-1)} & \gamma_{(K-2)} & \gamma_{(K-3)} & \cdots & \gamma_0 \end{pmatrix} \tag{39}$$

wherein the generic element in the (i, j)th position is $\gamma_{|i-j|} = \mathcal{C}(y_i, y_j)$, the covariance of y_i and y_j.

We have

$$\gamma_\lambda = \sigma^2_\omega \sum_j a_j a_{j+\lambda} \tag{40}$$

where $\lambda = 0, \ldots, K-1$ and $\lambda + j \leq K - 1$.

The correlation coefficient ρ_λ is computed as

$$\rho_\lambda = \frac{\gamma_\lambda}{\gamma_0} \tag{41}$$

and is elaborated into

$$\rho_\lambda = \begin{cases} 1, & \lambda = 0 \\ \dfrac{\sum_j a_j a_{j+\lambda}}{\sum_i a_i^2}, & \lambda = 1, \ldots, (K-1), \text{ and } \lambda + j \leq K - 1 \\ 0, & \lambda > K - 1 \end{cases}. \tag{42}$$

In general, a process like this is normalized by setting either $a_0 = 1$ or $\sigma^2_\omega = 1$. We make the choice of $\sigma^2_\omega = 1$ and estimate the remaining parameters. We call $\mathcal{H} = (\delta, a_0, a_1, \ldots, a_{K-1})$ as our set of *hyperparameters* that are to be estimated. Now, the problem is to estimate the hyperparameters $\mathcal{H}$ given the data $\boldsymbol{Y}$ under the constraints that

$$\left.\begin{array}{l} \delta > 0 \\ |\rho_\lambda| \leq 1, \quad \text{for } \lambda = 1, \ldots, K-1 \\ \boldsymbol{\Psi}^*_\eta \succ 0 \\ \boldsymbol{\Psi_{CC}} \succ 0 \\ \boldsymbol{\zeta}_0 \succ 0 \end{array}\right\}. \tag{43}$$

The first two equations give the element level constraints on the hypercovariance matrix and the rest of the equations provide the higher or matrix-level constraints.

In the following sections, we derive an expectation maximization algorithm for estimating the hyperparameters $\mathcal{H}$. In this process, we derive the much needed joint density of data $(\boldsymbol{Y'_B}, \boldsymbol{Y'_C})'$, and $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$, $\boldsymbol{\zeta}$, and the joint posterior density of $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$, $\boldsymbol{\zeta}$ given data $(\boldsymbol{Y'_B}, \boldsymbol{Y'_C})'$. This also serves to illustrate that the objective function derived in earlier sections by starting with data $\boldsymbol{X}$ and then transforming the objective function to the space spanned by the columns of $(\boldsymbol{B}\ \boldsymbol{C})$ is exactly the same as if we derived the objective function by first transforming the data $\boldsymbol{X}$ to the space spanned by the columns of $(\boldsymbol{B}\ \boldsymbol{C})$ as $\boldsymbol{Y}$ and then worked directly in this space. Directly working in the space spanned by the columns of $(\boldsymbol{B}\ \boldsymbol{C})$ leads to the result in a more straight forward manner.

C. Posterior Density of $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$

In this section, we derive an expression for the *joint posterior density* of $\boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}$, and $\boldsymbol{\zeta}$ given the data $\boldsymbol{Y_B}$ and $\boldsymbol{Y_C}$. In this treatment, we approximate $\boldsymbol{\Lambda}$ by its estimate $\widehat{\boldsymbol{\Lambda}}$ given by [rewritten from (27)]

$$\widehat{\boldsymbol{\Lambda}} = \boldsymbol{Y_B} - \boldsymbol{\tau}' \boldsymbol{Y_C}. \tag{44}$$

Theorem 3 (Joint Posterior Density of $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$): Given

$$\boldsymbol{Y_B} | \boldsymbol{Y_C} \sim \mathcal{N}(\boldsymbol{\Lambda} + \boldsymbol{\tau}' \boldsymbol{Y_C}, \boldsymbol{\zeta} \otimes \boldsymbol{I}_N) \text{ (see Theorem 7)} \tag{45}$$

$$\boldsymbol{Y_C} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_{CC}} \otimes \boldsymbol{I}_N) \tag{46}$$

$$\boldsymbol{\Sigma_{CC}} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi_{CC}}) \tag{47}$$

$$\boldsymbol{\zeta} \sim \mathcal{IW}(\delta + Q, \boldsymbol{\zeta}_0) \tag{48}$$

$$\boldsymbol{\tau} | \boldsymbol{\zeta} \sim \mathcal{N}\left(\boldsymbol{\tau}_0, \boldsymbol{\Psi}^{-1}_{\boldsymbol{CC}} \otimes \boldsymbol{\zeta}\right) \tag{49}$$

$\boldsymbol{\Lambda}$ is assumed to be *a priori* uniform and it is *a priori* independent of $\boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta}$. (50)

Then, the joint posterior density of $(\boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta})$ is given by

$$p(\boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta} | \boldsymbol{\Lambda} = \widehat{\boldsymbol{\Lambda}}, \boldsymbol{Y_B}, \boldsymbol{Y_C}) = p(\boldsymbol{\Sigma_{CC}} | \boldsymbol{Y_B}, \boldsymbol{Y_C}, \delta, \boldsymbol{\Psi_{CC}}) \times p(\boldsymbol{\tau} | \boldsymbol{Y_B}, \boldsymbol{Y_C}, \boldsymbol{\tau}_0, \boldsymbol{\Psi_{CC}}, \boldsymbol{\zeta}) \times p(\boldsymbol{\zeta} | \boldsymbol{Y_B}, \boldsymbol{Y_C}, \delta, \boldsymbol{\zeta}_0) \tag{51}$$

where

$$\boldsymbol{\Sigma_{CC}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\delta,\boldsymbol{\Psi_{CC}} \sim \mathcal{IW}\left(\delta+N,\boldsymbol{\Psi_{CC}}+\boldsymbol{Y_C}\boldsymbol{Y_C}'\right) \quad (52)$$

$$\boldsymbol{\tau}|\boldsymbol{Y_B},\boldsymbol{Y_C},\boldsymbol{\tau}_0,\boldsymbol{\Psi_{CC}},\boldsymbol{\zeta} \sim \mathcal{N}\left(\boldsymbol{\tau}_0,\boldsymbol{\Psi}_{\boldsymbol{CC}}^{-1}\otimes\boldsymbol{\zeta}\right) \quad (53)$$

$$\boldsymbol{\zeta}|\boldsymbol{Y_B},\boldsymbol{Y_C},\delta,\boldsymbol{\zeta}_0 \sim \mathcal{IW}(\delta+Q+N,\boldsymbol{\zeta}_0). \quad (54)$$

Proof: For the proof, see Appendix VI ■

D. Hyperparameter Estimation via the Expectation Maximization Algorithm

In the previous section, we carried out the analysis under the assumption that the hyperparameters $\mathcal{H}$ of the $\mathcal{GIW}$ distribution were known. This works well if we know the values of $\mathcal{H}$ *a priori*. However, in reality, we do not know these values. One must resort to some form of estimation procedure. One of the estimation procedures called the *EM* algorithm was proposed by Dempster [35] for this purpose. The EM algorithm works well when the probability density function for the incomplete data (the covariance matrix, in our case) cannot be computed easily. In that case, the conditional expectation of the logarithm of the probability density function of the incomplete data given the observables and the current point estimate of the parameters, is maximized to obtain the estimate of the parameters for the next iteration. Recall that we placed a structure on $\boldsymbol{\Psi}_\eta^*$ in Section III-B based on the filter coefficients a_i; $i=0,\ldots,K-1$. This translates into $\boldsymbol{\Psi_{CC}}$, $\boldsymbol{\tau}_0$, and $\boldsymbol{\zeta}_0$ being functions of these parameters. In this section, we design an algorithm using EM theory for estimating the hyperparameters, $\mathcal{H}=\{\delta,a_0,a_1,\ldots,a_{K-1}\}$ simultaneously, from the observed data, under the constraints imposed in (43). Additional examples of the usage of EM theory can be found in [36]–[38].

1) Background of EM Theory: Let $\boldsymbol{y}$ be the vector of hidden parameters. The vector $\boldsymbol{x}$ is called the *incomplete data*. The $\boldsymbol{z}=(\boldsymbol{x},\boldsymbol{y})$ is called the *complete data*. Then, the joint density function or the *complete data density function* is

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}). \quad (55)$$

This defines the *complete data likelihood*. This function is a random variable since the missing information $\boldsymbol{y}$ is unknown, random and presumably governed by an underlying distribution. Therefore, this function can be thought of as a function of $\boldsymbol{y}$ given that $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are constants.

The EM algorithm then finds the expected value of the complete data log-likelihood $\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})$ with respect to the unknown data $\boldsymbol{y}$ given the observed data $\boldsymbol{x}$ and the current parameter estimate $\boldsymbol{\theta}^{(t)}$. Therefore, let us define

$$\begin{aligned}\mathfrak{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) &= \boldsymbol{E}\left[\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\right] \\ &= \int_{\boldsymbol{y}} \log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}) p\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\right) d\boldsymbol{y} \end{aligned} \quad (56)$$

where $\boldsymbol{\theta}^{(t)}$ is the current parameter estimate that is used to evaluate the expectation and $\boldsymbol{\theta}$ are the parameters that we find to maximize $\mathfrak{Q}$. The function $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^{(t)})$ is the marginal distribution of the unobserved data and is dependent on both the observations $\boldsymbol{x}$ and the current parameter estimates. Convergence properties of EM algorithm are well understood [35], [39]. Algorithm III.1 describes the steps of the basic EM algorithm.

Algorithm III.1: Basic EM Algorithm

Once $\mathfrak{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is defined the EM algorithm has two steps, the *evaluation of expectation step* or the *E-step* and the *maximization of expection step* or the *M-step*. These two steps of the algorithm are described below.

E-step Given the data $\boldsymbol{x}$ and the parameter estimates $\boldsymbol{\theta}^{(t)}$ at iteration t, evaluate the expectation

$$\mathfrak{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \boldsymbol{E}\left[\log p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\right]. \quad (57)$$

Go to M-step.

M-step Given the expectation computed in E-step, optimize $\mathfrak{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to find $\boldsymbol{\theta}^{(t+1)}$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathfrak{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^t). \quad (58)$$

Set $t=t+1$ and go to E-step.

The E-step and M-step are carried out to convergence.

