On Posterior Distributions for Signals in Gaussian Noise With Unknown Covariance Matrix

Lennart Svensson and Magnus Lundberg

***Abstract*—A Bayesian approach to estimate parameters of signals embedded in complex Gaussian noise with unknown color is presented. The study specifically focuses on a Bayesian treatment of the unknown noise covariance matrix making up a nuisance parameter in such problems. By integrating out uncertainties regarding the noise color, an enhanced ability to estimate both the signal parameters as well as properties of the error is exploited. Several noninformative priors for the covariance matrix, such as the reference prior, the Jeffreys prior, and modifications to this, are considered. Some of the priors result in analytical solutions, whereas others demand numerical approximations. In the linear signal model, connections are made between the standard Adaptive Maximum Likelihood (AML) estimate and a Bayesian solution using the Jeffreys prior. With adjustments to the Jeffreys prior, correspondence to the regularized solution is also established. This in turn enables a formal treatment of the regularization parameter. Simulations indicate that significant improvements, compared to the AML estimator, can be obtained by considering both the derived regularized solutions as well as the one obtained using the reference prior. The simulations also indicate the possibility of enhancing the predictions of properties of the error as uncertainties in the noise color are acknowledged.**

***Index Terms*—Adaptive beamforming, Bayesian estimation, Jeffreys prior, nuisance parameters, posterior distribution, reference prior, regularization.**

I. Introduction

ESTIMATING parameters of signals embedded in additive complex Gaussian noise is one of the most frequently encountered problems in statistical signal processing. Depending on the signal model, the problem finds a variety of applications. Some of the most extensively studied areas include adaptive beamforming [1]–[3] and direction-of-arrival estimation [3], [4]. In most applications, the ever-increasing demand of functionality and performance has led to a trend in which the number of features and degrees of freedom continue to increase, resulting in problems of high dimensionality. A typical example is Space-Time-Adaptive-Processing (STAP) [5], [6], where the traditional *temporal-then-spatial* processing technique is replaced by a *joint* approach that drastically increases dimensionality. A crucial issue in many of these applications is the lack of knowledge regarding the noise covariance matrix. Although of no interest in itself, it is essential for the inference of the desired signal parameters. To aid the estimation procedure, information regarding the noise color is commonly provided by a training data set containing *noise-only* measurements. For small training sets, however, large uncertainties regarding the noise color most often lead to substantial degradation in the estimator performance. For applications with high dimensionality, it is therefore crucial that the information provided by the training data set is exploited in the best possible way.

The traditional approach to treat the problem of unknown noise color is to use the noise-only measurements to form the *sample covariance* matrix [7]; this is an estimate that is known to be optimal in the *maximum likelihood (ML)* sense, rendering good asymptotic properties. This estimate is then used in place of the true underlying covariance [2], [6] through the *Certainty Equivalence (CE)* principle. Although not explicitly stated, this strategy is also frequently considered for algorithms derived under a white noise assumption. These algorithms are often claimed to be equally applicable in colored noise scenarios due to prewhitening using the sample covariance matrix; see [8] for an array processing example. Unfortunately, despite good asymptotic behavior, the sample covariance estimate has certain undesirable properties for small sample sizes. This is particularly well studied for the case of real matrices [9]. The main weakness noted in [9], and references therein, is a significant spread in the distribution of the eigenvalues, especially when some of the eigenvalues of the underlying covariance are close to identical. See [10] for related results for the complex setting. A frequently used attempt to improve on the performance that has proven successful in many applications is to update the sample covariance estimate according to the regularization procedure [3]. In recent years, adjustments based on truncation of the Multistage Wiener Filter (MWF) have also received considerable interest [5], [11]. Both of these approaches, however, have the drawback of containing a design parameter that is not straightforward to assign.

Besides methods based on the CE principle, there exist numerous other alternatives in the literature. Within the signal processing community, perhaps the most recognized alternative is *joint ML*, where the parameter of interest and the covariance are estimated jointly [12]. Alternatives based on specific physical insight enabling parametrization of the covariance are also commonly encountered [13], [14]. Nevertheless, the vast majority of solutions concern different types of classical treatments. The very foundations of the classical treatment of a nuisance parameter (in this case the noise color) are questioned by most Bayesians [15]. The main arguments concern the fact that uncertainties in the nuisance parameter are not acknowledged. It is argued that since one cannot claim to know the nuisance parameters, all possible values should be taken into

Manuscript received April 29, 2003; revised September 25, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Athina Petropulu.

The authors are with the Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Göteborg, Sweden (e-mail: lennart.svensson@s2.chalmers.se; mlg@s2.chalmers.se).

Digital Object Identifier 10.1109/TSP.2005.853102

account. The potential drawbacks of the classical approach are illustrated by examples where the classical CE approach leads to inconsistencies. Besides this criticism, which is of a general and philosophical character, it is, of course, particularly troublesome that the standard estimate of the nuisance parameter has well-documented weaknesses for the current scenario.

In this paper, we explore the possible advantages of a Bayesian treatment of the noise covariance matrix. By doing so, our aim is twofold. First, we hope to improve the estimator performance for small sets of training data (all methods considered here have about the same performance for large sets of data). Second, by incorporating the uncertainties in the covariance matrix, we hope to achieve a more appropriate description of the likelihood function. A desirable consequence of the latter, which we will discuss further later on, is a more accurate description of posterior uncertainties. Having said that, it is important to note that applying the Bayesian methodology is not done without effort. Generally speaking, the Bayesian approach leads to two other difficulties in that 1) one has to find a reasonable prior for the nuisance parameter, and 2) the solution includes a marginalization with respect to the covariance matrix; therefore, computation can be very demanding. An essential part of this study is therefore to propose well-performing prior distributions and to investigate how to implement them; some priors will enable analytical solutions, whereas others require numerical evaluations.

To maintain a general perspective, we do not assume to have any prior knowledge regarding the covariance matrix. To reflect this ignorance, we apply noninformative priors. Deriving such priors is far from straightforward. In fact, formal ways of obtaining them are still under investigation [16]. Among the priors that we investigate are common noninformative alternatives such as the Jeffreys prior [17], [18] and the Reference prior [9], [19]. Observations made in connection to this will also lead us to propose priors of our own. While performing this study, we do not put any restrictions on the signal model. Thus, the paper is not intended for any particular application, even though the results may be more useful in high-dimensional settings. We will, however, put some extra focus on the linear signal model. This is done partly because it is an important model, which is well studied from a classical perspective [20], [21] and partly because its simplicity enables the derivation of further results.

A. Notation

Throughout this paper, bold lowercase symbols represent complex-valued vectors, whereas bold uppercase symbols denote complex matrices. Superscript H denotes Hermitian transpose, $|\cdot|$ defines the determinant, and etr$(\cdot)$ stands for $\exp(\mathrm{tr}(\cdot))$, where tr is the trace operator. The matrix denoted $\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is diagonal of size $p \times p$ with $\lambda_1, \lambda_2, \ldots, \lambda_p$ along the diagonal, whereas $\mathrm{vec}(\mathbf{A})$ is the vectorized $\mathbf{A}$, i.e., $\mathrm{vec}(\mathbf{A}) = [\mathbf{A}_1^H, \mathbf{A}_2^H, \ldots, \mathbf{A}_p^H]^H$, where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p]$. Further, $\otimes$ denotes the Kronecker product, δ_{kl} is Kronecker's delta function, and Γ is the gamma function. Two distributions are central in the paper.

- If x is complex Gaussian of dimension $p \times 1$, i.e., if $\mathbf{x} \sim \mathcal{CN}_p(\boldsymbol{\mu}, \mathbf{R})$, we denote its density as

$$\mathcal{CN}_p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{R}) = \pi^{-p} |\mathbf{R}^{-1}| \,\mathrm{etr}\left\{-(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^H \mathbf{R}^{-1}\right\}.$$

- If $\mathbf{R}$ is complex Wishart of dimension $p \times p$, i.e., if $\mathbf{R} \sim \mathcal{CW}_p(\boldsymbol{\Sigma}, M)$, we denote its density as

$$\mathcal{CW}_p(\mathbf{R}|\boldsymbol{\Sigma}, M) = \frac{|\mathbf{R}|^{M-p} \mathrm{etr}\left\{-\boldsymbol{\Sigma}^{-1}\mathbf{R}\right\} |\boldsymbol{\Sigma}^{-1}|^M}{c(p, M)}$$

where $c(p, M) = \pi^{p(p-1)/2} \prod_{i=1}^{p} \Gamma(M - p + i)$.

II. Data Model

Consider the following model of the p-dimensional complex measurement vector at time k:

$$\mathbf{x}_k = \boldsymbol{s}_k(\boldsymbol{\theta}) + \mathbf{n}_k. \tag{1}$$

Here, the parameters of interest are contained in the G-dimensional parameter vector $\boldsymbol{\theta}$. The measurement noise $\mathbf{n}_k$ is complex Gaussian distributed with zero mean and unknown covariance matrix $\mathbf{R}$. Further, it is assumed independent (in time) and identically distributed (i.i.d.), i.e., $\mathrm{E}\left\{\mathbf{n}_k \mathbf{n}_l^H\right\} = \delta_{kl}\mathbf{R}$. To aid the estimation procedure, knowledge regarding noise color is provided from a set of i.i.d. *noise-only* training data samples $\mathbf{z}_m$. These samples are drawn from the same distribution and are assumed to be independent of the measurement noise $\mathbf{n}$. As a consequence, $\mathbf{z}_m \sim \mathcal{CN}_p(\mathbf{0}, \mathbf{R})$, and $E\{\mathbf{z}_m n_k^H\} = 0$. For notation, we collect the N_1 primary data into the measurement matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_1}]$ of size $p \times N_1$, while denoting the corresponding signal and noise matrices $\mathbf{S}(\boldsymbol{\theta}) = [\mathbf{s}_1(\boldsymbol{\theta}), \mathbf{s}_2(\boldsymbol{\theta}), \ldots, \mathbf{s}_{N_1}(\boldsymbol{\theta})]$ and $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_{N_1}]$, respectively. We can now express (1) in the equivalent form

$$\mathbf{X} = \mathbf{S}(\boldsymbol{\theta}) + \mathbf{N}. \tag{2}$$

The secondary data consists of N_2 training noise samples, which are collected into the matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_2}]$ of size $p \times N_2$.

A. Linear Signal Model

In general, the signal $\mathbf{S}(\boldsymbol{\theta})$ can have any dependence on $\boldsymbol{\theta}$, depending on the application. However, in this paper, some of the results are specialized to the linear model, which is one of the most commonly used models. Here, the signal lies in a p-dimensional subspace spanned by the columns of some known matrix $\mathbf{H}$ of size $p \times L$. For this model, the parameter vector $\boldsymbol{\theta}$ is divided into N_1 vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{N_1}$ of size $L \times 1$, and the parameter vector $\boldsymbol{\theta}_k$ now determines the unknown signal amplitudes at time k. The signal at time instant k can then be written as[1]

$$\boldsymbol{s}_k(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}_k. \tag{3}$$

[1]Note that $\mathbf{s}_k$ is, in fact, only dependent of $\boldsymbol{\theta}_k$.

With slight abuse of notation,[2] we use $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{N_1}]$, and therefore, the modeled signal can be expressed as

$$\mathbf{S}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}. \tag{4}$$

Important examples of this model arise naturally in applications employing antenna arrays such as sonar, radar, and wireless communications, where the task is to extract the desired signals by suppressing interfering signals and noise. Here, the columns of $\mathbf{H}$ model the *steering vectors*, whereas the parameters model the strength of the desired signals.

If the covariance $\mathbf{R}$ is known, the ML estimate of the parameter $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\mathbf{H}^H\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{R}^{-1}\mathbf{X} \tag{5}$$

(see [20]). In our setting, however, the noise covariance is not known, and one has to treat it as a nuisance parameter. As mentioned, the joint ML and the CE principle are common classical ways to approach this. In fact, for the linear model, it can be shown that the resulting estimators for $\boldsymbol{\theta}$ coincide. This estimator, which is commonly referred to as the *Adaptive Maximum Likelihood (AML)* estimator, is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{AML}} = (\mathbf{H}^H\hat{\mathbf{R}}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\hat{\mathbf{R}}^{-1}\mathbf{X} \tag{6}$$

where

$$\hat{\mathbf{R}} = \frac{1}{N_2}\sum_{k=1}^{N_2}\mathbf{z}_k\mathbf{z}_k^H = \frac{1}{N_2}\mathbf{Z}\mathbf{Z}^H. \tag{7}$$

For large sets of training data (large N_2 compared to p), this procedure performs well since the estimated covariance is close to the true one. However, as discussed in the introduction, it is also known to have certain weaknesses for small data sets and a commonly used procedure to improve on it is regularization. Here, the covariance estimate is adjusted according to

$$\hat{\mathbf{R}}_\alpha = \frac{1}{N_2}(\mathbf{Z}\mathbf{Z}^H + \alpha\mathbf{I}) \tag{8}$$

where $\mathbf{I}$ is the identity matrix. With an appropriate choice of α, this strategy often offers significant improvements compared with AML. Traditionally, the parameter α is a design parameter chosen based on assumptions of the signal and noise environment; see [22] and references therein. In reality, such knowledge may not be available. Some alternative approaches on how to select α from data have been proposed [3], [23]. Nevertheless, these techniques are more or less *ad hoc*, promoting the need for a more formal treatment. We will return to this in Section VI.

B. General Signal Models

Besides the linear model, there exists a tremendous variety of important models with applications in the most varying fields. It is impossible within this framework to give a fair and extensive overview. The many important nonlinear signal models include sinusoids, chirps, and harmonics, among several others. We note that for some of these models, it is common that not only the parameters of interest, but also the order G, is considered unknown. One such typical case is the superposition of an unknown number of sinusoids, which is frequently considered in array processing. Such issues, involving model selection, are not considered in this paper. See the strategies in, for instance, [24] and [25]. Besides that, many of the results derived here apply to a general signal model. However, we frequently specialize to the linear model, for which further results are derived.

[2]Clearly, $\boldsymbol{\theta}$ is now a matrix of size $L \times N_1$ instead of vector.

III. Problem Formulation

The main purpose of this paper is to promote and investigate Bayesian treatment[3] of the noise covariance $\mathbf{R}^{-1}$. Here, instead of estimating the nuisance parameter through the CE or Joint ML approach, we consider the integrated likelihood

$$f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \int f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{R}^{-1})\pi(\mathbf{R}^{-1}|\boldsymbol{\theta})d\mathbf{R}^{-1} \tag{9}$$

(see [15]). This likelihood can be interpreted as an average over all conditional likelihoods given $\mathbf{R}^{-1}$, weighted by their corresponding prior probabilities. The main advantage of this strategy, compared to traditional ones, is that the integrated likelihood automatically incorporates the uncertainties regarding the noise color. As a consequence, this likelihood will typically have an increased spread as a function of $\boldsymbol{\theta}$. Once derived, it can either be used within the classical framework, treating $\boldsymbol{\theta}$ deterministically, or as we choose, used in a fully Bayesian approach considering $\boldsymbol{\theta}$ to be stochastic.