There is another variation of the algorithm called the *generalized EM algorithm* (GEM) in which, instead of maximizing $\mathfrak{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, we just find some $\boldsymbol{\theta}^{(t+1)}$ such that

$$\mathfrak{Q}\left(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}\right) \geq \mathfrak{Q}\left(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}\right) \quad (59)$$

in the M-step. This because the EM algorithm in its original form is guaranteed to increase the likelihood. We will use the GEM algorithm in the hyperparameter estimation to be discussed in the next section. Recall that the estimates of the hyperparameters must satisfy the constraints given in (43). Therefore, in the next section, we modify the GEM algorithm to incorporate these constraints during the M-step and we christen this approach GCEM algorithm.

2) Hyperparameter Estimation Using the GCEM Algorithm: In this section, we derive $\mathfrak{Q}(\mathcal{H}|\mathcal{H}^{(t)})$ for estimating the hyperparameters defined by $\mathcal{H}=\{\delta,a_0,a_1,\ldots,a_{K-1}\}$. Since $\mathcal{H}$ are the hyperparameters defining $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^*$ partitioned as $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$, and $\boldsymbol{\zeta}$, we conveniently choose these as the hidden variables. We remove the facet model coefficients $\boldsymbol{\Lambda}$ by conditioning as $\boldsymbol{\Lambda}=\widehat{\boldsymbol{\Lambda}}$. This is valid as we are assuming $\boldsymbol{\Lambda}$ to be *a priori* uniform, and, therefore, there are no hyperparameters of $\boldsymbol{\Lambda}$ to be estimated. The data matrix is $\boldsymbol{Y}=(\boldsymbol{Y_B}',\boldsymbol{Y_C}')'$.

Note that $\boldsymbol{\Psi_{CC}}$, $\boldsymbol{\tau}_0$ and $\boldsymbol{\zeta}_0$ are functions of $(a_0,a_1,\ldots,a_{K-1})$. The objective function $\mathfrak{Q}(\mathcal{H}|\mathcal{H}^{(t)})$,

the terms $\log p(\boldsymbol{Y_B}|\boldsymbol{Y_C},\boldsymbol{\tau},\boldsymbol{\zeta})$ and $\log p(\boldsymbol{Y_C}|\boldsymbol{\Sigma_{CC}})$ are independent of $\mathcal{H}$; therefore, we can write it as

$$\begin{aligned}\mathfrak{Q}^*\left(\mathcal{H}|\mathcal{H}^{(t)}\right) &= \boldsymbol{E}\left[\log p(\boldsymbol{\Sigma_{CC}}|\mathcal{H})|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad + \boldsymbol{E}\left[\log p(\boldsymbol{\tau}|\boldsymbol{\zeta},\mathcal{H})|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad + \boldsymbol{E}\left[\log p(\boldsymbol{\zeta}|\mathcal{H})|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]. \end{aligned} \tag{60}$$

The above equation shows that we take expectations of the logarithms of the prior probability density functions of $(\boldsymbol{\Sigma_{CC}},\boldsymbol{\tau},\boldsymbol{\zeta})$ with respect to the posterior density of $(\boldsymbol{\Sigma_{CC}},\boldsymbol{\tau},\boldsymbol{\zeta})$ (computed using Theorem 3), respectively. The expectations are computed using the properties of matrix normal, Wishart and $\mathcal{IW}$ distributions given in Appendices I and II. The objective function to be maximized is

$$\begin{aligned}&\mathfrak{Q}^*\left(\mathcal{H}|\mathcal{H}^{(t)}\right) \\ &= -\frac{\delta(M+Q)}{2}\log 2 - \sum_{i=1}^{Q}\log\Gamma\left[\frac{\delta+Q-i}{2}\right] \\ &\quad - \sum_{i=1}^{M}\log\Gamma\left[\frac{\delta+Q+N+M-i}{2}\right] \\ &\quad + \left(\frac{\delta+Q+M-1}{2}\right)\log\det\boldsymbol{\Psi_{CC}} \\ &\quad - \left(\frac{\delta+2Q}{2}\right)\boldsymbol{E}\left[\log\det\boldsymbol{\Sigma_{CC}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{E}\left[\boldsymbol{\Sigma_{CC}^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\boldsymbol{\Psi_{CC}}\right) \\ &\quad - \left(\frac{\delta+2(M+Q)+N}{2}\right)\boldsymbol{E}\left[\log\det\boldsymbol{\zeta}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{E}\left[\boldsymbol{\tau\zeta^{-1}\tau'}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\boldsymbol{\Psi_{CC}}\right) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{E}\left[\boldsymbol{\tau\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\boldsymbol{\tau_0'\Psi_{CC}}\right) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\tau_0 E}\left[\boldsymbol{\zeta^{-1}\tau'}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\boldsymbol{\Psi_{CC}}\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\tau_0 E}\left[\boldsymbol{\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\boldsymbol{\tau_0'\Psi_{CC}}\right) \\ &\quad + \left(\frac{\delta+Q+N+M-1}{2}\right)\log\det\boldsymbol{\zeta_0} \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{E}\left[\boldsymbol{\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right]\right) \end{aligned} \tag{61}$$

where

$$\begin{aligned}&\boldsymbol{E}\left[\log\det\boldsymbol{\Sigma_{CC}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = -Q\log 2 - \sum_{i=1}^{Q}\mathcal{D}\left(\frac{\delta^{(t)}+N+Q-i}{2}\right) \\ &\qquad + \log\det\left(\boldsymbol{\Psi_{CC}^{(t)}} + \boldsymbol{Y_C Y_C'}\right) \end{aligned} \tag{62}$$

$$\begin{aligned}&\boldsymbol{E}\left[\boldsymbol{\Sigma_{CC}^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = \left(\delta^{(t)}+N+Q-1\right)\left(\boldsymbol{\Psi_{CC}^{(t)}} + \boldsymbol{Y_C Y_C'}\right)^{-1} \end{aligned} \tag{63}$$

$$\begin{aligned}&\boldsymbol{E}\left[\log\det\boldsymbol{\zeta}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = -M\log 2 - \sum_{i=1}^{M}\mathcal{D}\left(\frac{\delta^{(t)}+Q+N+M-i}{2}\right) \\ &\qquad + \log\det\boldsymbol{\zeta_0^{(t)}} \end{aligned} \tag{64}$$

$$\begin{aligned}&\boldsymbol{E}\left[\boldsymbol{\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = \left(\delta^{(t)}+Q+N+M-1\right)\left(\boldsymbol{\zeta_0^{(t)}}\right)^{-1} \end{aligned} \tag{65}$$

$$\begin{aligned}&\boldsymbol{E}\left[\boldsymbol{\tau\zeta^{-1}\tau'}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = \left(\delta^{(t)}+Q+N+M-1\right)\boldsymbol{\tau_0^{(t)}}\left(\boldsymbol{\zeta_0^{(t)}}\right)^{-1}\boldsymbol{\tau_0^{(t)'}} \\ &\qquad + M\left(\boldsymbol{\Psi_{CC}^{(t)}}\right)^{-1} \end{aligned} \tag{66}$$

$$\begin{aligned}&\boldsymbol{E}\left[\boldsymbol{\tau\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}\right] \\ &\quad = \left(\delta^{(t)}+Q+N+M-1\right)\boldsymbol{\tau_0^{(t)}}\left(\boldsymbol{\zeta_0^{(t)}}\right)^{-1}. \end{aligned} \tag{67}$$

Note that $\boldsymbol{E}[\boldsymbol{\zeta^{-1}\tau'}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}] = \boldsymbol{E}[\boldsymbol{\tau\zeta^{-1}}|\boldsymbol{Y_B},\boldsymbol{Y_C},\mathcal{H}^{(t)}]'$.

In the GCEM algorithm, the M-step involves the constraints that are to be satisfied by the hyperparameters. Therefore, we implement generalized M-step by converting the problem to an unconstrained one using SUMT to compute some $\mathcal{H}^{(t+1)}$ which satisfies the condition (59) with $\mathfrak{Q}$ replaced by $\mathfrak{Q}^*$. This optimization procedure is discussed in the next section.

IV. Nonlinear Programming

A. Introduction

In this section, we discuss the nonlinear programming method incorporated in solving the maximization (minimization) problem for estimating the hyperparameters. Further details can be obtained from the excellent reference text [40].

B. Implementation of Generalized Constrained M-Step Using SUMT

We use the *interior point* or *barrier function* methods pioneered by Fiacco and McCormick [41] in solving the problem at hand. Since these methods involve minimization of a sequence of unconstrained problems, they are referred to as SUMTs. In the following sections, we develop the theory necessary for implementing the nonlinear programming algorithm using barrier functions to estimate the hyperparameters $\mathcal{H}^{(t+1)}$ in the *generalized constrained M-step*.

1) BFGS Method: Once the constrained optimization problem is transformed into a sequence of unconstrained problems using barrier functions, we can use any unconstrained optimization algorithm to solve this sequence of problems. We use the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [42] for multidimensional minimization by replacing the standard backtracking line search algorithm by that suggested by Moré and Thuente [43], [44] which uses the gradient information and guarantees sufficient function decrease at any chosen step size.

2) Barrier Function: Barrier function used to implement the constraints given in (43) is a logarithmic barrier function defined as

$$\Theta(\mathcal{H})=\begin{cases}\log\delta^{-1}+\sum_{\lambda=1}^{K-1}\log\left(1-\rho_\lambda^2\right)^{-1} \\ +\log\det\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*-1} \\ +\log\det\boldsymbol{\Psi}_{\boldsymbol{CC}}^{-1}+\log\det\boldsymbol{\zeta}_0^{-1} \\ +\infty, \end{cases} \begin{array}{l} \text{if Condition}-1 \\ \\ \text{otherwise} \end{array} \tag{68}$$

where Condition-1 is

$$\begin{aligned} \delta &>0 \\ |\rho_\lambda| &\le 1 \\ \text{for } \lambda &=1,\ldots,K-1 \\ \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{*} &\succ 0 \\ \boldsymbol{\Psi}_{\boldsymbol{CC}} &\succ 0 \\ \boldsymbol{\zeta}_0 &\succ 0. \end{aligned} \tag{69}$$

Since the maximization of $\mathfrak{Q}^*(\mathcal{H}|\mathcal{H}^{(t)})$ is the same as the minimization of $-\mathfrak{Q}^*(\mathcal{H}|\mathcal{H}^{(t)})$ (negative of the original objective function), we can write the GCEM algorithm as given in Algorithm IV.1.

Algorithm IV.1: GCEM Algorithm for the Estimation of $\mathcal{H}$ With M-Step Using SUMT

Let $\varepsilon > 0$ be a termination scalar, and choose a point $\mathcal{H}^{(1)} \in \chi$ with $\boldsymbol{g}(\boldsymbol{x}^{(1)}) < \boldsymbol{0}$. Let $\mu^{(1)} > 0$, $\theta \in (0,1)$, let $t=1$, and go to E-step.

E-step Given the data $(\boldsymbol{Y_B},\boldsymbol{Y_C})$ and the parameter estimates $\mathcal{H}^{(t)}$ at iteration t, evaluate the expectation

$$\mathfrak{Q}^*\left(\mathcal{H}|\mathcal{H}^{(t)}\right). \tag{70}$$

Go to M-step.

M-step Given the expectation computed in E-step, minimize using SUMT, the auxiliary function

$$-\mathfrak{Q}^*\left(\mathcal{H}|\mathcal{H}^{(t)}\right)+\mu^{(t)}\Theta(\mathcal{H}) \quad \mu>0 \tag{71}$$

to find some $\mathcal{H}^{(t+1)}$ such that

$$\mathfrak{Q}^*\left(\mathcal{H}^{(t+1)}|\mathcal{H}^{(t)}\right)\ge\mathfrak{Q}^*\left(\mathcal{H}^{(t)}|\mathcal{H}^{(t)}\right). \tag{72}$$

If $\mu^{(t)}\Theta(\mathcal{H}^{(t+1)}) < \varepsilon$, *stop*, achieved convergence. Otherwise, set $\mu^{(t+1)} = \theta\mu^{(t)}$, $t=t+1$ and go to E-step.

C. Final Algorithm for Hyperparameter Estimation

In this section, we summarize the algorithmic steps required to carry out the estimation of the hyperparameters, $\mathcal{H}$. Since the GCEM algorithm converges to a local maximum (and global maximum if only one maximum is present) when the objective function is extremely nonlinear. One way to solve this problem is to perform the optimization by restarting several times from randomly selected initial values for the parameters being estimated and finding the (possibly, local) maximum in each case. Then, the parameter set that gives the maximum among the many estimated local maxima is said to be the *global maximum*. This is detailed in Algorithm IV.2. The next section discusses the performance evaluation of the hyperparameter estimation and the covariance matrix estimation algorithms and discuss the results.

Algorithm IV.2: Overall Algorithm for Hyperparameter Estimation

1: **for** $i = 0$ to $i = MAXRAND$ in increments of 1 **do**
2: $\mathcal{H}^{(1)} \leftarrow$ Random Initialization that is strictly feasible.
3: Initialize $t=1$.
4: Given $\mathcal{H}^{(t)}$, apply Algorithm IV.1 to estimate the result $\mathcal{H}^{(f)}$ via GCEM.
5: Save the function value $\mathfrak{Q}^*(\mathcal{H}^{(f)}|\mathcal{H}^{(t)})$, and the parameters $\mathcal{H}^{(f)}$.
6: Go to step 1.
7: **end for**
8: $\mathcal{H} = \arg\max_{\mathcal{H}^{(f)}} \mathfrak{Q}^*(\mathcal{H}^{(f)}|\mathcal{H}^{(t)})$ is the global maximum of $\mathfrak{Q}^*(\mathcal{H}|\mathcal{H}^{(t)})$.

V. Performance Evaluation of Hyperparameter Estimation

In this section, we discuss the protocol used for performing experiments and validating the estimates computed by the GCEM algorithm discussed in the previous sections. We do this indirectly by performing hypothesis testing of the estimated facet model coefficients which are the functions of the estimated population covariance matrix which is in turn a function of the estimated hyperparameters. To do this, we first need a way of computing the covariance matrix of the facet model coefficients. This is described in the next section.

A. Uncertainty in the Estimated Facet Model Coefficients

In this section, we express the uncertainty in the estimated facet model coefficients in terms of the input perturbation described in terms of the noise covariance matrix $\boldsymbol{\Sigma_\eta}$. From (20), it is clear that $\widehat{\boldsymbol{\alpha}}_n$ is distributed *a posteriori*, with mean $\boldsymbol{\alpha}_n$ (the true unperturbed value) and the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\alpha}}}$. This is illustrated in this section.

Theorem 4 (Mean and Covariance of the Estimate $\widehat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$): Given that an estimate $\widehat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ is given by the (20), the *mean* and *covariance matrix* of the estimate $\widehat{\boldsymbol{\Lambda}}$, respectively, are

$$\boldsymbol{\mu}_{\widehat{\boldsymbol{\Lambda}}}=\boldsymbol{\Lambda}; \quad \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\alpha}}}=\boldsymbol{\zeta}. \tag{73}$$

Proof: For the proof, see Appendix VII. ■

B. Verification Through Hypothesis Testing

The approach we take in testing that the software algorithm is producing the right answers is to test the statistical properties of the answers. In other words, we can statistically test whether the statistical properties of the answers obtained are similar the statistical properties we expect. These expectations involve whether the mean of the computed estimates is sufficiently close to the population mean and whether the estimated covariance matrix of the estimates is sufficiently close to the population covariance matrix or both. In our tests, we test to see if the mean of the estimates of the facet model coefficients is close to the assumed mean of the population from which the facet model coefficients arose. We assume that we know the covariance matrix of the facet model coefficients. Using the closed form expression, given by (73), we can compute the expected covariance matrix of the facet model coefficients using the assumed true noise covariance matrix of the data. In doing this, we perform several *monte carlo* type experiments and test our hypothesis about the outcome. For more details on hypothesis testing one may consult [45] and [46].

1) Hypothesis Testing: We want to test the hypothesis that $H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ called the *null hypothesis* against the *alternative hypothesis*, $H_A : \boldsymbol{\alpha} \neq \boldsymbol{\alpha}_0$ with the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ is assumed known and fixed at $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}$. To test this hypothesis a general method is as follows: a significance level, θ is selected. When the test is run, a test statistic, say ϕ, is computed. This test statistic is typically chosen so that in the case that the hypothesis is true, the test statistic will tend to have its values, distributed around zero, in accordance with a known distribution. If the test statistic has a value, say, higher than a given ϕ_0, we reject the hypothesis that the computed estimated has statistically behaved as we expected. If we do not reject this hypothesis, it is, in effect, accepting the *null hypothesis*. The value ϕ_0 is chosen so that the probability that we reject the hypothesis, given that the hypothesis is true is less than the significance level θ.

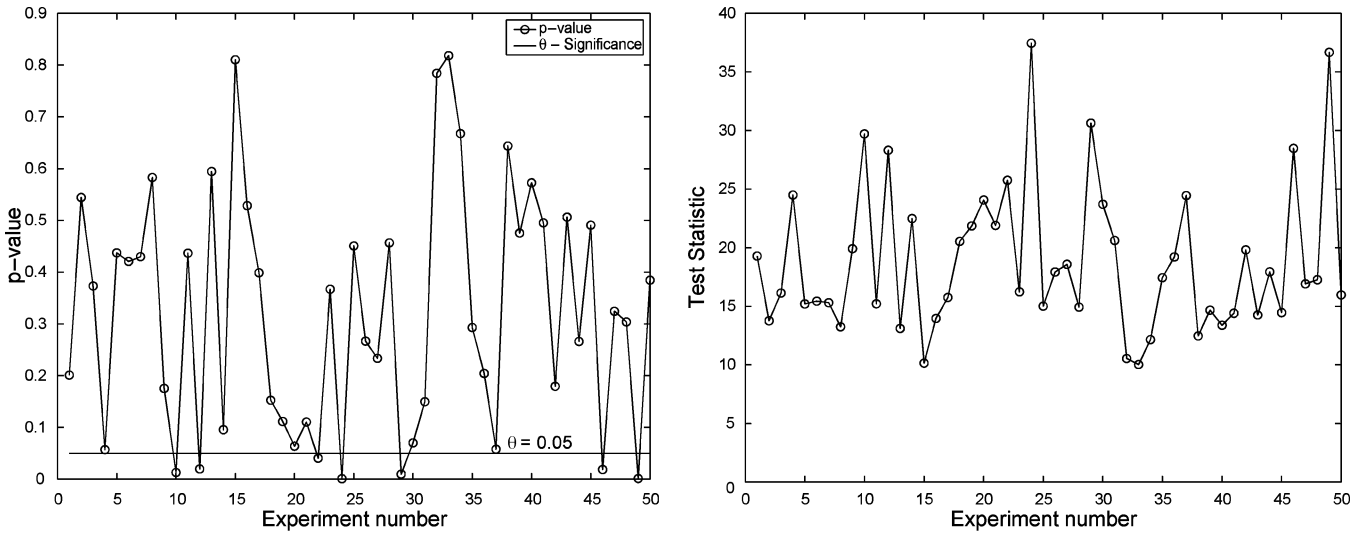
We set up an experiment in which we know what the correct answer α_0 for the no noise ideal case would be. We generate the perturbed data by adding a normally distributed vector from a population having a zero mean and a *given* covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, according to our noise model $\boldsymbol{x}_i = \boldsymbol{B}\boldsymbol{\alpha}_0 + \boldsymbol{\eta}_i$. By using (73), we derive the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}$ of the estimates of $\boldsymbol{\alpha}$ from $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. Using the perturbed data, we run our noise covariance matrix estimation procedure and compute the estimates of $\boldsymbol{\alpha}$. If we repeat this experiment many times by just changing the perturbed realizations and leaving everything else the same, the experiment produces estimates $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ that will come from a normal population having mean $\boldsymbol{\alpha}_0$, the correct answer for the ideal no noise case, and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}$. Now, we want to test the hypothesis that the observations $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ come from a normal distribution with mean $\boldsymbol{\alpha}_0$ with known covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}$.

Define the test statistic, $\phi = N(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}^{-1}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ where $\bar{\boldsymbol{\alpha}} = (1/N)\sum_{i=1}^{N} \widehat{\boldsymbol{\alpha}}_i$. Distribution ϕ under *null hypothesis* is Chi-squared [45] given by $\phi \sim \chi_M^2$ where M is the dimension of the vector $\boldsymbol{\alpha}_i$. The distribution under the alternative hypothesis is a noncentral Chi-squared and is given by $\phi \sim \chi_{M,d}^2$ where $d = N(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)'\boldsymbol{\Sigma}_{\boldsymbol{\alpha}1}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)$ is the noncentrality parameter.