By treating the covariance in a Bayesian manner, we have two aims. First, and most importantly, we want to estimate $\boldsymbol{\theta}$ and investigate if the performance can be improved, compared to the classical approaches. Second, we want to be able to estimate properties of the error. We are, for instance, interested in the question *"How big is the squared magnitude of the error for this particular set of data?"*. Answering this question can indeed be of great interest in many applications. Important examples arise in post-processing tasks such as tracking and signal detection, where it is often useful to know the quality of the estimate for the data at hand. To be able to describe the posterior uncertainties in the parameter of interest (given that we treat $\boldsymbol{\theta}$ stochastically), properties of the likelihood other than the position of the maximum are also important. By incorporating the uncertainties in $\mathbf{R}^{-1}$ through (9), we believe that a more appropriate description of the overall likelihood for $\boldsymbol{\theta}$ can be achieved.

One can, of course, try to approach the above issues also using classical methods. The second issue, however, cannot easily be addressed from this perspective. In fact, conditioning on measured data leads to philosophical problems if we want to study uncertainties within the classical framework. The reason is that the estimate can no longer be treated stochastically, and since all underlying parameters are treated deterministically, there are no uncertainties left to consider.[4] A commonly used *ad hoc* solution is, however, to estimate the underlying parameters and

[3]Since the likelihood is expressed more easily in terms of $\mathbf{R}^{-1}$, we regard this to be our nuisance parameter instead of $\mathbf{R}$.

[4]These difficulties are highly connected to the ones from which classical methods suffer in the context of the post-processing task in hypotheses testing [26], [27].

evaluate the squared magnitude of the error averaged over data by using the estimated parameters in place of the true ones. In the Bayesian framework, on the other hand, considering posterior uncertainties is something very natural. Here, the parameter of interest, and not the estimate, is considered to be stochastic. Therefore, the estimation error is also stochastic, enabling the opportunity to estimate its properties.

In principle, one may be interested in any property of the error. In a multidimensional setting, for instance, it may be desirable to estimate the direction of the error vector. We, however, focus on estimating the squared error, which appears to be a natural measure on the estimation quality. More precisely, given an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, we want an estimate $\hat{\sigma}^2$ of the squared error $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2$ for the particular data at hand. The performance of the different estimators $\hat{\sigma}^2$ are considered in the mean square error (MSE) sense, i.e., in terms of $E\{(\hat{\sigma}^2 - |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2)^2\}$. This measure is by no means the only possible one to study. Nevertheless, we believe it is reasonable and serves the purpose of illustrating the different methods' ability to describe posterior uncertainties.

One problem with comparing different methods is that they are estimating different errors. In principle, different Bayesian methods correspond to different choices of prior, and for every particular choice of prior, both $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ are derived. Depending on the ability to estimate $\boldsymbol{\theta}$, the task of estimating $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2$ will differ in difficulty in the sense that it will affect the feasibility to achieve a small MSE. If the fluctuations (the variance) in $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2$ are small, it is much easier to obtain a small MSE in $\hat{\sigma}^2$ compared to when the fluctuations are large. Therefore, to compare the quality of different estimators, we normalize our performance measure (the MSE of $\hat{\sigma}^2$) with the variance of $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2$. We refer to the resulting measure as the *Mean Square Error Prediction Factor (MSEPF)*

$$\text{MSEPF} = \frac{E\left\{\left(\hat{\sigma}^2 - |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2\right)^2 \middle| \boldsymbol{\theta}, \mathbf{R}\right\}}{Var\left\{|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2 \middle| \boldsymbol{\theta}, \mathbf{R}\right\}}. \tag{10}$$

Note that one reasonable estimator $\hat{\sigma}^2$ is one that is near the mean square error $E\{|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2 | \boldsymbol{\theta}, \mathbf{R}\}$, rendering a MSEPF value close to one independently of the variance of $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2$.

In conclusion, to fulfill our aims of estimating both the parameter $\boldsymbol{\theta}$ as well as the quality of the estimate, we model $\boldsymbol{\theta}$ stochastically. In order to find the estimates, we require the posterior distribution $f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})$, or to be more precise, we need at least to be able to evaluate properties of this distribution such as the mean and covariance.

IV. Choice of Priors

In order to carry out a Bayesian analysis, we need a prior distribution on the parameters $\boldsymbol{\theta}$ and $\mathbf{R}^{-1}$. Finding a reasonable prior distribution is a vital part of Bayesian analysis. It is through the choice of prior that the statistical properties of the resulting estimator are determined. As we assume to have no prior information regarding the parameters, we will resort to the use of noninformative priors. These priors are motivated to a great extent by their ability to describe our ignorance regarding the parameters. In addition, they can be used as default priors when it is not feasible to elicit our priors subjectively due to, for example, high dimensionality or in some other way complicated settings.[5] Finding a noninformative prior is far from easy. In fact, it has yet to be agreed upon as how to measure the information in a prior. Therefore, given two different priors, it is generally not possible to decide which one contains the least information. However, many serious attempts have been made to develop schemes to derive good noninformative (or default) priors; see [16], [17], and [19]. We will discuss a few of these below.

We assume that our prior distributions can be factorized as

$$\pi(\boldsymbol{\theta}, \mathbf{R}^{-1}) = \pi(\boldsymbol{\theta})\, \pi(\mathbf{R}^{-1}). \tag{11}$$

Hence, we assume that the parameter of interest $\boldsymbol{\theta}$ does not contain any information *a priori* about the nuisance parameter $\mathbf{R}^{-1}$ and vice versa. This assumption is in agreement with most common methods to derive noninformative priors and is therefore not to be considered *ad hoc*. As our aim is to propose and compare different methods to treat uncertainties regarding the noise color, we will limit our discussion to the choice of $\pi(\mathbf{R}^{-1})$. We are, however, interested in carrying out a complete Bayesian analysis at least for the linear model and, therefore, need a reasonable prior $\pi(\boldsymbol{\theta})$ for this scenario. Doing a complete study on how to choose this prior is beyond the scope of this paper. We only consider the uniform prior, which is the one suggested by most formal methods to derive noninformative prior distributions. For the covariance matrix, on the other hand, we will, in principle, consider four priors. The first three can be joined in a uniform framework, whereas the last is treated separately.

A. Jeffreys Prior

Undoubtedly the most commonly used noninformative (or default) prior is the *uniform prior* $\pi(\mathbf{R}^{-1}) \propto 1$. Even though it is intuitive at first, there are some well-documented difficulties with the uniform prior. Perhaps the most recognized is the lack of invariance to parametrization [16]. Indeed, it is not appealing that the choice of parametrization may influence the resulting estimate.

Jeffreys acknowledged this problem and proposed a different prior, which is commonly referred to as the *Jeffreys prior*; see, e.g., [16] and [17]. This prior, which we denote $\pi_J(\mathbf{R}^{-1})$, can generally be described as follows. Let $\boldsymbol{\phi}$ denote the parameters describing $\mathbf{R}^{-1}$. To be precise, one should thus write $\mathbf{R}^{-1}(\boldsymbol{\phi})$ instead of $\mathbf{R}^{-1}$, but the dependence on $\boldsymbol{\phi}$ will be dropped for notational convenience. Denoting the Fisher information matrix as $\mathbf{I}(\boldsymbol{\phi})$, where

$$\mathbf{I}(\boldsymbol{\phi})_{ij} = -E\left\{\frac{\partial^2 l}{\partial \phi_i \partial \phi_j}\right\} \tag{12}$$

and l is the log-likelihood, $\pi_J(\mathbf{R}^{-1})$ is defined as

$$\pi_J\left(\mathbf{R}^{-1}\right) \propto |\mathbf{I}(\boldsymbol{\phi})|^{1/2}. \tag{13}$$

[5] One could question if this is possible even in the simplest of settings.

Among the many intuitively appealing properties of the Jeffreys prior is the invariance to reparametrization. The standard parametrization is to use

$$\boldsymbol{\phi} = [\mathbf{R}_{11}^{-1}, \mathbf{R}_{22}^{-1}, \ldots, \mathbf{R}_{pp}^{-1}, \mathrm{Re}\{\mathbf{R}_{12}^{-1}\}, \mathrm{Im}\{\mathbf{R}_{12}^{-1}\}, \mathrm{Re}\{\mathbf{R}_{13}^{-1}\}, \mathrm{Im}\{\mathbf{R}_{13}^{-1}\}, \ldots, \mathrm{Re}\{\mathbf{R}_{p-1\,p}^{-1}\}, \mathrm{Im}\{\mathbf{R}_{p-1\,p}^{-1}\}]^T.$$

For this parametrization and the complex data model, the prior is given by

$$\pi_J\left(\mathbf{R}^{-1}\right) \propto |\mathbf{R}^{-1}|^{-p} \tag{14}$$

(see [18]). Here, the differential volume element is given by $d\mathbf{R}^{-1} = d\boldsymbol{\phi} = d\mathbf{R}_{11}^{-1}\, d\mathbf{R}_{22}^{-1} \cdots d\mathrm{Re}\{\mathbf{R}_{p-1\,p}^{-1}\}\, d\mathrm{Im}\{\mathbf{R}_{p-1\,p}^{-1}\}$. For $\mathbf{Z}$ complex Gaussian as above, the use of the Jeffreys prior results in a posterior that is complex Wishart distributed according to $\mathbf{R}^{-1}|\mathbf{Z} \sim \mathcal{CW}_p((\mathbf{Z}\mathbf{Z}^H)^{-1}, N_2)$ [28]. Hence, the posterior mean [29] will equal the standard sample covariance estimate in (7). This supports the use of the Jeffreys prior as the resulting estimate is commonly used. At the same time, it also gives us reason to question it since the sample covariance is known to have a significant spread in the distribution of the eigenvalues. In fact, it has previously been reported that the Jeffreys prior can have problems in multidimensional settings [19]. All in all, this leads us to study its behavior with respect to the eigenvalues in the current setting.

As the prior is given with respect to the parametrization $\boldsymbol{\phi}$ defined above, it is very difficult to draw any conclusions about the distribution of the eigenvalues from the expression in (14). However, as $\mathbf{R}^{-1}$ is a positive-semidefinite Hermitian matrix, it can be written as $\mathbf{R}^{-1} = \mathbf{B}\mathbf{D}\mathbf{B}^H$, where $\mathbf{D}$ is a diagonal matrix that contains the eigenvalues d_i of $\mathbf{R}^{-1}$ along the diagonal, and $\mathbf{B}$ is a unitary matrix. Parameterizing $\mathbf{R}^{-1}$ in terms of $\mathbf{B}$ and $\mathbf{D}$ requires the change-of-variables formula

$$\pi_J(\mathbf{D}, \mathbf{B})\, d\mathbf{D}\, d\mathbf{B} = \pi_J(\mathbf{R}^{-1})\, d\mathbf{R}^{-1} \tag{15}$$

where

$$d\mathbf{R}^{-1} = \Pi_{i<j}(d_i - d_j)^2\, d\mathbf{D}\, d\mathbf{B} \tag{16}$$

(see [10]). Hence, in the eigenvalue parametrization, we have

$$\pi_J(\mathbf{D}, \mathbf{B}) \propto \frac{\Pi_{i<j}(d_i - d_j)^2}{\Pi_{i=1}^{p} d_i^p}. \tag{17}$$

The distribution is now represented in such a way that it is possible to interpret some properties regarding the distribution in respect to the eigenvalues d_i. Clearly, the prior is zero when two or more of the eigenvalues are identical and small when they are close to identical. These properties will carry over to the posterior distribution. We regard this as a deficiency of this prior since it is likely to lead to estimators with poor performance when the true covariance matrix has these properties. Further, saying *a priori* that the eigenvalues are well separated is not what we consider noninformative. To illustrate this phenomenon, we study the following example.

Example 1: Consider a two-dimensional (2-D) ($p = 2$) white noise scenario in which the true covariance matrix is given by

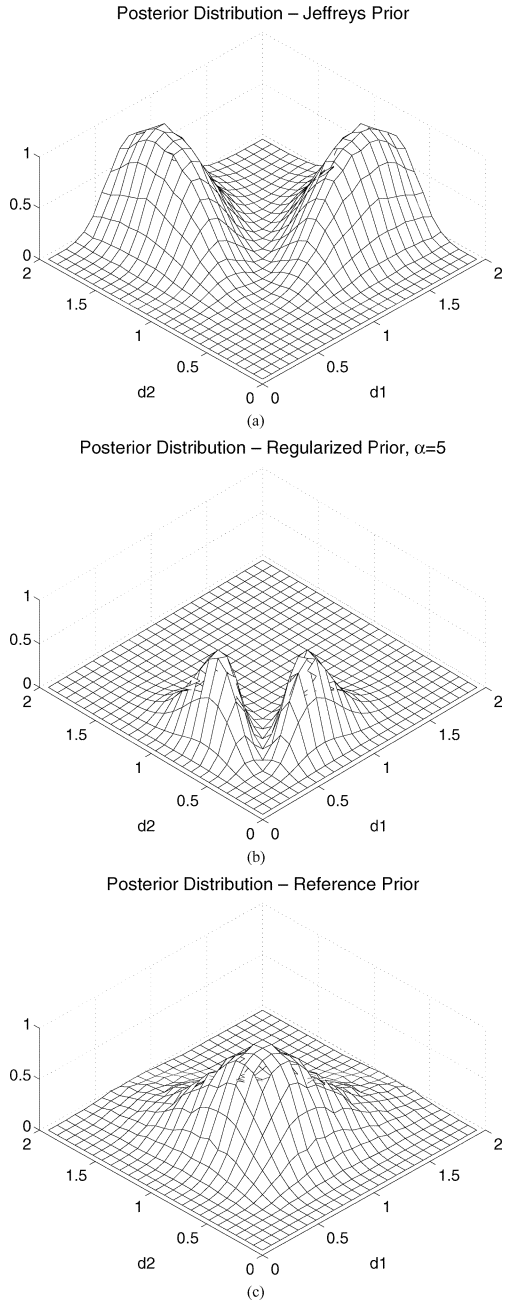


Fig. 1. Posterior distributions of the eigenvalues of $\mathbf{R}^{-1}$ when the underlying covariance matrix has eigenvalues (1,1). (a) Result using the Jeffreys prior. (b) Result using the regularization prior, with $\alpha = 5$. (c) Result using the reference prior.

$\mathbf{R} = \mathrm{diag}(1, 1)$. We generate $N_2 = 10$ training data samples $\mathbf{z}_k \sim \mathcal{CN}_p(0, \mathbf{R})$ and draw a large number of samples $\mathbf{R}_k^{-1}$ from the posterior distribution $f(\mathbf{R}^{-1}|\mathbf{Z})$, using the Jeffreys prior on $\mathbf{R}^{-1}$. The inherent problem of the Jeffreys prior is illustrated by plotting a histogram for the eigenvalues of $\mathbf{R}_k^{-1}$. This is done in Fig. 1(a). It can be observed that even though we have five times as many training data as the dimension of the matrix, very few samples are found in the region close to the point (1,1), which corresponds to the true eigenvalues. This is in agreement with the observations made from (17). ■

Obviously, the Jeffreys prior has certain undesirable properties regarding its treatment of the eigenvalues. An interesting

observation in connection to this is that also the sample covariance is complex Wishart distributed [7]. Thereby, the above example also illustrates the previously mentioned weaknesses of this estimate. Despite its drawbacks, it is important to note that for applications where the eigenvalues are of minor importance, employing the Jeffreys prior on the unknown covariance may still be a suitable choice. The reason for this is that this prior, although in some sense unsound with respect to the eigenvalues, has the advantage of enabling analytical marginalization of the covariance matrix [18]. Due to this beneficial property, it is of interest to circumvent its weakness regarding the distribution of the eigenvalues while maintaining the possibility of analytical marginalization.

B. Family of Jeffreys-Like Priors

To propose adjustments to the Jeffreys prior, it is important to identify those modifications that maintain the property of analytical marginalization. In essence, the marginalization procedure involves computation of the integral

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z}) &= \int f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\mathbf{R}^{-1}) f(\mathbf{R}^{-1}|\mathbf{X},\mathbf{Z}) d\mathbf{R}^{-1} \quad (18) \\ &\propto \int f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\mathbf{R}^{-1}) f(\mathbf{X}|\mathbf{R}^{-1}) f(\mathbf{Z}|\mathbf{R}^{-1}) \pi(\mathbf{R}^{-1}) d\mathbf{R}^{-1} \quad (19) \end{aligned}$$

where the fact that $\mathbf{X}$ and $\mathbf{Z}$ are independent given the covariance $\mathbf{R}^{-1}$, i.e., that $f(\mathbf{X},\mathbf{Z}|\mathbf{R}^{-1}) = f(\mathbf{X}|\mathbf{R}^{-1}) f(\mathbf{Z}|\mathbf{R}^{-1})$, is used in the second step. As $f(\mathbf{Z}|\mathbf{R}^{-1}) \propto |\mathbf{R}^{-1}|^{N_2} \text{etr}\{-\mathbf{R}^{-1}\mathbf{Z}\mathbf{Z}^H\}$, we can include factors of the form $|\mathbf{R}^{-1}|^K \text{etr}\{-\mathbf{R}^{-1}\boldsymbol{\Lambda}\}$, where $\boldsymbol{\Lambda}$ is a positive-semidefinite Hermitian matrix, in the prior without losing the ability to integrate out $\mathbf{R}^{-1}$ analytically. Thus, we define a family of priors

$$\pi(\mathbf{R}^{-1}|\boldsymbol{\Lambda},K) \propto |\mathbf{R}^{-1}|^{-K} \text{etr}\{-\mathbf{R}^{-1}\boldsymbol{\Lambda}\} \quad (20)$$

where $\boldsymbol{\Lambda}$ is a positive-semidefinite Hermitian matrix. Note that this family includes the family of complex Wishart distributions, which is the conjugate prior for $\mathbf{R}^{-1}$, i.e., when the prior is a Wishart distribution, so is the *a posteriori* distribution [30]. The scalar K relates to p and $\tilde{K}$ through $K = p - \tilde{K}$. The family includes both the Jeffreys prior and the uniform prior and enables analytical marginalization of $\mathbf{R}^{-1}$. The conditioning on K and $\boldsymbol{\Lambda}$ denotes the specific choice of prior within the family. As a guideline on how to choose K and $\boldsymbol{\Lambda}$, we recall that if $\tilde{K} > 0$ and $\boldsymbol{\Lambda}$ is positive definite

$$\arg\max_{\mathbf{R}^{-1}} |\mathbf{R}^{-1}|^{\tilde{K}} \text{etr}\{-\mathbf{R}^{-1}\boldsymbol{\Lambda}\} = \left(\frac{\boldsymbol{\Lambda}}{\tilde{K}}\right)^{-1} \quad (21)$$

(see, e.g., [7]). Hence, to attain an increase compared to the Jeffreys prior in a certain region, one can choose K and $\boldsymbol{\Lambda}$ so that $(\boldsymbol{\Lambda}/(K-p))^{-1}$ is in this region. We note that including the factor $\text{etr}\{-\mathbf{R}^{-1}\boldsymbol{\Lambda}\}$ in the prior in principle means changing $\mathbf{Z}\mathbf{Z}^H$ to $\mathbf{Z}\mathbf{Z}^H + \boldsymbol{\Lambda}$ [see (19)], which corresponds to having additional training data $\tilde{\mathbf{Z}}$ for which $\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^H = \boldsymbol{\Lambda}$.

In the light of the previous discussion, there are, of course, various ways to choose K and $\boldsymbol{\Lambda}$. We focus on the choice $\boldsymbol{\Lambda} = \alpha\mathbf{I}$, the main reason being that it has interesting connections to classical estimators. Compared to the Jeffreys prior, this prior is reinforced in a region where all eigenvalues are close to each other. For such a choice, the posterior distribution of $\mathbf{R}^{-1}|\mathbf{Z}$ is complex Wishart according to $\mathcal{CW}_p\{(\mathbf{Z}\mathbf{Z}^H + \alpha\mathbf{I})^{-1}, N_2 - K + p\}$, and the usage of this prior gives [29]

$$E\{\mathbf{R}^{-1}|\mathbf{Z}\} = \left(\frac{\mathbf{Z}\mathbf{Z}^H + \alpha\mathbf{I}}{N_2 - K + p}\right)^{-1}. \quad (22)$$

To prevent the mean of $\mathbf{R}^{-1}$ from being either too large or too small, K and α should be chosen jointly. However, as we will see, in the case of the linear signal model, K is of little importance. Therefore, to simplify things for this scenario, we simply use $K = 0$, rendering a prior that can be normalized to a proper distribution. For notation, let $\pi_\alpha(\mathbf{R}^{-1})$ denote this normalized prior as

$$\pi_\alpha(\mathbf{R}^{-1}) = \mathcal{CW}_p\left(\mathbf{R}^{-1}\left|\frac{1}{\alpha}\mathbf{I}, p\right.\right) = \frac{\text{etr}\{-\mathbf{R}^{-1}\mathbf{I}\alpha\}\,\alpha^{p^2}}{c(p,p)} \quad (23)$$

which by construction is included in the family (20). Note that this prior, in contrast to the ones discussed previously, is normalized to have volume one. Clearly, what we have proposed here is merely an *ad hoc* suggestion that does not solve the problem; the prior is still zero if two or more of the eigenvalues are identical. To study the proposed adjustments, we return to the white noise example.

Example 2: Again, consider the 2-D white noise scenario. A large number of samples $\mathbf{R}_k^{-1}$ is generated from the posterior distribution, given the same training data samples $\mathbf{z}_k$. Although this time, instead of the Jeffreys prior, the prior $\pi_\alpha(\mathbf{R}^{-1})$ is used with $\alpha = 5$. Fig. 1(b) shows the corresponding histogram. In comparison to the former histogram, it has a similar behavior, and thus the same difficulties, even though the peaks are now closer together. ■

As noted, the proposed adjustment does not remove the problem. However, it compensates for it and does so without significantly complicating the solution. Considering the linear signal model and $\pi(\boldsymbol{\theta}) \propto 1$, this prior is of particular interest as the resulting estimator is the standard so-called regularized estimator. Reflecting on this, we hereafter refer to this prior as the regularization prior.

C. Reference Prior

As discussed above, the Jeffreys prior has certain weaknesses in this scenario. An alternative prior is the *Reference prior*,[6] which was derived by Bernardo and Berger in [19]. It was developed to cover for weaknesses of the Jeffreys prior in multidimensional settings. The prior is based on a reasonable information measure, but since it was first proposed, it has been refined in order to cover for new paradoxes as they were discovered. By

[6]Some authors use different notation. In [16], reference prior is used to denote default priors in general. We use the notation as proposed in [31].

now, reference priors have been applied in various settings with good results.

The reference prior has previously been derived for real covariance matrices [9], whereas to our knowledge, it has yet to be derived for the complex data setting. In order to obtain the reference prior for $\mathbf{R}^{-1}$, we basically follow the approach given in [9]. The derivation of the reference prior includes ordering of the parameters in accordance to relevance. This can differ from problem to problem. Following [9], we say that the eigenvalues contained in $\mathbf{D}$ are of greater importance than $\mathbf{B}$.

Theorem 1: The reference prior for the parameter $(\mathbf{D}, \mathbf{B})$ is as follows, providing $\mathbf{D}$ is considered to be more important than $\mathbf{B}$ and the d_i are ordered monotonically (either increasing or decreasing):

$$\pi_{\text{Ref}}(\mathbf{D}, \mathbf{B})\, d\mathbf{D}\, d\mathbf{B} \propto \frac{1}{(\Pi_{i=1}^{p} d_i)\Pi_{i<j}(d_i - d_j)^2}\, d\mathbf{R}^{-1}. \quad (24)$$

Thus, the reference prior can be expressed as

$$\pi_{\text{Ref}}(\mathbf{R}^{-1}) \propto \frac{1}{(\Pi_{i=1}^{p} d_i)\Pi_{i<j}(d_i - d_j)^2}. \quad (25)$$

Proof of Theorem 1: Presenting the proof is beyond the scope of this paper. Those interested should see [32].

From (16) and (24), we observe that the reference prior in the eigenvalue parameterization is given by

$$\pi_{\text{Ref}}(\mathbf{D}, \mathbf{B}) \propto \frac{1}{\Pi_{i=1}^{p} d_i} \quad (26)$$

which, in contrast to the previously discussed priors, is not zero when two or more of the eigenvalues are close to identical. Once again, the white noise example is used for illustration.

Example 3: This time, the samples $\mathbf{R}_k^{-1}$ are generated using the reference prior, and the resulting histogram is plotted in Fig. 1(c). Clearly, this prior does not have the same difficulties regarding the distribution of the eigenvalues. Here, we have many samples in the region close to the true eigenvalues (1,1).■

This prior appears to have a favorable distribution with respect to the eigenvalues d_i. Unfortunately, to our knowledge, there does not exist an analytical solution to the marginalization with respect to $\mathbf{R}^{-1}$ using this prior. It is therefore likely that the implementation of the corresponding estimator is very complex for large dimensions p. A possible framework for implementation is discussed in the next section.

In conclusion, we study the priors $\pi_J(\mathbf{R}^{-1})$, $\pi_\alpha(\mathbf{R}^{-1})$, $\pi(\mathbf{R}^{-1}|\mathbf{\Lambda}, K)$, and $\pi_{\text{Ref}}(\mathbf{R}^{-1})$. We note that $\pi_J(\mathbf{R}^{-1})$ and $\pi_{\text{Ref}}(\mathbf{R}^{-1})$ both require that $N \geq p$ in order for the posterior to be proper and, thereby, applicable. The priors $\pi_J(\mathbf{R}^{-1})$ and $\pi_\alpha(\mathbf{R}^{-1})$ are both special cases of the family $\pi(\mathbf{R}^{-1}|\mathbf{\Lambda}, K)$. All priors in this family have the advantage of offering analytical marginalization. Major drawbacks concerning the eigenvalue distribution of the Jeffreys prior $\pi_J(\mathbf{R}^{-1})$ is somewhat helped by introducing $\pi_\alpha(\mathbf{R}^{-1})$. The reference prior $\pi_{\text{Ref}}(\mathbf{R}^{-1})$ seems to solve the issue but does not, on the other hand, offer an analytical solution. It is understood that when deriving posterior distributions and expected means, these will depend on the choice of prior. To specify which prior is used, we condition on $\mathbf{\Lambda}$ and K if a prior from the family $\pi(\mathbf{R}^{-1}|\mathbf{\Lambda}, K)$ is employed, on J if the Jeffreys prior is used, on α if the regularization prior is used, and on Ref if applying the reference prior.

V. Posterior Distributions

A central component in Bayesian analysis is the posterior distribution for the parameter of interest. Given this information, one can calculate different estimates such as the MMSE and the MAP estimate as well as find measures of posterior uncertainties regarding the parameter. In this section, we show how to calculate the posterior distributions for the priors discussed in the previous section. This is done both for a general signal model and for the specific case of a linear signal model with a uniform prior on $\boldsymbol{\theta}$. As we will see, some parts and scenarios have analytical solutions, whereas others demand numerical approximations.

A. Family of Jeffreys-Like Priors

It is well known that the Jeffreys prior $\pi_J(\mathbf{R}^{-1})$ enables analytical marginalization of the covariance matrix; see [18]. Due to its special structure, this does not only apply to the Jeffreys prior but, in fact, for the whole family described in (20). We start by deriving the integrated likelihood considering this family of priors.

Theorem 2: Suppose we observe the data matrices $\mathbf{X} = \mathbf{S}(\boldsymbol{\theta}) + \mathbf{N}$ and $\mathbf{Z}$ according to the assumptions in Section II. Further assume that $N_1 + N_2 + \text{rank}(\mathbf{\Lambda}) \geq p$. Considering the family of Jeffreys-like priors $\pi(\mathbf{R}^{-1}|\mathbf{\Lambda}, K)$, the integrated likelihood is then given by

$$f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{\Lambda}, K) \propto |\mathbf{Q}|^{-M} \quad (27)$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^H + (\mathbf{X} - \mathbf{S}(\boldsymbol{\theta}))(\mathbf{X} - \mathbf{S}(\boldsymbol{\theta}))^H + \mathbf{\Lambda}$, and $M = N_1 + N_2 - K + p$.

Proof of Theorem 2: See Appendix A.

Finding analytical expressions of the integrated likelihood is indeed advantageous, especially as the needed multidimensional integrals can be very computationally demanding to evaluate numerically. The integrated likelihood itself can, in fact, be of great interest. For instance, doing maximum likelihood estimation using this likelihood is known to be a sound way to treat the nuisance parameters [15].

Even so, the aim of this paper is not to study integrated likelihood methods but to invoke a fully Bayesian approach. Using Theorem 2, we can easily find an expression for the resulting posterior distribution.