2) Experiments: We performed the experiment several times for different known values of $\boldsymbol{\alpha}_0$, and different values of noise covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. We chose the significance level $\theta = 0.05$. We choose an $\boldsymbol{\alpha}_0$ and a $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and we generated $N = 2000$ perturbed realizations $\boldsymbol{x}_i$ for each run of the algorithm. We performed 250 such runs with this combination of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. We then repeated the experiments for different combinations of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and at each run of the algorithm we performed the hypothesis testing. For noise estimation we chose a neighborhood size of 5×5, resulting in an observation vector size $K = 25$. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is of size 25×25. The DOP basis in this neighborhood has a total of 25 coefficients. We use a fourth order DOP to represent the signal space. That is, the first 15 coefficients are used to represent the signal space. These 15 coefficients constitute the vector $\boldsymbol{\alpha}$. The remaining ten coefficients are used to represent the noise space. We do not write down the true noise covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ here, due to space constraints. Known noise covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is obtained as follows: first, a hyperparameter vector is generated at random, so that each component is uniformly distributed. For the test described here, we generated δ to be uniformly distributed in [0,1000] and let it be denoted by δ_1. The remaining 25 parameters (correlation filter coefficients) are generated so that each component is uniformly distributed in $[-10,10]$. Given the hyperparameter vector, we create the hypercovariance matrix $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^*$ which is premultiplied by $(\boldsymbol{B}\ \boldsymbol{C})$ and post-multiplied by its transpose to get $\boldsymbol{\Psi}_{\boldsymbol{\eta}}$. We derive a random sample from an $\mathcal{IW}$ distribution with number of degrees of freedom parameter δ_1 and scale matrix $\boldsymbol{\Psi}_{\boldsymbol{\eta}}$. This random sample constitutes our true covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. This is valid because the $\mathcal{GIW}$ distribution in the form that we used exactly corresponds to $\mathcal{IW}$ distribution. The known mean vector $\boldsymbol{\alpha}_0$ is chosen at random so that each of its components is sampled uniformly from the interval $[-100\ 100]$. We generate noise sample vectors $\boldsymbol{\eta}_i$ from a Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and add it to $\boldsymbol{B}\boldsymbol{\alpha}_0$ to generate perturbed input vectors $\boldsymbol{x}_i$. Results from the first 50 sample experiments are discussed here for brevity. Fig. 2(a) shows the p-values plotted against the experiment number. The horizontal line shows the significance level $\theta = 0.05$ chosen for the experiment. All the points that fell below this line indicate a failure of the test. Fig. 2(b) shows the test statistic ϕ plotted against the experiment number.

VI. Applications

As an application of the noise covariance estimation procedure, we designed a new ridge operator that is based on the facet model. The ridge operator uses the integral of the second directional derivative of the facet model in estimating the optimal ridge direction. Once the ridge direction is found it is easy to label a pixel to be a ridge or not, using the first and second directional derivatives. The ridge operator is abbreviated as ISDDRO. The ISDDRO operator is designed to work both under additive colored noise as well as white noise assumptions. Under colored noise assumption, we used the noise covariance matrix estimation procedure discussed in this paper to estimate the facet model coefficients which are then used in ridge detection.



(a) (b)

Fig. 2. Hypothesis testing. (a) Plot of p-value. (b) Plot of test statistic, ϕ.

The results of this application are reported in other papers [32], [47]. We evaluated the ridge detectors based on the *ridge orientation estimate*. The *mean orientation bias* which is the difference between the sample mean of the estimates and true orientation, and the *orientation standard deviation* which is the deviation of the estimates from their sample mean are chosen as the objective measures of performance. To summarize the results from the said references, using the chosen performance measures, ISDDRO operator that uses the colored noise assumption (ISDDRO-CN) displays superior noise sensitivity characteristics as compared to the same operator under white noise assumption (ISDDRO-WN) and both versions out performed the most competing operator MLSEC [48]. In fact, ISDDRO under colored noise assumption performs very well under high noise conditions where the other two operators failed.

VII. Conclusion

In this paper, we derived the theory and formulated the problem of noise covariance estimation as problem of minimizing an *objective* or *criterion* function in a Bayesian framework. The problem formulation was such that we decomposed the noise covariance matrix into its components in the space spanned by the columns of the signal space basis matrix $\boldsymbol{B}$ and its complement space spanned by the columns of the matrix $\boldsymbol{C}$ orthonormal to $\boldsymbol{B}$. This facilitated theorizing and proving that the same result for $\boldsymbol{\alpha}_n$ as in the case of white noise is obtained when the signal space and its complement space have no correlation between them, which implied that all energy of the noise in the space spanned by the columns of $\boldsymbol{B}$ goes into the computed $\boldsymbol{\alpha}_n$, but the noise in the space spanned by the columns of $\boldsymbol{C}$ will only influence the estimate of $\boldsymbol{\alpha}_n$ to the extent that there is coupling between the signal space and its complement space. We also derived an expression for $\boldsymbol{\alpha}_n$ for the general case of correlation between the two spaces. To reduce the dimensionality and to ease the optimization we introduced a correlation model for the two spaces that was necessary for the estimation of the population covariance matrix and derived the constraints which the hyperparameters must obey. We also derived a GCEM algorithm for the estimation of the hyperparameters. We put forward a new way of implementing the constrained M-step of the GCEM algorithm using SUMT. We evaluated the performance of our noise covariance matrix estimation algorithm using statistical hypothesis testing and concluded that the algorithm performs as expected. We also demonstrated, by designing a new ridge detector, that we can design feature extractors with better noise sensitivity characteristics if we assume additive colored noise perturbation as compared to the additive white noise assumption.

Appendix I
Matrix Normal Distribution

In this section, we give only the major results about matrix normal distribution. Further details can be found in [33], [45].

Proposition 1 (Matrix Normal Distribution): A $M \times N$ random matrix $\boldsymbol{X}$ is said to have the matrix normal distribution $\mathcal{N}(\boldsymbol{X}_0, \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon})$, if it has the joint density

$$p(\boldsymbol{X}) = \left(\frac{1}{(2\pi)^{\frac{NM}{2}} \det \boldsymbol{\Lambda}^{\frac{N}{2}} \det \boldsymbol{\Upsilon}^{\frac{M}{2}}} \right) \times \operatorname{etr}\left\{ -\frac{1}{2} \left[\boldsymbol{\Lambda}^{-1} (\boldsymbol{X} - \boldsymbol{X}_0) \right] \left[(\boldsymbol{X} - \boldsymbol{X}_0) \boldsymbol{\Upsilon}^{-1} \right]' \right\} \quad (74)$$

$\boldsymbol{\Lambda}$ is the covariance matrix within each column, $\boldsymbol{\Upsilon}$ is the covariance matrix between columns and $\boldsymbol{X}_0$ is the mean matrix.

Proposition 2 (Matrix Normal Distrib., With i.i.d. Columns): Let the normal random matrix $\boldsymbol{X}$ be given in its component form as, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$, where $\boldsymbol{x}_i$ is a column vector. Let a_{ij} be an element of $\boldsymbol{\Lambda} \in \mathbb{R}^{M \times M}$ and b_{ij} be an element of $\boldsymbol{\Upsilon} \in \mathbb{R}^{N \times N}$. If $\boldsymbol{x}_i$ are independent, then $\boldsymbol{\Upsilon} = \boldsymbol{I}_N$, where $I_N \in \mathbb{R}^{N \times N}$ is an identity matrix. Therefore, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}_0, \boldsymbol{\Lambda} \otimes \boldsymbol{I}_N)$ and the density function is given by

$$p(\boldsymbol{X}) = \left(\frac{1}{(2\pi)^{\frac{NM}{2}} \det \boldsymbol{\Lambda}^{\frac{N}{2}}} \right) \times \operatorname{etr}\left\{ -\frac{1}{2} \left[\boldsymbol{\Lambda}^{-1} (\boldsymbol{X} - \boldsymbol{X}_0) \right] [\boldsymbol{X} - \boldsymbol{X}_0]' \right\} \quad (75)$$

which is equivalent to

$$p(\boldsymbol{X}) = \left(\frac{1}{(2\pi)^{\frac{NM}{2}} \det \boldsymbol{\Lambda}^{\frac{N}{2}}}\right) \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{x}_{0i})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_{0i})\right\} \quad (76)$$

where $\boldsymbol{x}_{0i}$ are the column vectors of the mean matrix $\boldsymbol{X}_0$.

Proposition 3 (Properties): The proposition states some properties of matrix normal distribution.

- $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}_0, \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon})$ iff $\boldsymbol{X}' \sim \mathcal{N}(\boldsymbol{X}_0', \boldsymbol{\Upsilon} \otimes \boldsymbol{\Lambda})$.
- $\boldsymbol{E}[\boldsymbol{X}] = \boldsymbol{X}_0$, $\mathrm{var}[\mathrm{vec}(\boldsymbol{X})] = \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon}$, and $\mathrm{var}[\mathrm{vec}(\boldsymbol{X}')] = \boldsymbol{\Upsilon} \otimes \boldsymbol{\Lambda}$.
- $\mathrm{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = a_{ij}\boldsymbol{\Upsilon}$, and $\mathrm{cov}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = b_{ij}\boldsymbol{\Lambda}$.
- For any matrices $\boldsymbol{C} \in \mathbb{R}^{K \times M}$ and $\boldsymbol{D} \in \mathbb{R}^{N \times J}$, $\boldsymbol{CXD} = \mathcal{N}(\boldsymbol{CX}_0\boldsymbol{D}, \boldsymbol{C\Lambda C}' \otimes \boldsymbol{D}'\boldsymbol{\Upsilon D})$.

where $\mathrm{var}(\boldsymbol{x})$ is the variance of $\boldsymbol{x}$, and $\mathrm{vec}(\boldsymbol{X})$ is the vector resulting by stacking up one column of the matrix $\boldsymbol{X}$ on top of another.

Appendix II
Wishart and Inverted Wishart Distributions

In this section, we present a summary of results on *Wishart* and *IW* distributions. Details can be found in [33], [36], [45], and the classic paper by J. Wishart [49]–[51].

Proposition 4 (Wishart Distribution): A $K \times K$ random matrix $\boldsymbol{\Sigma}$ has a *Wishart distribution* $\boldsymbol{\Sigma} \sim \mathcal{W}(\nu, \boldsymbol{\Psi})$, if its joint density has the form

$$p(\boldsymbol{\Sigma}) = \frac{1}{2^{\frac{\nu K}{2}}\boldsymbol{\Gamma}_K\left(\frac{\nu}{2}\right)} \det \boldsymbol{\Psi}^{-\frac{\nu}{2}} \det \boldsymbol{\Sigma}^{\frac{\nu-K-1}{2}} \times \mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\right\} \quad (77)$$

where ν is the number of degrees of freedom, $\boldsymbol{\Gamma}_K(t)$ is a multidimensional Gamma function given by, $\boldsymbol{\Gamma}_K(t) = \pi^{K(K-1)/4}\prod_{i=1}^{K}\Gamma[t - ((i-1)/2)]$.

Proposition 5 ($\mathcal{IW}$ Distribution): A $K \times K$ matrix $\boldsymbol{\Sigma}$ has an *$\mathcal{IW}$ distribution* $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$, if it has the density function of the following form:

$$p(\boldsymbol{\Sigma}) = \frac{1}{2^{\frac{(\delta+K-1)K}{2}}\boldsymbol{\Gamma}_K\left(\frac{\delta+K-1}{2}\right)} \det \boldsymbol{\Psi}^{\frac{\delta+K-1}{2}} \times \det \boldsymbol{\Sigma}^{-\left(\frac{\delta+2K}{2}\right)} \mathrm{etr}\left\{-\frac{1}{2}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\} \quad (78)$$

where δ is the parameter of the distribution which is chosen because it is invariant under change of dimension [11], [33], and $\nu = \delta + K - 1$ and $\boldsymbol{\Gamma}_K(t)$ is a multidimensional gamma function as defined in Proposition 4.

Proposition 6 (Properties): Some properties of the *Wishart* and $\mathcal{IW}$ distributions are given below.

- $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$ iff $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{\Psi}^{-1})$.
- If $\boldsymbol{\Sigma} \sim \mathcal{W}(\nu, \boldsymbol{\Psi})$, then $\boldsymbol{E}[\boldsymbol{\Sigma}] = \nu\boldsymbol{\Psi}$, and $\boldsymbol{E}[\boldsymbol{\Sigma}^{-1}] = (1/(\delta - 2))\boldsymbol{\Psi}^{-1}$ provided $(\delta - 2) > 0$.
- If $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$, then $\boldsymbol{E}[\boldsymbol{\Sigma}] = (1/(\delta - 2))\boldsymbol{\Psi}$, and $\boldsymbol{E}[\boldsymbol{\Sigma}^{-1}] = \nu\boldsymbol{\Psi}^{-1}$.
- If $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$, then, $\boldsymbol{E}[\log\det\boldsymbol{\Sigma}] = -K\log 2 - \sum_{i=1}^{K}\mathcal{D}[((\delta + K - i)/2)] + \log\det\boldsymbol{\Psi}$, where $\mathcal{D}(x) = (d/dx)\log\boldsymbol{\Gamma}(x)$ is the *digamma* function.

The following proposition states that the $\mathcal{IW}$ distribution is closed under linear transformations.

Proposition 7 (Linear Transformation of $\mathcal{IW}$): Let $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$. Let $\boldsymbol{T}$ be a given transformation matrix. Then, $\boldsymbol{T}'\boldsymbol{\Sigma T} \sim \mathcal{IW}(\delta, \boldsymbol{T}'\boldsymbol{\Psi T})$.

Theorem 5 (Orthonormal Transformation of $p(\boldsymbol{\Sigma})$): Let $p(\boldsymbol{\Sigma})$ denote a matrix variate distribution. Let $\boldsymbol{T}$ be an orthonormal transformation matrix. Then, $p(\boldsymbol{\Sigma}) = p(\boldsymbol{T}'\boldsymbol{\Sigma T})$.

Proof: The proof of this theorem is self evident as the determinant of the Jacobian matrix $\det \boldsymbol{J} = \det \boldsymbol{T}'\boldsymbol{T} = 1$, since $\boldsymbol{T}$ is an orthonormal matrix with unity determinant. ∎

Appendix III
Generalized Inverted Wishart Distribution

In this section, we discuss the $\mathcal{GIW}$ distribution introduced by Brown *et al.* [19].

Let $\boldsymbol{X} \in \mathbb{R}^{K \times N}$ be a data matrix containing N, K-dimensional independent observation vectors arranged in columns. Then, according to Section I, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{X}_0, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_N)$.

Let the matrix $\boldsymbol{\Sigma}$ be distributed as $\mathcal{IW}(\delta, \boldsymbol{\Psi})$ where δ is as defined in Section II, and $\boldsymbol{\Psi}$ is the scale matrix.

Let $\boldsymbol{\Sigma}$ be partitioned naturally as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \quad (79)$$

Conformably, let the data matrix $\boldsymbol{X}$ be partitioned as $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$. Let $\boldsymbol{\Psi}$ be partitioned conformably as

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix}. \quad (80)$$

Using Bartlett [34] decomposition, we have

$$\boldsymbol{\Sigma} = \boldsymbol{T\Omega T}' \quad (81)$$

where $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\zeta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ and $\boldsymbol{T} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\tau}' \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}$. Hence

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\zeta} + \boldsymbol{\tau}'\boldsymbol{\Sigma}_{22}\boldsymbol{\tau} & \boldsymbol{\tau}'\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{22}\boldsymbol{\tau} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad (82)$$

where $\boldsymbol{\Sigma}_{22} : Q \times Q$, $\boldsymbol{\zeta} : M \times M$ and $\boldsymbol{\tau} : Q \times M$ and $\boldsymbol{\tau} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$; $\boldsymbol{\zeta} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$; $\boldsymbol{\tau}_0 = \boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21}$; $\boldsymbol{\zeta}_0 = \boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21}$; $K = M + Q$.

Proposition 8 ($\mathcal{GIW}$ Distribution): Let $\boldsymbol{\Sigma} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi})$. If $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ have the conformable partitioning given in (79) and (80), respectively, then in terms of its block partitions $\boldsymbol{\Sigma}$ has a $\mathcal{GIW}$ *model* [19] given as

1) $\boldsymbol{\Sigma}_{22} \sim \mathcal{IW}(\delta, \boldsymbol{\Psi}_{22})$, is distributed independently of $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$;
2) $\boldsymbol{\zeta} \sim \mathcal{IW}(\delta + Q, \boldsymbol{\zeta}_0)$;
3) $\boldsymbol{\tau}|\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\tau}_0, \boldsymbol{\Psi}_{22}^{-1} \otimes \boldsymbol{\zeta})$.

It is clear that $\boldsymbol{\tau}$ is the *slope* of the best linear predictor of the observables $\boldsymbol{X}_1$ based on the observables $\boldsymbol{X}_2$ and $\boldsymbol{\zeta}$ is the residual covariance matrix of the resulting prediction errors.

Under the above assumptions $\boldsymbol{\Sigma}$ is said to have the $\mathcal{GIW}$ distribution with parameter δ and the scale matrix $\boldsymbol{\Psi}$. This is denoted as $\boldsymbol{\Sigma} \sim \mathcal{GIW}(\delta, \boldsymbol{\Psi})$. Under this notation, it is assumed that $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ have the aforementioned partitioned structure. The $\mathcal{GIW}$ in the above form exactly corresponds to the standard $\mathcal{IW}(\delta, \boldsymbol{\Psi})$ on the unpartitioned $\boldsymbol{\Sigma}$. We still use the term $\mathcal{GIW}$ to refer to this special case of $\mathcal{GIW}$ just to differentiate this hierarchical structure from the standard $\mathcal{IW}$.

Appendix IV
Optimization of Scalar Functions of Matrices

In this section, we state and prove [45], the following theorem, which we use in this paper.

Theorem 6 (Maximum of a Scalar Function of a Matrix): If $\boldsymbol{D} \in \mathbb{R}^{M \times M}$ is positive definite and C is a constant, then the *maximum* of

$$f(\boldsymbol{G}) = -C \log\det \boldsymbol{G} - \mathrm{tr}(\boldsymbol{G}^{-1}\boldsymbol{D}) \tag{83}$$

with respect to positive definite matrices $\boldsymbol{G}$ exists, and occurs at

$$\boldsymbol{G} = \frac{1}{C}\boldsymbol{D} \tag{84}$$

and has the value of $f((1/C)\boldsymbol{D}) = MC \log C - C \log\det \boldsymbol{D} - MC$.

Proof: Let $\boldsymbol{D} = \boldsymbol{E}\boldsymbol{E}'$ and $\boldsymbol{E}'\boldsymbol{G}^{-1}\boldsymbol{E} = \boldsymbol{H}$. Then, $\boldsymbol{G} = \boldsymbol{E}\boldsymbol{H}^{-1}\boldsymbol{E}'$, and $\det \boldsymbol{G} = \det \boldsymbol{E} \det \boldsymbol{H}^{-1} \det \boldsymbol{E}' = \det \boldsymbol{H}^{-1} \det(\boldsymbol{E}\boldsymbol{E}') = (\det \boldsymbol{D}/\det \boldsymbol{H})$, and $\mathrm{tr}(\boldsymbol{G}^{-1}\boldsymbol{D}) = \mathrm{tr}(\boldsymbol{G}^{-1}\boldsymbol{E}\boldsymbol{E}') = \mathrm{tr}(\boldsymbol{E}'\boldsymbol{G}^{-1}\boldsymbol{E}) = \mathrm{tr}\boldsymbol{H}$. Then, the function to be maximized (with respect to the positive definite $\boldsymbol{H}$) is

$$f = -C \log\det \boldsymbol{D} + C \log\det \boldsymbol{H} - \mathrm{tr}\boldsymbol{H}. \tag{85}$$

Let $\boldsymbol{H} = \boldsymbol{T}\boldsymbol{T}'$, where $\boldsymbol{T}$ is lower triangular. Then, the maximum of

$$\begin{aligned} f &= -C \log\det \boldsymbol{D} + C \log(\det \boldsymbol{T})^2 - \mathrm{tr}(\boldsymbol{T}\boldsymbol{T}') \\ &= -C \log\det \boldsymbol{D} + \sum_{i=1}^{M} \left(C \log t_{ii}^2 - t_{ii}^2\right) - \sum_{i>j} t_{ij}^2 \end{aligned} \tag{86}$$

occurs at $t_{ii}^2 = C$, $t_{ij} = 0$, $i \neq j$, that is, at $\boldsymbol{H} = C\boldsymbol{I}$. Then, $\boldsymbol{G} = (1/C)\boldsymbol{E}\boldsymbol{E}' = (1/C)\boldsymbol{D}$. ∎

Appendix V
Conditional Density of $\boldsymbol{Y_B}$ Given $\boldsymbol{Y_C}$

In this section, we derive the conditional density of $\boldsymbol{Y_B}$ given $\boldsymbol{Y_C}$. This is necessary for computing the joint posterior density of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$, and $\boldsymbol{\zeta}$ given the data $(\boldsymbol{Y}_{\boldsymbol{B}}', \boldsymbol{Y}_{\boldsymbol{C}}')'$.