Corollary 1: For any signal model $\mathbf{S}(\boldsymbol{\theta})$ and prior distribution $\pi(\boldsymbol{\theta})$, the posterior distribution, using $\pi(\mathbf{R}^{-1}|\mathbf{\Lambda}, K)$ is given by

$$f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \mathbf{\Lambda}, K) = \frac{f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{\Lambda}, K)\pi(\boldsymbol{\theta})}{\int f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \mathbf{\Lambda}, K)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto |\mathbf{Q}|^{-M}\, \pi(\boldsymbol{\theta}). \quad (28)$$

Except for the scaling factor, this gives analytical expressions for the posterior distribution of $\boldsymbol{\theta}$ for any signal model $\mathbf{S}(\boldsymbol{\theta})$, any prior in the family (20), and any prior $\pi(\boldsymbol{\theta})$. The virtues of this result are enforced by the fact that the scaling factor is often not needed. Common examples include settings where a maximum *a posteriori* (MAP) estimate is desired or where a Markov

Chain Monte Carlo (MCMC) method is applied. Despite its usefulness, however, for a large dimension of $\boldsymbol{\theta}$, it is still important to obtain analytical expressions for the posterior so that the need for numerical methods can be avoided. For a general signal model $\mathbf{S}(\boldsymbol{\theta})$ and arbitrary prior $\pi(\boldsymbol{\theta})$, this is not possible. However, for the special case of a linear model and a uniform prior on $\boldsymbol{\theta}$, closed-form expressions (including the scaling factor) can be found for any prior in the family $\pi(\mathbf{R}^{-1}|\boldsymbol{\Lambda}, K)$.

Theorem 3: If $L \leq p$ and $N_2 + \text{rank}(\boldsymbol{\Lambda}) \geq p$, then under the linear model $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$, a prior $\pi(\mathbf{R}^{-1}|\boldsymbol{\Lambda}, K)$ and a uniform prior on $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}) \propto 1$, the posterior distribution of $\boldsymbol{\theta}$ is given by

$$f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\Lambda}, K) = \lambda \frac{|\mathbf{V}|^{N_1}|\mathbf{T}|^{M-L}}{|\mathbf{T} + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^H \mathbf{V} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})|^M}. \quad (29)$$

Here, $M = N_1 + N_2 + p - K$, $\lambda = \pi^{-LN_1} c(L, M)/c(L, M - N_1)$, $\mathbf{V} = \mathbf{H}^H \mathbf{U}^{-1} \mathbf{H}$, $\mathbf{T} = \mathbf{I} + \mathbf{X}^H (\mathbf{U}^{-1} - \mathbf{U}^{-1}\mathbf{H}(\mathbf{H}^H \mathbf{U}^{-1}\mathbf{H})^{-1}\mathbf{H}^H \mathbf{U}^{-1})\mathbf{X}$, $\mathbf{U} = \mathbf{Z}\mathbf{Z}^H + \boldsymbol{\Lambda}$, and $\hat{\boldsymbol{\theta}} = (\mathbf{H}^H \mathbf{U}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{U}^{-1} \mathbf{X}$. Moreover, the posterior mean and covariance of $\boldsymbol{\theta}$ are given by

$$E\{\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\Lambda}, K\} = (\mathbf{H}^H \mathbf{U}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{U}^{-1} \mathbf{X} \quad (30)$$

$$\text{Cov}\{\text{vec}(\boldsymbol{\theta})|\mathbf{X}, \mathbf{Z}, \boldsymbol{\Lambda}, K\} = \frac{1}{M - N_1 - L} \mathbf{T} \otimes \mathbf{V}^{-1}. \quad (31)$$

Proof of Theorem 3: See Appendix B.

We note that the different $\boldsymbol{\theta}_k$'s are correlated in the posterior sense, which is not the case for known noise covariance. The theorem also confirms the previous statement that the actual value of K is of marginal importance, as the posterior mean here is independent of K. Furthermore, to simplify the understanding of the above theorem, we observe that if $N_2 > 0$, a natural estimate of the covariance matrix is $\hat{\mathbf{R}}_{\boldsymbol{\Lambda}} = (1/N_2)\mathbf{U}$. Using this notation, $\mathbf{U}^{-1}$ can be replaced by $(1/N_2)\hat{\mathbf{R}}_{\boldsymbol{\Lambda}}^{-1}$, which is easier to interpret.

By studying some particular choices of prior distributions for $\mathbf{R}^{-1}$, within the family described in (20), we will see that the corresponding MMSE estimates using these priors can be connected to well-known classical estimates. We first consider the Jeffreys prior, which corresponds to $K = p$ and $\boldsymbol{\Lambda} = \mathbf{0}$. For this specific choice, $\mathbf{U} = N_2 \hat{\mathbf{R}}$, and we note that the MMSE estimator using the Jeffreys prior simply reproduces the classical AML, or CE, estimator in (6) and (7). Nevertheless, integrating out the uncertainties in $\mathbf{R}^{-1}$ results in a different posterior covariance for $\boldsymbol{\theta}$. If we study (31) using $K = p$ and $\boldsymbol{\Lambda} = \mathbf{0}$ for the simple case when $N_1 = L = 1$ we see that

$$\text{Cov}\{\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, J\} = \frac{N_2}{N_2 - 1} \frac{1 + \frac{1}{N_2}\sigma_\perp^2}{\mathbf{H}^H \hat{\mathbf{R}}^{-1} \mathbf{H}} \quad (32)$$

where $\sigma_\perp^2 = \mathbf{X}^H (\hat{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\mathbf{H}(\mathbf{H}^H \hat{\mathbf{R}}^{-1}\mathbf{H})^{-1}\mathbf{H}^H \hat{\mathbf{R}}^{-1})\mathbf{X}$. This can be compared to a method where the Bayesian analysis only includes the parameter $\boldsymbol{\theta}$, and the covariance is treated using the classical CE approach

$$\text{Cov}\{\boldsymbol{\theta}|\mathbf{X}, \mathbf{R} = \hat{\mathbf{R}}\} = \frac{1}{\mathbf{H}^H \hat{\mathbf{R}}^{-1} \mathbf{H}}. \quad (33)$$

It is interesting to note that the covariance using the Jeffreys prior is always larger than that of the CE approach. This is natural as the Bayesian treatment of the nuisance parameter acknowledges the uncertainties in $\mathbf{R}^{-1}$.

The posterior covariances above can also be put in connection to the classical performance measures. As mentioned in Section III, these are substantially different measures that evaluate the mean performance of the estimator for a particular scenario. Reed *et al.* [2] derived the covariance for the present setting

$$\text{Cov}\{\hat{\boldsymbol{\theta}}|\mathbf{R}, \boldsymbol{\theta}\} = \frac{N_2}{N_2 - p + 1} \frac{1}{\mathbf{H}^H \mathbf{R}^{-1} \mathbf{H}}. \quad (34)$$

We note that this measure acknowledges the uncertainty in the covariance, as it averages the distribution of the ML estimate over the distribution of the sample covariance estimate. To provide an answer to the question *"How big is the squared magnitude of the error for this particular set of data?"*, the first two measures above can easily be used as estimates. To apply the third one, on the other hand, the underlying parameters describing the scenario must first be estimated. If the sample covariance estimate is used, we see that both (32) and (34) acknowledge the uncertainties in the covariance and, therefore, yield larger estimates than (33). Still, when the uncertainties in the covariance decrease, as $N_2 \to \infty$, all measures coincide.

The AML estimator is not the only classical estimator that can be connected to an MMSE approach using the family of Jeffreys-like Priors. In Section IV-B, we claimed that the usage of the prior $\pi_\alpha(\mathbf{R}^{-1})$ corresponding to $\boldsymbol{\Lambda} = \alpha\mathbf{I}$ and $K = 0$ results in the regularized estimate. This can be verified by studying Theorem 3 for this specific choice of $\boldsymbol{\Lambda}$ and by comparing the corresponding results to (7) and (8). As before, the MMSE estimator is merely a reproduction of a classical estimator. In fact, the posterior distribution of $\boldsymbol{\theta}$ is symmetric around its mean so that many other estimates, such as, for example, the MAP estimate, also render the same solution. The covariance expressions, corresponding to (32) and (33), are also the same, except that $\hat{\mathbf{R}}$ is replaced by $\hat{\mathbf{R}}_\alpha$. The classical performance measure in (34), on the other hand, to our knowledge, does not even have an analytical equivalence for the regularized estimator. This makes the covariance expression in (31) even more useful as one can no longer rely on this classical alternative. Even so, perhaps the main contribution with the connection between the regularized and a Bayesian solution is that this enables a proper treatment of the parameter α, as this parameter is now simply a nuisance parameter in our model. This will be discussed in more detail in Section VI. Note also that the regularized prior $\pi_\alpha(\mathbf{R}^{-1})$, as opposed to the Jeffreys prior, can be applied even when $N_2 = 0$.

B. Reference Prior

In the previous section, we derived analytical expressions for the integrated likelihood. We also saw that for the particular case of a linear model with a uniform prior on $\boldsymbol{\theta}$, the posterior distribution and some of its properties could be derived analytically. Unfortunately, for the reference prior, no analytical solution has been found for the marginalization integral. Instead, we have to resort to the use of numerical methods. One family of suitable methods is the family of MCMC methods. An MCMC method works by generating a Markov chain whose stationary distribution is the distribution of interest [33]. Through this framework, samples $\boldsymbol{\theta}_m$ from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})$ are gener-

ated.[7] These can be used to approximate relevant features of the posterior distribution, such as the posterior mean or variance. Here, we describe, in a general fashion, how such a method can be designed. The algorithm is not designed for any particular signal model but can be applied for any model $\mathbf{S}(\boldsymbol{\theta})$.

Our approach is to generate samples $(\boldsymbol{\theta}_m, \mathbf{R}_m^{-1})$ from the joint distribution $f(\boldsymbol{\theta}, \mathbf{R}^{-1}|\mathbf{X}, \mathbf{Z}, \text{Ref})$. This is accomplished using a Metropolis-Hastings (MH) strategy that is closely related to *Gibbs sampling* [26]. We use the following.

1) Set $m = 0$ and initiate $\boldsymbol{\theta}_0$. We use $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_{AML}$.
2) Generate $\mathbf{R}_m^{-1} \sim f(\mathbf{R}^{-1}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}_{m-1}, \text{Ref})$.
3) Generate $\boldsymbol{\theta}_m \sim f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \mathbf{R}_m^{-1})$.
4) Set $m = m + 1$, and go back to step 2.

The desired samples are the $\boldsymbol{\theta}_m$ samples obtained in step 3. To generate samples in steps 2 and 3, the MH algorithm is generally used; see [33] and below. However, step 3 can, for certain scenarios, be accomplished much more easily as $\boldsymbol{\theta}_m \sim f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \mathbf{R}_m^{-1})$ has a special structure. For instance, a linear signal model and a uniform prior on $\boldsymbol{\theta}$ will result in a Gaussian posterior. MH is the most commonly used mechanism to produce a Markov chain with a specified stationary distribution and is thus an essential part of most MCMC methods. An important part of MH is to propose a new candidate state in the chain. How this is done will have a great influence on the efficiency of the algorithm. In step 2) above, designing this part is particularly complicated in the original parametrization. The reason is that the distribution here is singular when two or more of the eigenvalues are identical, which is clear from (25) and the relation $f(\mathbf{R}^{-1}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}_{m-1}, \text{Ref}) \propto f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_{m-1}, \mathbf{R}^{-1})\, \pi_{\text{Ref}}(\mathbf{R}^{-1})$. One way to solve the problem is to instead parameterize $\mathbf{R}^{-1}$ in terms of its eigenvalues and parameters describing the eigenvectors. The advantage is clearly indicated by comparing (25) and (26). We will address the problem of specifying such a parametrization in a future publication [32]. Nevertheless, once a suitable parametrization is specified, it is possible to present a pseudo-algorithm. Let $\boldsymbol{\rho}$ denote the parameters in the considered parametrization $\mathbf{R}^{-1}(\boldsymbol{\rho})$, and let $\left|\partial \mathbf{R}^{-1}/\partial \boldsymbol{\rho}\right|$ denote the Jacobian for the change of variables. Further, if $q(\boldsymbol{\rho}_k|\boldsymbol{\rho}_{k-1})$ denotes the conditional distribution from which new states are proposed, MH can be described as follows.

1) Set $k = 1$ and initiate $\boldsymbol{\rho}_0$. We use $\boldsymbol{\rho}_0$ so that $\mathbf{R}^{-1}(\boldsymbol{\rho}_0) = \hat{\mathbf{R}}^{-1}$.
2) Generate $\boldsymbol{\rho}_{New} \sim q(\boldsymbol{\rho}_{New}|\boldsymbol{\rho}_{k-1})$.
3) Let

$$\gamma = \frac{f(\mathbf{R}^{-1}(\boldsymbol{\rho}_{New})|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \text{Ref}) q(\boldsymbol{\rho}_{k-1}|\boldsymbol{\rho}_{New}) \left|\frac{\partial \mathbf{R}^{-1}}{\partial \boldsymbol{\rho}}\right|_{\boldsymbol{\rho}=\boldsymbol{\rho}_{k-1}}}{f(\mathbf{R}(\boldsymbol{\rho}_{k-1})|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \text{Ref}) q(\boldsymbol{\rho}_{New}|\boldsymbol{\rho}_{k-1}) \left|\frac{\partial \mathbf{R}^{-1}}{\partial \boldsymbol{\rho}}\right|_{\boldsymbol{\rho}=\boldsymbol{\rho}_{New}}}.$$

4) Take

$$\boldsymbol{\rho}_k = \begin{cases} \boldsymbol{\rho}_{\text{New}}, & \text{with probability } \min(\gamma, 1). \\ \boldsymbol{\rho}_{k-1}, & \text{otherwise.} \end{cases}$$

5) Set $\mathbf{R}_k^{-1} = \mathbf{R}^{-1}(\boldsymbol{\rho}_k)$, $k = k + 1$, and go back to step 2.

We should point out that the use of any valid parametrization will yield the same asymptotic distribution. However, the choice of parametrization, and the means by which one proposes new states in the chain, will affect the convergence rate of the accompanying posterior moment estimates [32], [33].

VI. Treating the Regularization Parameter α

Despite its weaknesses for small training sets, the AML estimator, which is given by (6) and (7), is widely used for the linear model. Two motivations for this are its nice asymptotic properties and its comparatively low complexity. It is important, however, to find alternatives to the AML estimator that not only improve the estimator performance but do so without significantly increasing the complexity. As mentioned, one approach that has proved successful for many scenarios with low sample support is *regularization* [3]. Here, the sample covariance matrix in (7) is replaced by an adjusted estimate given in (8).