Theorem 7 (Conditional Density of $\boldsymbol{Y_B}$ Given $\boldsymbol{Y_C}$): If

$$\begin{aligned} \boldsymbol{Y} &= \begin{pmatrix} \boldsymbol{Y_B} \\ \boldsymbol{Y_C} \end{pmatrix} \\ &\sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_{BB}} & \boldsymbol{\Sigma_{BC}} \\ \boldsymbol{\Sigma_{CB}} & \boldsymbol{\Sigma_{CC}} \end{pmatrix} \otimes \boldsymbol{I}_N \right) \end{aligned} \tag{87}$$

and

$$\boldsymbol{Y_B} \sim \mathcal{N}(\boldsymbol{\Lambda}, \boldsymbol{\Sigma_{BB}} \otimes \boldsymbol{I}_N) \tag{88}$$

$$\boldsymbol{Y_C} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_{CC}} \otimes \boldsymbol{I}_N) \tag{89}$$

then

$$\boldsymbol{Y_B} | \boldsymbol{Y_C} \sim \mathcal{N}(\boldsymbol{\Lambda} + \boldsymbol{\tau}'\boldsymbol{Y_C}, \boldsymbol{\zeta} \otimes \boldsymbol{I}_N) \tag{90}$$

where $\boldsymbol{\zeta} = \boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}$ and $\boldsymbol{\tau} = \boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}$.

Proof: Let us define the following linear transformations of the observations:

$$\boldsymbol{Y_B^\dagger} = \boldsymbol{Y_B} + \boldsymbol{A}\boldsymbol{Y_C} \tag{91}$$

$$\boldsymbol{Y_C^\dagger} = \boldsymbol{Y_C}. \tag{92}$$

We need to choose the transformation matrix $\boldsymbol{A}$ such that $\boldsymbol{Y_B^\dagger}$ and $\boldsymbol{Y_C^\dagger}$ are uncorrelated. This condition means that the cross-correlation must be zero, i.e., $\boldsymbol{E}[(\boldsymbol{Y_B^\dagger} - \boldsymbol{E}[\boldsymbol{Y_B^\dagger}])(\boldsymbol{Y_C^\dagger} - \boldsymbol{E}[\boldsymbol{Y_C^\dagger}])'] = 0$, where $\boldsymbol{E}[\boldsymbol{Y_B^\dagger}] = \boldsymbol{\Lambda}$ and $\boldsymbol{E}[\boldsymbol{Y_C^\dagger}] = \boldsymbol{0}$. Substituting for $\boldsymbol{Y_B^\dagger}$ and $\boldsymbol{Y_C^\dagger}$ from (92) and (91), and their expectations just computed, we get $\boldsymbol{E}[(\boldsymbol{Y_B} + \boldsymbol{A}\boldsymbol{Y_C} - \boldsymbol{\Lambda})\boldsymbol{Y_C}'] = \boldsymbol{0}$ and $\boldsymbol{E}[(\boldsymbol{Y_B} - \boldsymbol{\Lambda})\boldsymbol{Y_C}'] + \boldsymbol{A}\boldsymbol{E}[\boldsymbol{Y_C}\boldsymbol{Y_C}'] = \boldsymbol{0}$. Now, using the fact that $\boldsymbol{E}[(\boldsymbol{Y_B} - \boldsymbol{\Lambda})\boldsymbol{Y_C}'] = \boldsymbol{\Sigma_{BC}}$ and $\boldsymbol{E}[\boldsymbol{Y_C}\boldsymbol{Y_C}'] = \boldsymbol{\Sigma_{CC}}$, we get $\boldsymbol{\Sigma_{BC}} + \boldsymbol{A}\boldsymbol{\Sigma_{CC}} = \boldsymbol{0}$ and $\boldsymbol{A} = -\boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}$. Now, (92) and (91) can be written in matrix form as $\boldsymbol{Y}^\dagger = \begin{pmatrix} \boldsymbol{Y_B^\dagger} \\ \boldsymbol{Y_C^\dagger} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y_B} \\ \boldsymbol{Y_C} \end{pmatrix}$. Taking expectations of this result, we get

$$\boldsymbol{E}[\boldsymbol{Y}^\dagger] = \begin{pmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{0} \end{pmatrix}. \tag{93}$$

Covariance matrix $\boldsymbol{\Sigma_{Y^\dagger}}$ of $\boldsymbol{Y}^\dagger$ is given by

$$\begin{aligned} \boldsymbol{\Sigma_{Y^\dagger}} &= \boldsymbol{E}\left[\left(\boldsymbol{Y}^\dagger - \boldsymbol{E}[\boldsymbol{Y}^\dagger]\right) \left(\boldsymbol{Y}^\dagger - \boldsymbol{E}[\boldsymbol{Y}^\dagger]\right)' \right] \\ &= \begin{pmatrix} \boldsymbol{E}\left[\left(\boldsymbol{Y_B^\dagger} - \boldsymbol{E}\left[\boldsymbol{Y_B^\dagger}\right]\right) \left(\boldsymbol{Y_B^\dagger} - \boldsymbol{E}\left(\boldsymbol{Y_B^\dagger}\right)\right)' \right] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_{CC}} \end{pmatrix}. \end{aligned} \tag{94}$$

We used the fact that $\boldsymbol{Y_B^\dagger}$ and $\boldsymbol{Y_C^\dagger}$ are uncorrelated. Consider the term

$$\begin{aligned} &\boldsymbol{E}\left[\left(\boldsymbol{Y_B^\dagger} - \boldsymbol{E}\left[\boldsymbol{Y_B^\dagger}\right]\right) \left(\boldsymbol{Y_B^\dagger} - \boldsymbol{E}\left[\boldsymbol{Y_B^\dagger}\right]\right)' \right] \\ &\quad = \boldsymbol{E}\left[\left(\boldsymbol{Y_B} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C} - \boldsymbol{\Lambda}\right) \right. \\ &\qquad \left. \times \left(\boldsymbol{Y_B} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C} - \boldsymbol{\Lambda}\right)' \right] \\ &\quad = \boldsymbol{E}\left[(\boldsymbol{Y_B} - \boldsymbol{\Lambda})(\boldsymbol{Y_B} - \boldsymbol{\Lambda})'\right] \\ &\qquad - \boldsymbol{E}\left[(\boldsymbol{Y_B} - \boldsymbol{\Lambda})\boldsymbol{Y_C}'\right] \boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}} \\ &\qquad - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{E}\left[\boldsymbol{Y_C}(\boldsymbol{Y_B} - \boldsymbol{\Lambda})'\right] \\ &\qquad + \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{E}\left[\boldsymbol{Y_C}\boldsymbol{Y_C}'\right] \boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}} \\ &\quad = \boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}. \end{aligned} \tag{95}$$

Therefore

$$\boldsymbol{\Sigma_{Y^\dagger}} = \begin{pmatrix} \boldsymbol{\zeta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_{CC}} \end{pmatrix} \tag{96}$$

where $\boldsymbol{\zeta} = \boldsymbol{\Sigma_{BB}} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}$. Since uncorrelated normally distributed random variables are also *independent*, the conditional density of $\boldsymbol{Y_B^\dagger}$ given $\boldsymbol{Y_C}$ (note that $\boldsymbol{Y_C^\dagger} = \boldsymbol{Y_C}$) is same as the marginal density of $\boldsymbol{Y_B^\dagger}$. Therefore, $\boldsymbol{Y_B^\dagger}|\boldsymbol{Y_C} \sim \mathcal{N}(\boldsymbol{\Lambda}, \boldsymbol{\zeta})$. Given $\boldsymbol{Y_C}$, we have

$$\boldsymbol{Y_B} = \boldsymbol{Y_B^\dagger} + \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C}. \tag{97}$$

Note that $\boldsymbol{Y_C}$ is a constant in the conditional density of $\boldsymbol{Y_B}|\boldsymbol{Y_C}$. We have from (97) that

$$\begin{aligned}\boldsymbol{E}[\boldsymbol{Y_B}|\boldsymbol{Y_C}] &= \boldsymbol{E}\left[\boldsymbol{Y_B^\dagger}\right] + \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C} \\ &= \boldsymbol{\Lambda} + \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C}.\end{aligned} \tag{98}$$

The covariance matrix of the conditional density $\boldsymbol{Y_B}|\boldsymbol{Y_C}$ is

$$\boldsymbol{\Sigma_{Y_B|Y_C}} = \boldsymbol{E}\left[(\boldsymbol{Y_B} - \boldsymbol{E}[\boldsymbol{Y_B}|\boldsymbol{Y_C}])(\boldsymbol{Y_B} - \boldsymbol{E}[\boldsymbol{Y_B}|\boldsymbol{Y_C}])'\right]. \tag{99}$$

Substituting for $\boldsymbol{Y_B}$ and $\boldsymbol{E}[\boldsymbol{Y_B}|\boldsymbol{Y_C}]$ from (97) and (98), respectively, and using (93), (96), and (95) and simplifying, we get, $\boldsymbol{\Sigma_{Y_B|Y_C}} = \boldsymbol{\zeta}$. Therefore, we can write

$$\boldsymbol{Y_B}|\boldsymbol{Y_C} \sim \mathcal{N}\left(\boldsymbol{\Lambda} + \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C}, \boldsymbol{\zeta} \otimes \boldsymbol{I}_N\right) \tag{100}$$

where $\boldsymbol{\tau} = \boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{\Sigma_{CB}}$. ■

Appendix VI
Proof of Theorem 3

Proof: Joint density of $(\boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta})$ and the data $(\boldsymbol{Y_B}, \boldsymbol{Y_C})$ is given by

$$\begin{aligned}&p(\boldsymbol{Y_B}, \boldsymbol{Y_C}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta}, \delta, \boldsymbol{\Psi_{CC}}, \boldsymbol{\tau}_0, \boldsymbol{\zeta}_0) \\ &\quad = p(\boldsymbol{Y_B}|\boldsymbol{Y_C}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta}, \delta, \boldsymbol{\Psi_{CC}}, \boldsymbol{\tau}_0, \boldsymbol{\zeta}_0) \\ &\qquad \times p(\boldsymbol{Y_C}|\boldsymbol{\Lambda}, \boldsymbol{\Sigma_{CC}}, \boldsymbol{\tau}, \boldsymbol{\zeta}, \delta, \boldsymbol{\Psi_{CC}}, \boldsymbol{\tau}_0, \boldsymbol{\zeta}_0) \\ &\qquad \times p(\boldsymbol{\Lambda}) p(\boldsymbol{\Sigma_{CC}}|\delta, \boldsymbol{\Psi_{CC}}) \\ &\qquad \times p(\boldsymbol{\tau}|\boldsymbol{\zeta}, \delta, \boldsymbol{\Psi_{CC}}, \boldsymbol{\tau}_0, \boldsymbol{\zeta}_0) p(\boldsymbol{\zeta}|\delta, \boldsymbol{\zeta}_0) \\ &\quad = \left(\frac{1}{(2\pi)^{\frac{NM}{2}} \det \boldsymbol{\zeta}^{\frac{N}{2}}}\right) \\ &\qquad \times \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\zeta}^{-1}(\boldsymbol{Y_B} - \boldsymbol{\Lambda} - \boldsymbol{\tau}'\boldsymbol{Y_C}) \right. \\ &\qquad\qquad \left. \times (\boldsymbol{Y_B} - \boldsymbol{\Lambda} - \boldsymbol{\tau}'\boldsymbol{Y_C})'\right\} \\ &\qquad \times \frac{1}{(2\pi)^{\frac{NQ}{2}} \det \boldsymbol{\Sigma_{CC}}^{\frac{N}{2}}} \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{Y_C}\boldsymbol{Y_C'}\right\} \\ &\qquad \times \frac{\det \boldsymbol{\Psi_{CC}}^{\frac{\delta+Q-1}{2}} \det \boldsymbol{\Sigma_{CC}}^{-\frac{\delta+2Q}{2}}}{2^{\frac{(\delta+Q-1)Q}{2}} \boldsymbol{\Gamma}_Q\left(\frac{\delta+Q-1}{2}\right)} \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\Psi_{CC}}\boldsymbol{\Sigma_{CC}^{-1}}\right\} \\ &\qquad \times \frac{\det \boldsymbol{\zeta_0}^{\frac{\delta+Q+M-1}{2}} \det \zeta^{-\frac{\delta+Q+2M}{2}}}{2^{\frac{(\delta+Q+M-1)M}{2}} \boldsymbol{\Gamma}_M\left(\frac{\delta+Q+M-1}{2}\right)} \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\zeta}_0\boldsymbol{\zeta}^{-1}\right\} \\ &\qquad \times \left(\frac{1}{(2\pi)^{\frac{MQ}{2}} \det \boldsymbol{\Psi_{CC}}^{\frac{M}{2}} \det \boldsymbol{\zeta}^{\frac{Q}{2}}}\right) \\ &\qquad \times \operatorname{etr}\left\{-\frac{1}{2}\left[\boldsymbol{\Psi_{CC}}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\right]\left[(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\boldsymbol{\zeta}^{-1}\right]'\right\}.\end{aligned} \tag{101}$$

Now, substituting $\boldsymbol{\Lambda} = \widehat{\boldsymbol{\Lambda}}$ the first exponential in above equation becomes unity. Now, multiplying and dividing the (101) by

$$\frac{\det\left(\boldsymbol{\Psi_{CC}} + \boldsymbol{Y_C}\boldsymbol{Y_C'}\right)^{\frac{\delta+N+Q-1}{2}} \det \boldsymbol{\zeta_0}^{\frac{N}{2}}}{\boldsymbol{\Gamma}_Q\left(\frac{\delta+N+Q-1}{2}\right)\boldsymbol{\Gamma}_M\left(\frac{\delta+Q+N+M-1}{2}\right)} \tag{102}$$

and rearranging, we get

$$\begin{aligned}&= \left(\frac{\det\left(\boldsymbol{\Psi_{CC}} + \boldsymbol{Y_C}\boldsymbol{Y_C'}\right)^{\frac{\delta+N+Q-1}{2}} \det \boldsymbol{\Sigma_{CC}}^{-\frac{\delta+N+2Q}{2}}}{2^{\frac{(\delta+N+Q-1)Q}{2}} \boldsymbol{\Gamma}_Q\left(\frac{\delta+N+Q-1}{2}\right)}\right) \\ &\quad \times \operatorname{etr}\left\{-\frac{1}{2}\left(\boldsymbol{\Psi_{CC}} + \boldsymbol{Y_C}\boldsymbol{Y_C'}\right)\boldsymbol{\Sigma_{CC}^{-1}}\right\} \\ &\quad \times \left(\frac{\det \boldsymbol{\zeta_0}^{\frac{\delta+Q+N+M-1}{2}} \det \boldsymbol{\zeta}^{-\frac{\delta+Q+N+2M}{2}}}{2^{\frac{(\delta+Q+N+M-1)M}{2}} \boldsymbol{\Gamma}_M\left(\frac{\delta+Q+N+M-1}{2}\right)}\right) \\ &\quad \times \operatorname{etr}\left\{-\frac{1}{2}\boldsymbol{\zeta}_0\boldsymbol{\zeta}^{-1}\right\} \times \left(\frac{1}{(2\pi)^{\frac{MQ}{2}} \det \boldsymbol{\Psi_{CC}}^{\frac{M}{2}} \det \boldsymbol{\zeta}^{\frac{Q}{2}}}\right) \\ &\quad \times \operatorname{etr}\left\{-\frac{1}{2}\left[\boldsymbol{\Psi_{CC}}(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\right]\left[(\boldsymbol{\tau} - \boldsymbol{\tau}_0)\boldsymbol{\zeta}^{-1}\right]'\right\} \\ &\quad \times \frac{\text{Term} - 1}{\text{Term} - 2}\end{aligned} \tag{103}$$

where

$$\begin{aligned}\text{Term} - 1 &= \det \boldsymbol{\Psi_{CC}}^{\frac{\delta+Q-1}{2}} \boldsymbol{\Gamma}_Q\left(\frac{\delta+N+Q-1}{2}\right) \\ &\quad \times \boldsymbol{\Gamma}_M\left(\frac{\delta+Q+N+M-1}{2}\right)\end{aligned} \tag{104}$$

$$\begin{aligned}\text{Term} - 2 &= \pi^{\frac{NQ}{2}} \boldsymbol{\Gamma}_Q\left(\frac{\delta+Q-1}{2}\right)\boldsymbol{\Gamma}_M\left(\frac{\delta+Q+M-1}{2}\right) \\ &\quad \times \det\left(\boldsymbol{\Psi_{CC}} + \boldsymbol{Y_C}\boldsymbol{Y_C'}\right)^{\frac{\delta+N+Q-1}{2}} \det \boldsymbol{\zeta_0}^{\frac{N}{2}}.\end{aligned} \tag{105}$$

Integrating the above equation, in turn, over $\boldsymbol{\Sigma_{CC}}$, $\boldsymbol{\tau}$, and $\boldsymbol{\zeta}$, the first three terms yield unity as they are proper density functions leaving only the last term as the result, given by

$$= \frac{\text{Term} - 1}{\text{Term} - 2}. \tag{106}$$

Dividing the joint density in (103) by the term in (106), we get the result for the posterior density. ■

Appendix VII
Proof of Theorem 4

Proof: Mean of the $\widehat{\boldsymbol{\alpha}}_n$ is given by

$$\begin{aligned}\boldsymbol{E}(\widehat{\boldsymbol{\alpha}}_n) &= \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)\boldsymbol{E}(\boldsymbol{x}_n) \\ &= \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)\boldsymbol{E}(\boldsymbol{B}\boldsymbol{\alpha}_n + \boldsymbol{\eta}_n) \\ &= \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)\boldsymbol{B}\boldsymbol{\alpha}_n \\ &= \boldsymbol{\alpha}_n\end{aligned} \tag{107}$$

since $\boldsymbol{E}(\boldsymbol{\eta}_n) = \boldsymbol{0}$, $\boldsymbol{B'B} = \boldsymbol{I}$ and $\boldsymbol{C'B} = \boldsymbol{0}$. Therefore, $\widehat{\boldsymbol{\alpha}}_n$ is an *unbiased* estimator of $\boldsymbol{\alpha}_n$. Now, consider

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_n &= \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)\boldsymbol{x}_n \\ &= \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)(\boldsymbol{B}\boldsymbol{\alpha}_n + \boldsymbol{\eta}_n) \\ &= \boldsymbol{\alpha}_n + \left(\boldsymbol{B'} - \boldsymbol{\Sigma_{BC}}\boldsymbol{\Sigma_{CC}^{-1}}\boldsymbol{C'}\right)\boldsymbol{\eta}_n.\end{aligned} \tag{108}$$

The covariance matrix $\Sigma_{\widehat{\alpha}}$ of $\widehat{\alpha}_n$ is

$$\Sigma_{\widehat{\alpha}} = E(\widehat{\alpha}_n - \alpha_n)(\widehat{\alpha}_n - \alpha_n)'. \tag{109}$$

Substituting for $\widehat{\alpha}_n$ from (108), using the fact that $E(\eta_n \eta_n') = \Sigma_\eta$, and simplifying, we get

$$\Sigma_{\widehat{\alpha}} = \left(B' - \Sigma_{BC}\Sigma_{CC}^{-1}C'\right)\Sigma_\eta \times \left(B' - \Sigma_{BC}\Sigma_{CC}^{-1}C'\right)'. \tag{110}$$

We can express Σ_η in $(B\ C)$ space, by using the fact that $(B\ C)^{-1} = (B\ C)'$, we have

$$\begin{aligned}\Sigma_\eta &= (B\quad C)(B\quad C)'\Sigma_\eta(B\quad C)(B\quad C)' \\ &= (B\quad C)\begin{pmatrix}\Sigma_{BB} & \Sigma_{BC}\\ \Sigma_{CB} & \Sigma_{CC}\end{pmatrix}(B\quad C)'. \end{aligned}\tag{111}$$

After substituting for Σ_η from (111) into (110) and simplifying, we get the covariance matrix of the estimated α_n as

$$\Sigma_{\widehat{\alpha}} = \Sigma_{BB} - \Sigma_{BC}\Sigma_{CC}^{-1}\Sigma_{CB} = \zeta. \tag{112}$$

Note that, in the *white* noise case [52], due to the independence of the spaces spanned by the columns of B, and the columns of C, we have

$$\Sigma_{\widehat{\alpha}} = \Sigma_{BB}. \tag{113}$$

■

REFERENCES

[1] V. Chalana, "Deformable models for segmentation of medical ultrasound images," Ph.D. dissertation, Dept. Bioeng., Univ. Washington, Seattle, 1996.

[2] L. D. Cohen and I. Cohen, "Deformable models for 3-D medical images using finite elements and balloons," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1992, pp. 592–598.

[3] D. Terzopoulos and D. Metaxas, "Dynamic 3D models with local and global deformations: deformable superquadrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 7, pp. 703–714, Jul. 1991.

[4] T. McInerney and D. Terzopoulos, "A dynamic finite element surface model for segmentation and tracking in multidimensional medical images with application to cardiac 4D image analysis," *J. Comput. Med. Imag. Graph.*, to be published.

[5] L. D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-D and 3-D images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1131–1147, Oct. 1993.

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *Int. J. Comput. Vis.*, vol. 1, pp. 321–331, 1988.

[7] C. Stein, "Estimation of a covariance matrix," presented at the Rietz Lecture, Annu. Meeting Institute of Mathematics and Statistics, Atlanta, GA, 1975.

[8] ——, "Lectures on the theory of estimation of many parameters," in *Proc. Scientific Seminars of the Steklov Institute, Leningrad Division, Studies in the Statistical Theory of Estimation, Part I*, I. A. Ibraginov and M. S. Nikulin, Eds., 1977, pp. 4–65.