One problem with the regularization approach is to find an appropriate value for the parameter α. Traditionally, this is a design parameter set according to some general rule of thumb, based on prior knowledge on the signal and noise environment. In reality, such prior knowledge may be absent, and one would like to replace existing design methods by one that uses the given data in a systematic way. As we will see, such a design can be accomplished by considering a Bayesian approach using the prior $\pi_\alpha(\mathbf{R}^{-1})$. By comparison, it was identified in Section V-A, that the posterior mean, and thus the MMSE estimate of $\boldsymbol{\theta}$, using $\pi_\alpha(\mathbf{R}^{-1})$ is identical to the regularization solution in (6) and (8). Now, when reflecting on how to choose the parameter α in the prior, it should be noted that in this setting, α is merely a nuisance parameter. In fact, α enters the integrated likelihood $f(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \alpha)$ in the same manner as the covariance matrix did in, e.g., (9). Therefore, one faces the same variety of options on how to treat the nuisance parameter; one can either estimate the parameter or integrate out its dependence. The latter is the main focus of this paper, and we start by deriving the MMSE solution using such an approach.

To employ a full Bayesian approach, we consider the posterior distribution $f(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})$ using a prior $\pi(\alpha)$ on α. Marginalization with respect to α renders

$$f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\mathrm{I}_\alpha) = \int f(\boldsymbol{\theta},\alpha|\mathbf{X},\mathbf{Z}) d\alpha = \int f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\alpha) f(\alpha|\mathbf{X},\mathbf{Z})\, d\alpha \tag{35}$$

where I_α is introduced in accordance to previous notation to distinguish this posterior from the others. As a reflection, this approach can be seen as an alternative way to treat the prior on $\mathbf{R}^{-1}$. In essence, we use

$$\pi(\mathbf{R}^{-1}) = \pi(\mathbf{R}^{-1}|\alpha)\pi(\alpha). \tag{36}$$

[7] Note that, different from previous notation, the subindex m denotes the mth vector in a sequence of vectors and not the mth element of a vector.

This way to model the prior distribution is commonly known as a *Hierarchical Bayesian* method [26].

Finding the posterior distribution in (35) requires both $f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\alpha)$ and $f(\alpha|\mathbf{X},\mathbf{Z})$. The first is given by (29), recalling that here, $\boldsymbol{\Lambda}=\mathbf{I}\alpha$ and $K=0$, whereas the second can be expressed as

$$f(\alpha|\mathbf{X},\mathbf{Z})=\frac{f(\mathbf{X},\mathbf{Z}|\alpha)\pi(\alpha)}{\int f(\mathbf{X},\mathbf{Z}|\alpha)\pi(\alpha)\,d\alpha}. \tag{37}$$

Considering the previous discussions, choosing a prior $\pi(\alpha)$ is not a trivial task. For convenience, we use a uniform prior $\pi(\alpha)\propto 1$, hoping that the choice of prior is of minor importance. The appropriateness of this is further discussed below. In conclusion, to derive the posterior distribution, we require the likelihood $f(\mathbf{X},\mathbf{Z}|\alpha)$, which is given in the following theorem.

Theorem 4: If $L\leq p$, then given measurements $\mathbf{X}$ and $\mathbf{Z}$ according to Section II-A, the likelihood $f(\mathbf{X},\mathbf{Z}|\alpha)$, using a uniform prior on $\boldsymbol{\theta}$, is given by

$$f(\mathbf{X},\mathbf{Z}|\alpha)\propto\frac{\alpha^{p^2}}{|\mathbf{U}(\alpha)|^{N_1+N_2+p}\,|\mathbf{T}(\alpha)|^{N_1+N_2+p-L}\,|\mathbf{V}(\alpha)|^{N_1}} \tag{38}$$

where $\mathbf{U}(\alpha)=\mathbf{Z}\mathbf{Z}^H+\alpha\mathbf{I}$, $\mathbf{T}(\alpha)=\mathbf{I}+\mathbf{X}^H(\mathbf{U}^{-1}-\mathbf{U}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{U}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{U}^{-1})\mathbf{X}$, and $\mathbf{V}(\alpha)=\mathbf{H}^H\mathbf{U}^{-1}\mathbf{H}$. The specific dependence of α through $\mathbf{U}$ is here ignored in $\mathbf{T}(\alpha)$ and $\mathbf{V}(\alpha)$ for notational convenience.

Proof of Theorem 4: See Appendix C.

Note that if N_2 is positive, then $\mathbf{U}^{-1}$ can be replaced by $(1/N_2)\hat{\mathbf{R}}_\alpha^{-1}$, where $\hat{\mathbf{R}}_\alpha$ is given in (8). By combining (35), (37), and (38), the posterior can be derived. Unfortunately, no closed-form solution to the integral in (37) has been found. However, as α is real and one-dimensional (1-D), the drawback from having to rely on numerical methods is not that severe. If α is integrated out, the posterior mean is given by

$$\begin{aligned}\mathrm{E}\{\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\mathrm{I}_\alpha\}&=\int\boldsymbol{\theta}\int f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\alpha)f(\alpha|\mathbf{X},\mathbf{Z})\,d\alpha\,d\boldsymbol{\theta}\\&=\int\hat{\boldsymbol{\theta}}(\alpha)f(\alpha|\mathbf{X},\mathbf{Z})d\alpha\end{aligned} \tag{39}$$

where $\hat{\boldsymbol{\theta}}(\alpha)$ is the MMSE estimate of $\boldsymbol{\theta}$ for a given α; see Section V-A and (7) and (8). A natural interpretation is that the mean of the distribution, and therefore the MMSE estimate of $\boldsymbol{\theta}$ as well, is now the average over all regularized solutions weighted by their respective posterior probability.

An alternative to the fully Bayesian approach described in (35)–(39) is to estimate α and proceed by using this value to define a fixed regularization prior. One way to find an appropriate estimate is through the likelihood derived in Theorem 4

$$\hat{\alpha}_{\mathrm{ML}}=\arg\max_\alpha f(\mathbf{X},\mathbf{Z}|\alpha). \tag{40}$$

Again, the complex structure of (38) prevents us from finding an analytical solution. Nevertheless, since it only requires a 1-D maximization, the complexity is minor. Once an estimate is obtained, inference on $\boldsymbol{\theta}$ is based on the posterior distribution $f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\alpha=\hat{\alpha}_{\mathrm{ML}})$, thus rendering the estimate

$$\hat{\boldsymbol{\theta}}_{\hat{\alpha}}=\hat{\boldsymbol{\theta}}(\hat{\alpha}_{\mathrm{ML}}). \tag{41}$$

This method—to estimate a parameter in the prior from data—is known as an *Empirical Bayesian* approach [26]. These estimators have proved to yield good performance for many scenarios. A well-known example is the super efficient estimator commonly referred to as *Steins estimator* [26].

It should be noticed that so far, we have only considered *Bayesian inference* on $\boldsymbol{\theta}$. One may as well use the integrated likelihood directly, for example, through the use of joint ML

$$(\hat{\boldsymbol{\theta}},\hat{\alpha})=\arg\max_{\theta,\alpha} f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\alpha). \tag{42}$$

In fact, there exist a variety of possible approaches, depending on if α and $\boldsymbol{\theta}$ are modeled stochastically or deterministically. Nevertheless, as concluded above, we restrict our studies in this paper to the case when $\boldsymbol{\theta}$ is modeled stochastically.

The analysis above was performed under the assumption of a linear signal model. Regularization is a powerful tool not only for this case but has found applications to numerous methods employing an estimated covariance matrix. For a general signal model, the dependence on $\boldsymbol{\theta}$ is too complicated to integrate out analytically. One way to target this problem is to disregard the information contained in $\mathbf{X}$ while treating α. By doing so, we can form the likelihood function $f(\mathbf{Z}|\alpha)$.

Theorem 5: Given the measurement $\mathbf{Z}$ according to Section II-A, the likelihood $f(\mathbf{Z}|\alpha)$ is given by

$$f(\mathbf{Z}|\alpha)\propto\frac{\alpha^{p^2}}{|\mathbf{Z}\mathbf{Z}^H+\alpha\mathbf{I}|^{N_2+p}}. \tag{43}$$

Proof of Theorem 5: See Appendix D

This can now be used in place of $f(\mathbf{X},\mathbf{Z}|\alpha)$ using any of the methodologies described above, enabling a formal treatment of α, even in the case of a general signal model. One can, for instance, use a fully Bayesian approach similar to (39) or conduct an empirical Bayesian approach as in (41).

We will end this section with a brief study on the choice of prior on α. As mentioned above, $\pi(\alpha)\propto 1$ is chosen in an *ad hoc* manner and is not based on any formal rules. Although it is well known to the Bayesian community that priors on second-order parameters are generally of minor importance [34], it is still of interest to study the robustness to the choice of prior $\pi(\alpha)$ for this specific example. We recall from (39) that the resulting MMSE estimator is the average of $\hat{\theta}(\alpha)$ over the posterior distribution $f(\alpha|\mathbf{X},\mathbf{Z})$. Since $f(\alpha|\mathbf{X},\mathbf{Z})\propto f(\mathbf{X},\mathbf{Z}|\alpha)\,\pi(\alpha)$, we expect the estimate to be sensitive to the choice of prior if the estimate $\hat{\theta}(\alpha)$ varies significantly within the support of $f(\mathbf{X},\mathbf{Z}|\alpha)$, whereas it can be considered insensitive if the estimator is close to constant therein. In an initial study, we consider a small three-dimensional simulation example ($p=3$).

Example 4: A linear model is used with $\mathbf{H}=[1,\,1,\,1]^H$ and $\mathbf{R}=\mathrm{diag}(1,0.1,0.01)$. In the simulation, we employ five training data $N_2=5$ and one primary data $N_1=1$. Fig. 2 shows the integrated likelihood $f(\mathbf{X},\mathbf{Z}|\alpha)$ along with the estimation error $|\hat{\boldsymbol{\theta}}(\alpha)-\boldsymbol{\theta}|$ for two "typical" realizations of $\mathbf{X}$ and $\mathbf{Z}$. Both the error and the likelihood are normalized to yield a maximum value of one. The actual value is not interesting as we want to investigate the fluctuations of $\hat{\boldsymbol{\theta}}(\alpha)$ within the support of the likelihood function. In the first realization, the estimate varies

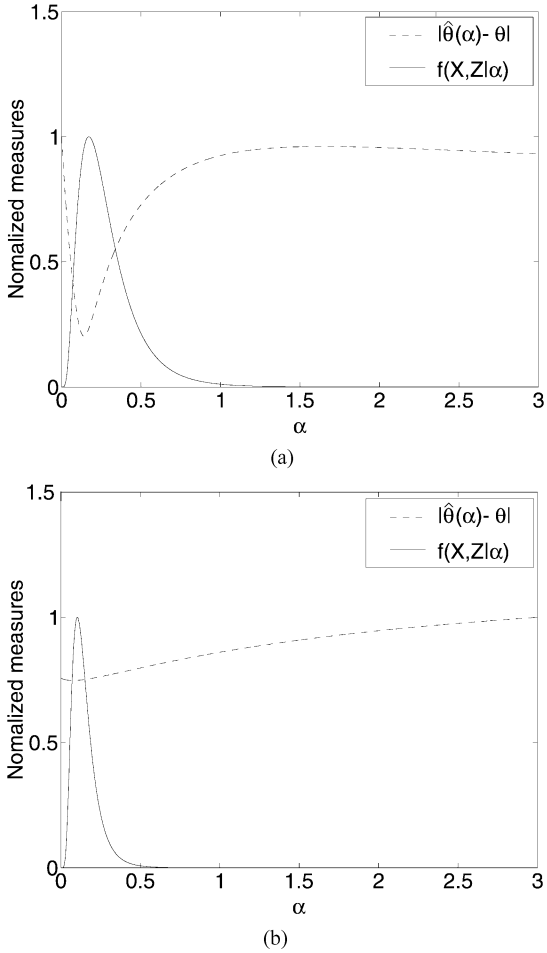


Fig. 2. Shape of the likelihood and the magnitude of the error as functions of α for two arbitrary realizations of $\mathbf{X}$ and $\mathbf{Z}$.

significantly within the support of the likelihood function while it is almost constant in the second.

Thus, even though we cannot draw any general conclusions, the choice of prior appears to have some, but a far from profound, influence. It is beyond the scope of this paper to fully evaluate its effect.

VII. Simulations

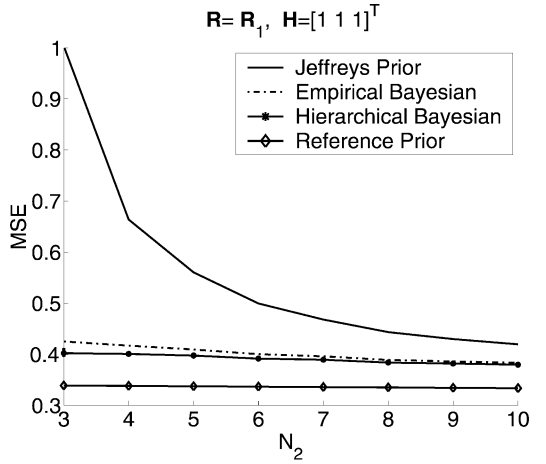
In this section, we study and evaluate the performance of the different methods discussed in the paper. The signal model considered is the linear model in (3). Due to the *Steins effect*, we know that if either N_1 or L is large, then the prior on $\boldsymbol{\theta}$ can have a significant influence on the performance [26]. The main purpose of this paper is not to discuss priors for the parameter of interest $\boldsymbol{\theta}$ but to study and evaluate the treatment of the nuisance parameter $\mathbf{R}^{-1}$. Therefore, we only evaluate the methods for small N_1 and L. In fact, we only study the simplest of all cases: $N_1 = L = 1$. The performance evaluation includes two parts. First, the quality of the different estimates of $\boldsymbol{\theta}$ are measured in terms of the mean square error. Second, the ability to estimate properties of the estimation error is studied. The performances are evaluated in the frequentist sense, i.e., we fix $\mathbf{R}$ and $\boldsymbol{\theta}$ and generate many different data sets $\mathbf{X}$ and $\mathbf{Z}$ to which the different methods are applied. More precisely, all estimators are evaluated using 30 000 sets of data for each scenario.