[9] S. P. Lin and M. D. Perlman, "An improved procedure for the estimation of a correlation matrix," in *Statistical Theory and Data Analysis*, K. Matusita, Ed. Amsterdam, The Netherlands: Elsevier, 1985, pp. 369–379.

[10] L. R. Haff, "Empirical bayes estimation of the multivariate normal covariance matrix," *Ann. Stat.*, vol. 8, pp. 586–597, 1980.

[11] J. M. Dickey, D. V. Lindley, and S. J. Press, "Bayesian estimation of the dispersion matrix of a multivariate normal distribution," *Commun. Stat. Theory Meth.*, vol. 14, no. 5, pp. 1019–1034, 1985.

[12] S. Wright, "The method of path coefficients," *Ann. Math. Stat.*, vol. 5, pp. 161–215, 1934.

[13] K. G. Jöreskog, *Structural Equation Models in the Social Sciences: Specification, Estimation and Testing*. Amsterdam, The Netherlands: North-Holland, 1977, pp. 265–286.

[14] C. Spearman, "Proof and measurement of association between two things," *Amer. J. Psych.*, vol. 15, pp. 72–202, 1904.

[15] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 763–767, Jul. 1996.

[16] J. P. Le Cadre, "Parametric methods for spatial signal processing in the presence of unknown colored noise fields," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 965–983, Jun. 1989.

[17] T. Leonard and J. S. J. Hsu, "Bayesian inference for a covariance matrix," *Ann. Stat.*, vol. 20, no. 4, pp. 1669–1696, 1992.

[18] M. J. Daniels and R. E. Kass, "Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models," Dept. Stat., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. 659, Nov. 1998.

[19] P. J. Brown, N. D. Le, and J. V. Zidek, *Aspects of Uncertainty: A Tribute to D. V. Lindley*. New York: Wiley, 1994, ch. Inference for a covariance matrix, pp. 77–92.

[20] R. M. Haralick and L. T. Watson, "A facet model for image data," *Comput. Graph. Image Process.*, vol. 15, pp. 113–129, 1981.

[21] R. M. Haralick, "A Bayesian approach to robust local facet estimation," in *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds. Dordrecht, The Netherlands: Reidel, 1987, pp. 85–97.

[22] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1992, vol. 1.

[23] R. M. Haralick, "Edge and region analysis for digital image data," *Comput. Vis., Graph., Image Process.*, vol. 12, pp. 60–73, 1980.

[24] ——, "The digital edge," in *Proc. IEEE Computer Soc. Conf. Pattern Recognition and Image Processing*, New York, 1981, pp. 285–294.

[25] ——, "Zero-crossing of second directional derivative edge operator," in *Proc. SPIE Symp. Robotic Vision*, Washington, DC, 1982, p. 23.

[26] ——, "Ridges and valleys on digital images," *Comput. Vis., Graph., Image Process.*, vol. 22, pp. 28–38, 1983.

[27] R. M. Haralick, L. T. Watson, and T. J. Laffey, "The topographic primal sketch," *Int. J. Robot. Res.*, vol. 2, pp. 50–72, 1983.

[28] R. M. Haralick, "Digital step edges from zero crossing of the second directional derivatives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 1, pp. 58–68, Jan. 1984.

[29] ——, "Cubic facet model edge detector and ridge-valley detector: implementation details," in *Pattern Recognition in Practice II*, E. S. Gelsema and L. N. Kanal, Eds. Amsterdam, The Netherlands: Elsevier, 1986, pp. 81–90.

[30] O. A. Zuniga and R. M. Haralick, "Integrated directional derivative gradient operator," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, no. 3, pp. 508–517, May/Jun. 1987.

[31] G. E. Forsythe, "Generation and use of orthogonal polynomials for data-fitting with a digital computer," *J. Soc. Ind. Appl. Math.*, vol. 5, pp. 74–88, 1957.

[32] D. Nadadur, "Noise Covariance Estimation in Low-level Computer Vision," Ph.D. dissertation, Dept. Elect. Eng., Univ. Washington, Seattle, Nov. 2001.

[33] A. P. Dawid, "Some matrix-variate distribution theory: notational considerations and a bayesian application," *Biometrika*, vol. 68, no. 1, pp. 265–274, 1981.

[34] M. S. Bartlett, "On the theory of statistical regression," in *Proc. Roy. Stat. Soc.*, vol. 53, Edinburgh, U.K., 1933, pp. 260–283.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, Dec. 1976.

[36] N. D. Le, L. Sun, and J. V. Zidek, "Bayesian spatial interpolation and backcasting using Gaussian-generalized inverted Wishart model," Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., 1999.

[37] B. M. Golam Kibria, "Multivariate Bayesian spatial interpolation using Gaussian-generalized inverted Wishart model," Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., 2000.

[38] C. F. Chen, "Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis," *J. Roy. Stat. Soc. B*, no. 41, pp. 235–248, 1979.

[39] G. J. McLachlan and T. Krishan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

[40] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed. New York: Wiley, 1993.

[41] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: Wiley, 1968.

[42] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[43] J. J. Moré and D. J. Thuente, Line search algorithms with guaranteed sufficient decrease, Math. Comput. Sci. Division, Argonne Nat. Lab., Preprint MCS-P330-1092, Oct. 1992.
[44] ——, "Line search algorithms with guaranteed sufficient decrease," *ACM Trans. Math. Software*, vol. 20, no. 3, pp. 286–307, 1994.
[45] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.
[46] T. Kanungo and R. M. Haralick, "Multivariate hypothesis testing for Gaussian samples," Intell. Syst. Lab., Dept. Elect. Eng., Univ. Washington, Seattle, Tech. Rep. ISL-TR-95-05, Oct. 5, 1995.
[47] D. Nadadur, R. M. Haralick, and D. E. Gustafson, Integrated second directional derivative ridge operator, to be published.
[48] A. M. López, F. Lumbreras, J. Serrat, and J. J. Villanueva, "Evaluation of methods for ridge and valley detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 4, pp. 327–335, Apr. 1999.
[49] J. Wishart, "The generalized product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20, no. A, pp. 32–52, 1928.
[50] ——, "Proofs of the distribution law of the second order moment statistics," *Biometrika*, vol. 35, pp. 55–57, 1948.
[51] J. Wishart and M. S. Bartlett, "The generalized product moment distribution in a normal system," in *Proc. Cambridge Philosophical Soc.*, vol. 29, 1933, pp. 260–270.
[52] D. Nadadur and R. M. Haralick, "A Bayesian framework for noise covariance estimation using the facet model," Siemens Medical Syst., Inc., Ultrasound Group (SMS-UG) and Intell. Syst. Lab. (ISL), Issaquah, WA, Tech. Rep. TR (DN) 2001-0001, 2001.

Desikachari Nadadur (M'00) was born in Cuddapah, Andhra Pradesh, India, in 1970. He received the B.Tech. degree in electronics and communications engineering from Sri Venkateswara University, Tirupati, Andhra Pradesh, in 1990, the M.S. degree in electrical engineering from Gonzaga University, Spokane, WA, in 1993, and the Ph.D. degree in electrical engineering, with a specialization in image processing and computer vision, from the University of Washington (UW), Seattle, in 2001, under the supervision of Prof. R. M. Haralick. His dissertation was titled "Noise Covariance Estimation in Low-Level Computer Vision."

Since 1999, he has been a Senior Scientist in the Advanced Imaging Applications Group of the Developing Competency Department, Siemens Medical Solutions, Inc., Ultrasound Division, Issaquah, WA. He was a Teaching Assistant with the Department of Electrical Engineering, Gonzaga University, Spokane, from 1991 to 1993. He was a predoctoral Research Associate II at the Intelligent Systems Laboratory, UW, from 1993 to 1997. He was a Software Engineer in the Advanced Technologies Group at Electronics for Imaging, Inc., Foster City, CA, from 1997 to 1999. He has coauthored several publications in the fields of image processing, mathematical morphology, and computer vision, and holds several patents. His current research interests include application of deformable, finite-element models, PDE and level-set methods to medical image segmentation, Bayesian and other statistical methods in image analysis, and performance characterization of vision algorithms.

Robert Martin Haralick (F'84) is a Distinguished Professor in the Department of Computer Science, Graduate Center, City University of New York. He was the Boeing Clairmont Egtvedt Professor in the Department of Electrical Engineering, University of Washington, Seattle. He is responsible for developing the grayscale co-occurrence texture analysis technique and the facet model technique for the image processing. He has worked on robust methods for photogrammetry and developed fast algorithms for solving the consistent labeling problems in high-level computer vision. He has developed shape analysis and extraction techniques using mathematical morphology, the morphological sampling theorem, and fast recursive morphology algorithms. Together with Prof. I. Phillips, he has developed a comprehensive ground-truthed set of some 1600 document image pages (most in English) and some 200 pages in Japanese in the area of document image understanding. He has also developed algorithms for document skew angle estimation, zone delineation, and word and text line bounding box delineation. His most recent research is in the area of computer vision performance characterization and covariance propagation. He has published more than 490 papers.

Dr. Haralick was made a Fellow of IEEE for his contributions in computer vision and image processing and a Fellow of the International Association for Pattern Recognition (IAPR) for his contributions in pattern recognition, image processing, and for services to IAPR. He was also the President of the IAPR.

David Earl Gustafson received the B.S. degree in physics from Hamline University, St. Paul, MN, and the Ph.D. degree in physics from the University of Virginia, Charlottesville.

He was a Nuclear Physics Postdoctoral Fellow, Florida State University, Gainesville, and a Postdoctoral Research Fellow at the Mayo Clinic, Rochester, MN. He is, at present, a Senior Director at Siemens Medical Solutions USA, Inc., Ultrasound Division, Issaquah, WA. He has extensive experience, including medical imaging research and development in both university academic and industrial settings. five years of experience in academia, including teaching in radiology residency and technologist programs, and 24 years in the medical imaging industry. His industry positions, which included group to department managerial roles for 20 years, have focused on conducting and directing product research and development, with a strong focus on extensive strategic advanced development carrying product features from inception to product release. His experience has also included manufacturing, marketing, and business management responsibility and management of multisite development teams, widely diversified geographically, including extensive R & D staff consisting of engineers and scientists, most with advanced degrees, and product development marketing personnel. His development projects have covered most modern medical imaging methods. Broad imaging and image processing product development experience includes: computed tomography, digital X-ray fluoroscopy, radiography and angiography products, mobile X-ray systems, image-intensified special procedures radiology and cardiology systems, mammography, ultrasound, PACS and connectivity, CAD, and nuclear medicine products.