To investigate the different methods' ability to handle uncertainties in the nuisance parameter $\mathbf{R}^{-1}$, we primarily vary the size of the training data set. Asymptotically, in the amount of training data, all methods perform equally well as the uncertainties can be neglected. Hence, we evaluate the different methods for small training sets implying large uncertainties. As noted in Section IV, the considered set of priors show significant differences when two or more of the eigenvalues of the covariance matrix are close to identical. To further investigate this issue, we find it interesting to study one scenario for which the eigenvalues are identical and one where they are well separated. For simplicity, we choose two low-dimensional examples according to $\mathbf{R}_1 = \text{diag}(1,1,1)$ and $\mathbf{R}_2 = \text{diag}(1, 0.15, 0.05)$. We observe that the prior distributions for the covariance matrix are all invariant to unitary transformations. It is thus sufficient to study the performance for diagonal matrices such as $\mathbf{R}_1$ and $\mathbf{R}_2$. However, unless the noise covariance matrix has eigenvalues with multiplicity, the performance depends on the angle between the signal mode $\mathbf{H}$ and the corresponding eigenvectors. We therefore study two different signal modes: one that aligns with one of the eigenvectors for both cases of $\mathbf{R}$—$\mathbf{H} = \mathbf{H}_2 = [1,0,0]^H$—and one that appears at an angle $\mathbf{H} = \mathbf{H}_1 = [1,1,1]^H$. During the simulations, we employ $\boldsymbol{\theta} = 3$, although we note that the specific value of $\boldsymbol{\theta}$ does not influence the MSE.

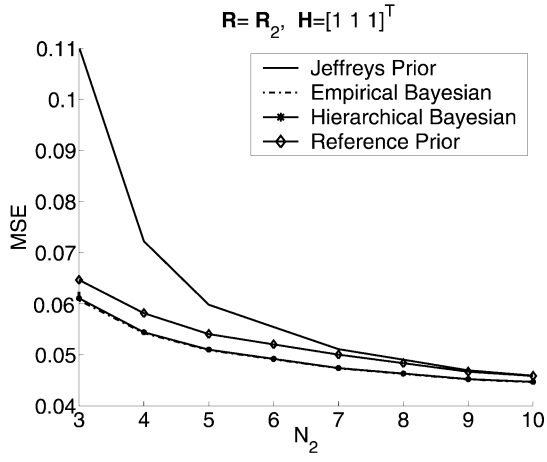
A. Estimator Performance

Consider four different estimators. The first two are the MMSE estimators of $\boldsymbol{\theta}$ using the *Jeffreys prior* and the *Reference prior*, respectively. The first of these is given as the conditional mean in (30), whereas the second is obtained through numerical evaluations; see Section V-B. The remaining two estimators are the results from the usage of the regularization prior. One is the MMSE estimator using $\pi_\alpha(\mathbf{R}^{-1})$ and a uniform prior on α, see (39). We refer to this as the *Hierarchical Bayesian* estimator, whereas the final estimator is the *Empirical Bayesian* method defined in (41). One purpose is to compare Bayesian methods to classical approaches. Nevertheless, as the classical AML estimator here coincides with the MMSE estimator using the Jeffreys Prior, it does not need to be treated separately.

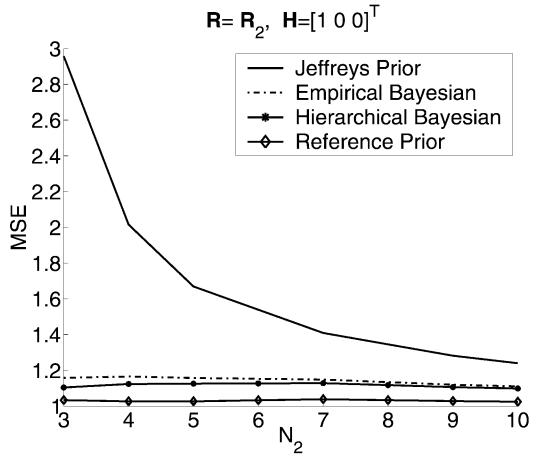
In Fig. 3(a), the performance of the different estimators is shown for the scenario defined by $\mathbf{R} = \mathbf{R}_1$ and $\mathbf{H} = \mathbf{H}_1$. As can be seen, it is possible to significantly improve on the AML estimator (or, equivalently, on the estimator originating from the Jeffreys prior). It is interesting to note that for this scenario, the performances of the other estimators are almost independent of the number of training data. Overall, the estimator corresponding to the reference prior has the best performance. Even for small data sets, its MSE is very close to $\left(\mathbf{H}^H\mathbf{R}^{-1}\mathbf{H}\right)^{-1} = 1/3$, which is the performance of the ML estimator with a known covariance matrix; see [20]. Hence, there is almost no degradation in performance due to uncertainties in $\mathbf{R}$, and even an infinite training set $\mathbf{Z}$ would not improve the estimates. Clearly, this is connected to the appealing properties of the reference prior for identical eigenvalues; see Section IV. In addition, the regularized solutions give good performance with MSEs fairly close to 1/3. Comparing the two, there is a slight



(a)



(b)



(c)

Fig. 3. Performance of the various estimators in terms of Mean Square Error (MSE). (a) White noise scenario $\mathbf{R} = \mathbf{R}_1$ and $\mathbf{H} = \mathbf{H}_1$. (b) Scenario with a noise covariance with significant eigenvalue spread $\mathbf{R} = \mathbf{R}_2$, and $\mathbf{H} = \mathbf{H}_1$. (c) Scenario where, again, $\mathbf{R} = \mathbf{R}_2$, but $\mathbf{H}$ is an eigenvector to $\mathbf{R}$, namely, $\mathbf{H} = \mathbf{H}_2$.

advantage, for small sets of training data, to integrate out the dependence on the regularization parameter through the Hierarchical Bayesian method compared to estimating it through the Empirical Bayesian approach.

Considering the properties of the different priors, the scenario defined by $\mathbf{R} = \mathbf{R}_1$ is very special; see Section IV. Fig. 3(b) presents the corresponding results when the true covariance instead is given by $\mathbf{R}_2$ (we still have $\mathbf{H} = \mathbf{H}_1$). As a comment, we note that this is a much simpler estimation problem as it includes two dimensions in which the noise has significantly lower variance than in the previous case. The MSE of the ML estimator for a known covariance matrix is here 0.037. Compared to the previous scenario, the performances of the different estimators are much more dependent on the amount of training data. Still, the estimator using the Jeffreys prior performs the worst, but the reference prior estimator is no longer the best. Instead, it is the Empirical and Hierarchical Bayesian solutions that show the best performances. For this case, it also appears that integrating out or estimating the regularization parameter yields the same performance.

Previously, we suggested that the angle between the signal mode and the noise eigenvector can influence the overall performance of the noise adaptive estimators. Above, the signal mode $\mathbf{H}_1$ was aligned with a possible eigenvector of $\mathbf{R}_1$, whereas this was not the case for $\mathbf{R}_2$. As a finale, we therefore address the scenario defined by $\mathbf{R} = \mathbf{R}_2$ and $\mathbf{H} = \mathbf{H}_2$ for which this is again true; see Fig. 3(c). By studying the results and comparing them to the corresponding ones in Fig. 3(a), the major trends and properties are all very similar. The reference prior again shows almost perfect performance, as the asymptotic MSE for this scenario is 1. Meanwhile, the regularized solutions again show good performance, whereas the method based on the Jeffreys prior performs significantly worse. The alignment of $\mathbf{H}$ with an eigenvector of $\mathbf{R}$ certainly appears to bring most of the estimation procedures closer to the asymptotic performance. In particular, the reference prior tends to produce very good results. In fact, the superiority of this method does not only extend to the above cases, but additional simulations indicate that the reference prior often performs best out of the derived approaches.

In conclusion, all estimators, except the one based on the Jeffreys prior, are well performing. Taking into account the tremendous difference in complexity, the Empirical and Hierarchical Bayesian solutions, implementing a regularized approach, are natural choices. Because of complexity concerns, the Reference prior only seems to be a suitable option for problems with small dimensionality. However, due to its appealing properties, it serves as a good benchmark if compared with other less-demanding approaches.

B. Error Prediction

In addition to good estimator performance, it is, in many applications, also desirable to gain knowledge regarding the error. Depending on the application, different features of the error can be of interest. Here, we study the ability to estimate the quality of the estimate of $\boldsymbol{\theta}$. We primarily do so using the MSEPF defined in (10).

We consider both Bayesian and Classical alternatives to estimate the squared errors. In the Bayesian framework, we use the MMSE estimate, which now is the posterior variance. In total, the study includes six different methods. Four of these are the natural Bayesian solutions that correspond to the four different estimators that we evaluated in the previous subsection. Again, the Jeffreys prior renders an analytical solution [see

(32)], whereas the posterior variance of the reference prior has to be found by numerical evaluations; see Section V-B. The corresponding estimate for the Empirical Bayesian estimator is given in (31), replacing $\mathbf{\Lambda}$ by $\mathbf{I}\hat{\alpha}$ and setting $K = 0$, whereas the posterior variance using the Hierarchical prior is found through numerical integration with respect to α. Note that all these are estimates of different entities $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2$ as $\hat{\boldsymbol{\theta}}$ differs. The final two are alternative methods to estimate the squared error of the AML estimator described in Section V-A. The first of these is the semi-Bayesian method in which the uncertainties in the covariance matrix are ignored. The expression for this estimator, which we refer to as *Bayesian CE*, is given in (33). Finally, a *Classical CE* approach is taken where the covariance expression in (34) is used while replacing the unknown parameters $\boldsymbol{\theta}$ and $\mathbf{R}$ with the AML and sample covariance estimates, respectively. It is important to note that the expression in (34) actually incorporates the uncertainties in the covariance matrix, even though it is done in a classical manner; see the discussion in Section V-A.

Fig. 4 shows the MSEPF for the same three scenarios that were studied before. Perhaps the most striking property is the poor performance of the reference prior in Fig. 4(b). A more detailed study reveals that the error realizations are all similar in magnitude. Therefore, the scaling factor $\mathrm{Var}\{|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2|\boldsymbol{\theta}, \mathbf{R}\}$ is small, suggesting that the MSEPF is sensitive for a bias in the error estimate (which is the case here). Still, this behavior is very disappointing, and we return to the understanding of this deficiency below. Apart from the reference prior, it appears that the methods that apply the Bayesian treatment of the uncertainties in the covariance matrix offer significant improvements compared to the ones that use the classical approach. The worst of the classical methods is the *Bayesian CE*, which completely ignores these uncertainties. This indicates the value of a proper treatment of the unknown noise color. Again, for large enough training data sets, all methods perform the same; here they all approach one as N_2 goes to infinity.

Of course, comparing performances between estimators that do not have the same task (they estimate different errors) can only be done on a survey level. Moreover, all aspects of these estimators cannot be illustrated by studying one single feature such as the MSEPF. To gain additional insight, we therefore find it interesting to further consider the three measures that do estimate the same error, namely, *Jeffreys, Classical CE*, and *Bayesian CE*. Studying Fig. 4 closely, we note that for small sets of training data, at least for $N_2 = 3$, the method based on the Jeffreys priors gives an MSEPF value of less than one. It even outperforms the estimator $\hat{\sigma}^2 = E\{|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2|\boldsymbol{\theta}, \mathbf{R}\}$ (note that this estimator cannot be implemented since it depends on the unknown parameters $\boldsymbol{\theta}$ and $\mathbf{R}$). To do so, it must have an ability to track the error and "not only" give ensemble properties. To study this in more detail, we consider the scenario given by $\mathbf{R} = \mathbf{R}_2$, $\mathbf{H} = \mathbf{H}_1$, and $N_2 = 3$, which is also interesting because of the poor performance of the reference prior estimator. We want to illustrate the ability to predict the errors by studying the connection between large $\hat{\sigma}^2$ and large $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^2$. We do this by 1) generating a large number of realizations and 2) sorting the error estimates from small to large and permuting the corresponding errors accordingly. The result for the *Classical CE*, *Bayesian CE*, and *Jeffreys* methods are shown in Fig. 5 using 30 000 realizations. As the Bayesian CE estimates are only scaled versions of the Classical CE estimates [see (33) and (34)], they can be plotted in the same figure. Originally, the error curves are very "noisy." To enhance visibility, we average the errors over 100

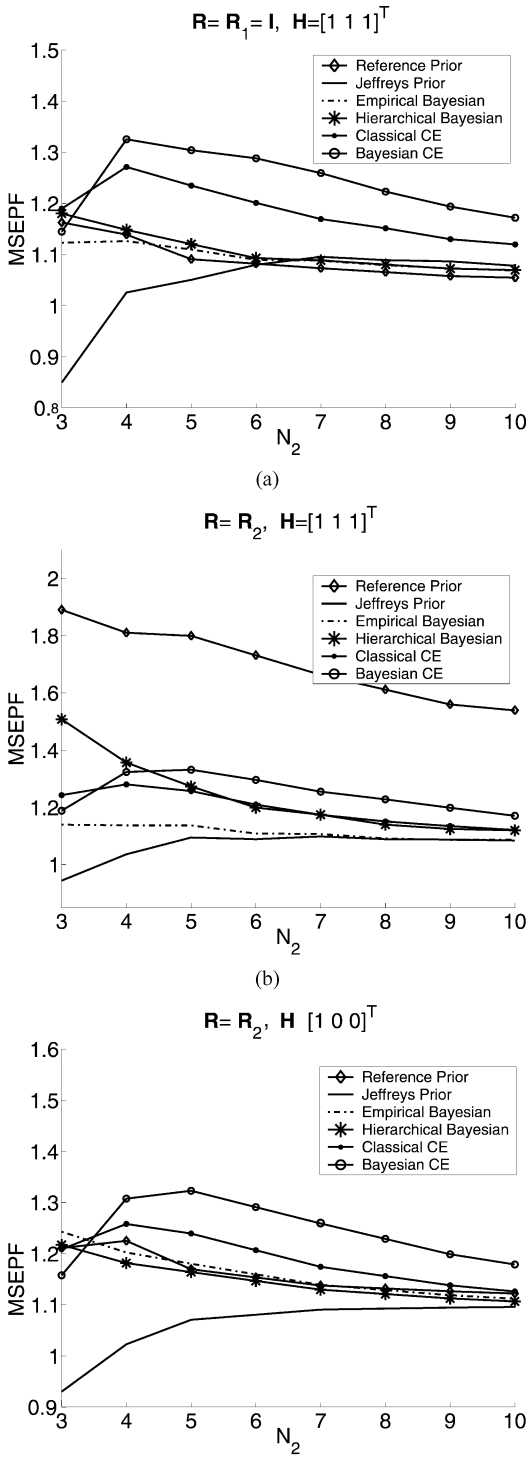


Fig. 4. Performance of the various methods to estimate the error in terms of the MSEPF measure. (a) White noise scenario $\mathbf{R} = \mathbf{R}_1$ and $\mathbf{H} = \mathbf{H}_1$. (b) Scenario with a noise covariance with significant eigenvalue spread $\mathbf{R} = \mathbf{R}_2$, and $\mathbf{H} = \mathbf{H}_1$. (c) Scenario where again, $\mathbf{R} = \mathbf{R}_2$, but $\mathbf{H}$ is an eigenvector to $\mathbf{R}$, namely, $\mathbf{H} = \mathbf{H}_2$.

realizations, where the averaging is done on consecutive samples after they are permuted. At first, this procedure might seem peculiar or even misleading. However, as the sorted estimates vary very slowly, the error curve can be seen as a measure on the average error corresponding to one particular error estimate. In Fig. 5(a), it is apparent that the Bayesian CE method generally underestimates the error. This is natural since a great source of performance degradation (the uncertainties in $\mathbf{R}$) is completely ignored. Nonetheless, the MSEPF [see (10)] is limited as a result of large variance of the error $\mathrm{Var}\left\{|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|^2\middle|\boldsymbol{\theta},\mathbf{R}\right\}$, compared to the entity[8] $E\left\{\left(|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|^2\right)^2\middle|\boldsymbol{\theta},\mathbf{R}\right\}$. The Classical CE estimates, on the other hand, are in a sense of the right magnitude but do not have any visible connection, sample by sample, to the error. In contrast, Fig. 5(b) shows that the estimator based on the Jeffreys prior evidently has an ability to track the error, especially for large errors. This ability is due to a capability of judging the amount of influence the uncertainties in the covariance matrix will have sample by sample. The fact that the Jeffreys estimator slightly underestimates the error can be connected to the discussion in Section IV. There, it was concluded that the Jeffreys prior barely has any mass in the region where two or more of the eigenvalues are close to identical. Since this region typically corresponds to a more difficult scenario, it is natural that this prior therefore leads to a slight underestimation of the error. The overall behavior of this estimator is still very satisfying and shows the possible advantage of the Bayesian treatment of the nuisance parameters. The ability to track the magnitude of the errors is, of course, also present for training sets larger than $N_2 = p$ but is then less evident. In fact, for sufficiently large training sets, all methods give approximately the same error prediction for *all* realizations. This stems from the fact that there is far less uncertainty in $\mathbf{R}^{-1}$ as N_2 becomes large; see the discussion below.

For completeness, Fig. 6 shows the corresponding sorted plots for the remaining methods. Note again that we cannot directly compare the ability of the different methods since they estimate errors of different parameter estimates. However, there are some properties that can be identified. Primarily, we observe that these Bayesian approaches tend to have significantly less ability to track the error. Only a small ability can be observed for the largest errors. This can be explained by the fact that here, there is far less uncertainty to consider. In principle, the error has two components: one due to the noise in $\mathbf{X}$ and one due to lack of knowledge about the covariance. It is clear that if we have perfect knowledge regarding $\mathbf{R}^{-1}$, it is not possible to track the magnitude of the error since the error is due to the noise outcome only. Then, the estimate of the error should ideally be the same for all realization: a straight line. What can be tracked, at least to some extent, is, instead, the amount of available and relevant information about the noise color. Since the performance of the Empirical Bayesian, Hierarchical Bayesian, and the Reference prior are all fairly close to the asymptotic performance even for small training data sets, the influence from the uncertainties in $\mathbf{R}^{-1}$ on the estimate is

[8]For this particular scenario, and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{AML}}$, the estimator $\hat{\sigma}^2 = 0$ would render MSEPF $= 1.22$.

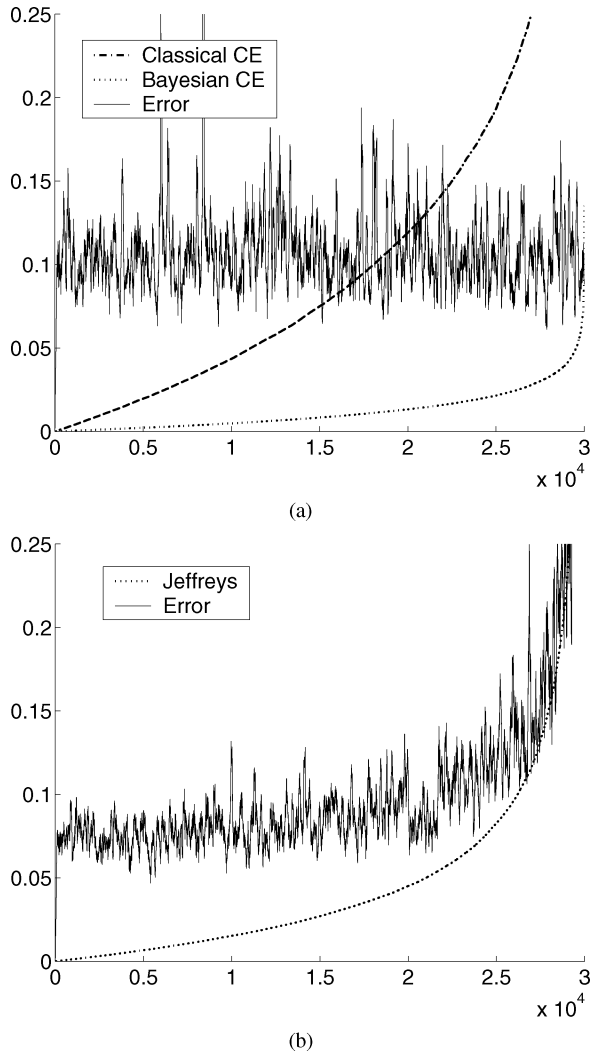
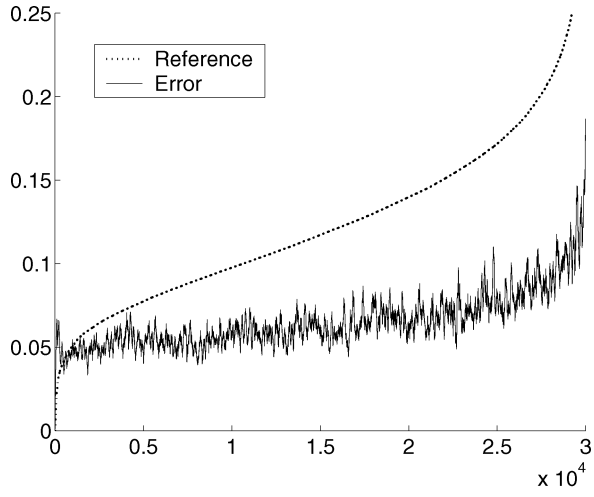


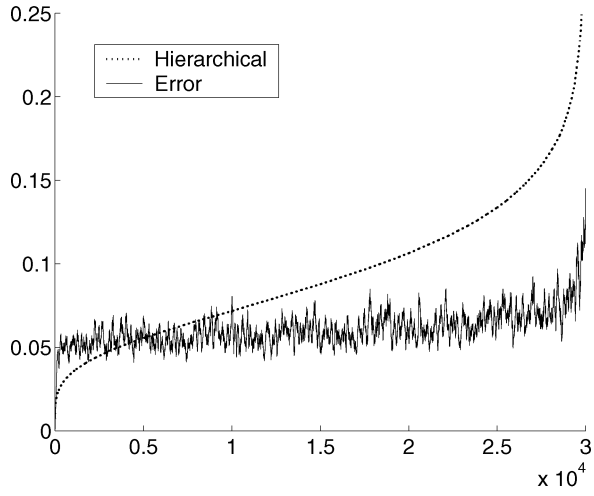
Fig. 5. Sorted error estimates and the corresponding errors. (a) Classical CE and the Bayesian CE approach. (b) Jeffreys prior method.

comparably small. Thereby, the error is again primarily due to the noise outcome, and it is understood that the possibility to track the magnitude of the error is only marginal. From Fig. 6, we also identify the tendency of the Reference prior to overestimate the error. Combined with a small normalization factor $\mathrm{Var}\left\{|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}|^2\middle|\boldsymbol{\theta},\mathbf{R}\right\}$, this renders the large MSEPF values observed in Fig. 4. Empirical observations indicate that the overestimation of the error is caused by having too much density in the region where the eigenvalues for $\mathbf{R}^{-1}$ are small. This suggests that the reference prior could be adjusted, and perhaps replaced by, for instance, $\pi(\mathbf{D},\mathbf{B}) \propto 1$. To investigate and evaluate such corrections to the reference prior is, however, left for future work.

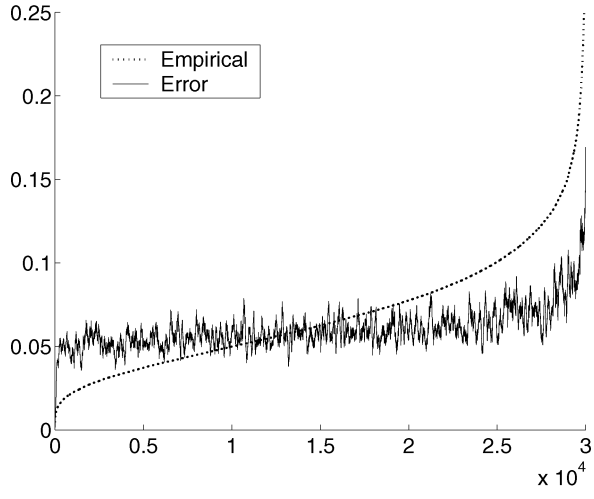
Comparing the performance of the different methods, both in terms of the ability to estimate the parameter and the ability to track the error, we observe that the method based on the Jeffreys prior shows superior performance in terms of tracking the magnitude of the error. Meanwhile, the other methods are significantly better at actually estimating the parameter itself. Although a good parameter estimate leaves less room for tracking the error, we would like to design a strategy that performs well in both categories. In most applications,



(a)



(b)



(c)

Fig. 6. Sorted error estimates and the corresponding errors. (a) Reference prior solution. (b) Hierarchical Bayesian method. (c) Empirical Bayesian estimates.

however, the ability to successfully estimate the parameter of interest is of much greater importance than approximating the error. Therefore, for the priors considered in this paper, the regularized solutions, along with the Reference prior, are preferred. However, due to computational issues, the regularized solutions are recommended.

VIII. Conclusions

We studied a Bayesian treatment of the unknown covariance matrix $\mathbf{R}$, making up the nuisance parameter in our model. This was performed for a number of different priors $\pi(\mathbf{R}^{-1})$, some of which enabled analytical marginalization. For the linear model, we showed the connection between the AML estimator and the MMSE estimator using the Jeffreys prior on $\pi(\mathbf{R}^{-1})$ and a uniform prior on the parameter of interest. By a slight adjustment to the Jeffreys prior, we also found a connection between the corresponding MMSE estimate and the regularized solution. This in turn enabled a formal treatment of the regularization parameter as it could be treated as a nuisance parameter. In addition to these, a prior called the Reference prior was derived. This prior resolved certain inconsistencies present in the previous proposals but, on the other hand, did not provide analytical solutions. Besides offering a tool to derive powerful estimators, simulations also showed that the Bayesian approach provides an enhanced possibility to estimate properties of the estimation error by treating the nuisance parameter in a sound way.

Appendix A
Proof of Theorem 2

The measurements are i.i.d. according to $f(\mathbf{x}_k|\mathbf{R}^{-1},\boldsymbol{\theta}) = \mathcal{CN}_p(\mathbf{x}_k|\mathbf{s}_k(\boldsymbol{\theta}_k),\mathbf{R})$ and $f(\mathbf{z}_k|\mathbf{R}^{-1}) = \mathcal{CN}_p(\mathbf{z}_k|\mathbf{0},\mathbf{R})$. Then, in matrix notation

$$f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\mathbf{R}^{-1}) = \pi^{-p(N_1+N_2)}|\mathbf{R}^{-1}|^{N_1+N_2} \times \text{etr}\left\{-\mathbf{R}^{-1}\left(\mathbf{Z}\mathbf{Z}^H + (\mathbf{X}-\mathbf{S}(\boldsymbol{\theta}))(\mathbf{X}-\mathbf{S}(\boldsymbol{\theta})^H\right)\right\}.$$

Using the notation

$$\mathbf{Q} = \mathbf{Z}\mathbf{Z}^H + (\mathbf{X}-\mathbf{S}(\boldsymbol{\theta}))(\mathbf{X}-\mathbf{S}(\boldsymbol{\theta}))^H + \boldsymbol{\Lambda}$$

and $M = N_1 + N_2 - K + p$, we can now marginalize with respect to $\mathbf{R}^{-1}$

$$\begin{aligned} f(\mathbf{X},&\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\Lambda},K) \\ &= \int f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\mathbf{R}^{-1})\pi(\mathbf{R}^{-1}|\boldsymbol{\Lambda},K)\,d\mathbf{R}^{-1} \\ &\propto \int \pi^{-p(N_1+N_2)}|\mathbf{R}^{-1}|^{M-p}\text{etr}(-\mathbf{R}^{-1}\mathbf{Q})\,d\mathbf{R}^{-1} \\ &\propto \int CW_p(\mathbf{R}^{-1}|\mathbf{Q}^{-1},M)d\mathbf{R}^{-1}c(p,M)|\mathbf{Q}|^{-M}\pi^{-p(N_1+N_2)} \\ &\propto |\mathbf{Q}|^{-M}. \end{aligned} \tag{44}$$

Implicitly, this assumes that $\mathbf{Q}$ is invertible. This is true, with probability 1, if and only if $N_1 + N_2 + \text{rank}\{\boldsymbol{\Lambda}\} \geq p$.

Appendix B
Proof of Theorem 3

By Theorem 2

$$f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\Lambda},K) \propto |\mathbf{Q}|^{-M} \tag{45}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^H + \boldsymbol{\Lambda} + (\mathbf{X}-\mathbf{H}\boldsymbol{\theta})(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})^H$, and $M = N_1 + N_2 - K + p$. To derive the determinant $|\mathbf{Q}|$, we define $\mathbf{U} = \mathbf{Z}\mathbf{Z}^H + \boldsymbol{\Lambda}$ and assume that $\mathbf{U}$ is invertible. This is true with probability 1 if and only if $N_2 + rank(\boldsymbol{\Lambda}) \geq p$. Employing

the general determinant relation $|\mathbf{I}+\mathbf{AB}| = |\mathbf{I}+\mathbf{BA}|$ then gives

$$|\mathbf{Q}| = |\mathbf{U}|\,|\mathbf{I}+(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})^H\mathbf{U}^{-1}(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})|. \quad (46)$$

Let $\hat{\boldsymbol{\theta}} = (\mathbf{H}^H\mathbf{U}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{U}^{-1}\mathbf{X}$, $\mathbf{V} = \mathbf{H}^H\mathbf{U}^{-1}\mathbf{H}$, and $\mathbf{T} = \mathbf{I}+\mathbf{X}^H(\mathbf{U}^{-1}-\mathbf{U}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{U}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{U}^{-1})\mathbf{X}$. Then

$$\begin{aligned}(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})^H\mathbf{U}^{-1}(\mathbf{X}-\mathbf{H}\boldsymbol{\theta}) &= \mathbf{X}^H\mathbf{U}^{-1}\mathbf{X}+\boldsymbol{\theta}^H\mathbf{H}^H\mathbf{U}^{-1}\mathbf{H}\boldsymbol{\theta}\\ &\quad -\boldsymbol{\theta}^H\mathbf{H}^H\mathbf{U}^{-1}\mathbf{X}-\mathbf{X}^H\mathbf{U}^{-1}\mathbf{H}\boldsymbol{\theta}\\ &= \mathbf{T}-\mathbf{I}+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \quad (47)\end{aligned}$$

and hence

$$\begin{aligned}|\mathbf{Q}| &= |\mathbf{U}|\,|\mathbf{T}+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})|\\ &= |\mathbf{U}|\,|\mathbf{T}|\,|\mathbf{I}+\mathbf{T}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})|\\ &= |\mathbf{U}|\,|\mathbf{T}|\,|\mathbf{I}+\mathbf{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\mathbf{T}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H|\\ &= |\mathbf{U}|\,|\mathbf{T}|\,|\mathbf{V}|\,|\mathbf{V}^{-1}+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\mathbf{T}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H|. \quad (48)\end{aligned}$$

Thus, for a uniform prior on $\boldsymbol{\theta}$, (45) renders

$$f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K) \propto |\mathbf{V}^{-1}+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\mathbf{T}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H|^{-M}. \quad (49)$$

Now introduce

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\mathbf{T}^{-1/2}. \quad (50)$$

According to the change-of-variable formula

$$f(\tilde{\boldsymbol{\theta}}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K) = f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K)\left|\frac{\partial\boldsymbol{\theta}}{\partial\tilde{\boldsymbol{\theta}}}\right| \quad (51)$$

where $|\partial\boldsymbol{\theta}/\partial\tilde{\boldsymbol{\theta}}| = |\mathbf{T}|^L$; see, e.g., [2]. Hence

$$f(\tilde{\boldsymbol{\theta}}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K) \propto |\mathbf{V}^{-1}+\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H|^{-M}. \quad (52)$$

We only miss the scaling factor for the distribution in (52) to be completely known. The easiest way to find it is to identify the expression as proportional to a well-known distribution. Before we do so, we need some more notation. Let $\tilde{\boldsymbol{\theta}}_k$ be the kth column of $\tilde{\boldsymbol{\theta}}$ so that $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{N_1}]$, and let $\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}$ be the vectorized version of $\tilde{\boldsymbol{\theta}}$, i.e., let $\tilde{\boldsymbol{\theta}}_{\mathrm{vec}} = [\tilde{\boldsymbol{\theta}}_1^H, \tilde{\boldsymbol{\theta}}_2^H, \ldots, \tilde{\boldsymbol{\theta}}_{N_1}^H]^H$. Now, (52) can be identified as the somewhat peculiar distribution achieved when the columns $\tilde{\boldsymbol{\theta}}_k$ are i.i.d. according to $\mathcal{CN}_L(0,\boldsymbol{\Sigma})$, whereas $\boldsymbol{\Sigma}^{-1} \sim \mathcal{CW}_L(\mathbf{V}, M-N_1)$. To clarify this, the columns are only independent given the covariance matrix $\boldsymbol{\Sigma}$, and will become dependent through the outcome of $\boldsymbol{\Sigma}^{-1}$. To justify this and to derive the distribution

$$\begin{aligned}f(\tilde{\boldsymbol{\theta}}) &= \int \Pi_{k=1}^{N_1}\mathcal{CN}_L(\tilde{\boldsymbol{\theta}}_k|0,\boldsymbol{\Sigma})\mathcal{CW}_L(\boldsymbol{\Sigma}^{-1}|\mathbf{V},N_2-K+p)d\boldsymbol{\Sigma}^{-1}\\ &= \int \frac{|\mathbf{V}^{-1}|^{M-N_1}|\boldsymbol{\Sigma}^{-1}|^{M-L}\mathrm{etr}\left\{-(\mathbf{V}^{-1}+\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H)\boldsymbol{\Sigma}^{-1}\right\}}{\pi^{L\,N_1}\,c(L,M-N_1)}d\boldsymbol{\Sigma}^{-1}\\ &= \int \mathcal{CW}_L(\boldsymbol{\Sigma}^{-1}|(\mathbf{V}^{-1}+\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H)^{-1},M)\,d\boldsymbol{\Sigma}^{-1}\\ &\quad \times \frac{\pi^{-L\,N_1}c(L,M)}{c(L,M-N_1)}\,|\mathbf{V}^{-1}|^{M-N_1}\,|\mathbf{V}^{-1}+\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H|^{-M} \quad (53)\\ &= \frac{\pi^{-L\,N_1}c(L,M)}{c(L,M-N_1)}\,|\mathbf{V}^{-1}|^{M-N_1}\,|\mathbf{V}^{-1}+\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H|^{-M} \quad (54)\\ &= \frac{\pi^{-L\,N_1}c(L,M)}{c(L,M-N_1)}\,|\mathbf{V}|^{N_1}\,|\mathbf{I}+\tilde{\boldsymbol{\theta}}^H\mathbf{V}\tilde{\boldsymbol{\theta}}|^{-M}. \quad (55)\end{aligned}$$

From (54), it is clear that this probability density function is proportional to $f(\tilde{\boldsymbol{\theta}}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K)$ in (52) and, hence, identical. By combining this result with (51) and denoting $\lambda = \pi^{-L\,N_1}c(L,M)/c(L,M-N_1)$, we obtain the desired result

$$f(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z},\boldsymbol{\Lambda},K) = \lambda\frac{|\mathbf{V}|^{N_1}|\mathbf{T}|^{M-L}}{|\mathbf{T}+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})|^M}. \quad (56)$$

Furthermore, a convenient way to derive the mean and covariance matrix of $\boldsymbol{\theta}_{\mathrm{vec}}$ is first to derive them for $\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}$ and then use (50). The mean and covariance of $\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}$, for a given covariance matrix $\boldsymbol{\Sigma}$, are given as $\mathbf{0}$ and $\mathbf{I}\otimes\boldsymbol{\Sigma}$, respectively. As the mean is independent of $\boldsymbol{\Sigma}$, we have $E\{\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}\} = \mathbf{0}$. Further, since $\boldsymbol{\Sigma}^{-1} \sim \mathcal{CW}_L(\mathbf{V}, M-N_1)$, then $E\{\boldsymbol{\Sigma}\} = (1/(M-N_1-L))\mathbf{V}^{-1}$; see, e.g., [29]. The covariance matrix of $\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}$ is thereby given as

$$\mathrm{Cov}\{\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}\} = E\{\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}^H\} = \frac{1}{M-N_1-L}\mathbf{I}\otimes\mathbf{V}^{-1}. \quad (57)$$

Moreover, writing (50) in vectorized form gives

$$\tilde{\boldsymbol{\theta}}_{\mathrm{vec}} = (\mathbf{T}^{-1/2}\otimes\mathbf{I})(\boldsymbol{\theta}_{\mathrm{vec}}-\hat{\boldsymbol{\theta}}_{\mathrm{vec}}). \quad (58)$$

In conclusion, we obtain the posterior mean

$$E\{\boldsymbol{\theta}_{\mathrm{vec}}\} = \hat{\boldsymbol{\theta}}_{\mathrm{vec}} + (\mathbf{T}^{-1/2}\otimes\mathbf{I})^{-1}E\{\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}\} = \hat{\boldsymbol{\theta}}_{\mathrm{vec}} \quad (59)$$

and covariance matrix

$$\begin{aligned}\mathrm{Cov}\{\boldsymbol{\theta}_{\mathrm{vec}}\} &= E\{(\boldsymbol{\theta}_{\mathrm{vec}}-\hat{\boldsymbol{\theta}}_{\mathrm{vec}})(\boldsymbol{\theta}_{\mathrm{vec}}-\hat{\boldsymbol{\theta}}_{\mathrm{vec}})^H\}\\ &= (\mathbf{T}^{-1/2}\otimes\mathbf{I})^{-1}E\{\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}\tilde{\boldsymbol{\theta}}_{\mathrm{vec}}^H\}(\mathbf{T}^{-1/2}\otimes\mathbf{I})^H \quad (60)\\ &= (\mathbf{T}^{-1/2}\otimes\mathbf{I})^{-1}\frac{1}{M-N_1-L}\mathbf{I}\otimes\mathbf{V}^{-1}(\mathbf{T}^{-1/2}\otimes\mathbf{I})^{-H} \quad (61)\\ &= \frac{1}{M-N_1-L}(\mathbf{T}^{1/2}\otimes\mathbf{I})(\mathbf{I}\otimes\mathbf{V}^{-1})(\mathbf{T}^{1/2}\otimes\mathbf{I})\\ &= \frac{1}{M-N_1-L}\,\mathbf{T}\otimes\mathbf{V}^{-1}. \quad (62)\end{aligned}$$

Appendix C
Proof of Theorem 4

First, we derive $f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\alpha)$. In principle, this is a special case of $f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\Lambda},K)$ given in Theorem 2. However, Theorem 2 only gives the expression up to a proportionality factor. The reason is that the prior and, hence, the integrated likelihood is not proper in general. As this factor depends on α, we need to calculate it. Using the notation $\mathbf{Q}(\alpha) = \mathbf{Z}\mathbf{Z}^H + (\mathbf{X}-\mathbf{H}\boldsymbol{\theta})(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})^H + \mathbf{I}\alpha$ and the prior $\pi_\alpha(\mathbf{R}^{-1}) = \mathrm{etr}\{-\mathbf{R}^{-1}\alpha\}\alpha^{p^2}/c(p,p)$, we have (63)–(66), shown at the top of the next page. Now, the likelihood $f(\mathbf{X},\mathbf{Z}|\alpha)$ is achieved through marginalization of $\boldsymbol{\theta}$

$$\begin{aligned}&f(\mathbf{X},\mathbf{Z}|\alpha)\\ &= \int f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \propto \int \frac{\alpha^{p^2}}{|\mathbf{Q}(\alpha)|^{N_1+N_2+p}}\,d\boldsymbol{\theta} \quad (67)\\ &\propto \frac{\alpha^{p^2}}{|\mathbf{U}(\alpha)|^{N_1+N_2+p}}\int |\mathbf{T}(\alpha)+(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\alpha)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})|^{-(N_1+N_2+p)}d\boldsymbol{\theta}. \quad (68)\end{aligned}$$

$$f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\alpha) = \int f(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta},\alpha,\mathbf{R}^{-1})\pi_\alpha(\mathbf{R}^{-1})\,d\mathbf{R}^{-1} \tag{63}$$

$$= \int \frac{|\mathbf{R}^{-1}|^{N_1+N_2}\text{etr}\{-\mathbf{R}^{-1}(\mathbf{Z}\mathbf{Z}^H + (\mathbf{X}-\mathbf{H}\boldsymbol{\theta})(\mathbf{X}-\mathbf{H}\boldsymbol{\theta})^H + \mathbf{I}\alpha)\}}{\pi^{p(N_1+N_2)}\,c(p,p)} d\mathbf{R}^{-1}\alpha^{p^2} \tag{64}$$

$$= \int CW_p(\mathbf{R}^{-1}|\mathbf{Q}(\alpha)^{-1}, N_1+N_2+p)\,d\mathbf{R}^{-1} \frac{\alpha^{p^2} c(p,N_1+N_2+p)}{\pi^{p(N_1+N_2)}\,c(p,p)\,|\mathbf{Q}(\alpha)|^{N_1+N_2+p}} \tag{65}$$

$$= \frac{c(p,N_1+N_2+p)\,\alpha^{p^2}}{\pi^{p(N_1+N_2)}\,c(p,p)\,|\mathbf{Q}(\alpha)|^{N_1+N_2+p}}. \tag{66}$$

Here, the last step follows from (48), and we specifically note that $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{T}$ depend on α. Using that the expression in (56) integrates to one along with $K = 0$ renders

$$\int |\mathbf{T}(\alpha) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^H\mathbf{V}(\alpha)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})|^{-(N_1+N_2+p)} d\boldsymbol{\theta} \propto |\mathbf{V}(\alpha)|^{-N_1}|\mathbf{T}(\alpha)|^{-(N_1+N_2+p-L)}. \tag{69}$$

In conclusion, by inserting the result into (67), we obtain the desired result

$$f(\mathbf{X},\mathbf{Z}|\alpha) \propto \frac{\alpha^{p^2}}{|\mathbf{U}(\alpha)|^{N_1+N_2+p}\,|\mathbf{T}(\alpha)|^{N_1+N_2+p-L}\,|\mathbf{V}(\alpha)|^{N_1}}. \tag{70}$$

Appendix D
Proof of Theorem 5

Using the prior $\pi_\alpha(\mathbf{R}^{-1}) = \text{etr}\{-\mathbf{R}^{-1}\alpha\}\alpha^{p^2}/c(p,p)$ gives

$$f(\mathbf{Z}|\alpha) = \int f(\mathbf{Z}|\alpha,\mathbf{R}^{-1})\pi_\alpha(\mathbf{R}^{-1})\,d\mathbf{R}^{-1}$$

$$= \int \frac{|\mathbf{R}^{-1}|^{N_2}\text{etr}\{-\mathbf{R}^{-1}(\mathbf{Z}\mathbf{Z}^H+\mathbf{I}\alpha)\}}{\pi^{pN_2}\,c(p,p)} d\mathbf{R}^{-1}\alpha^{p^2} \tag{71}$$

$$= \int \mathcal{CW}_p(\mathbf{R}^{-1}|(\mathbf{Z}\mathbf{Z}^H+\mathbf{I}\alpha)^{-1}, N_2+p)\,d\mathbf{R}^{-1} \times \frac{\alpha^{p^2}c(p,N_2+p)}{\pi^{pN_2}\,c(p,p)\,|\mathbf{Z}\mathbf{Z}^H+\mathbf{I}\,\alpha|^{N_2+p}} \tag{72}$$

$$= \frac{c(p,N_2+p)\,\alpha^{p^2}}{\pi^{pN_2}\,c(p,p)\,|\mathbf{Z}\mathbf{Z}^H+\mathbf{I}\,\alpha|^{N_2+p}}. \tag{73}$$

References

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] I. S. Reed, J. D. Mallett, and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-10, pp. 853–863, Nov. 1974.

[3] H. L. Van Trees, *Array Processing: Detection and Estimation Theory IV*. New York: Wiley, 2002.

[4] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[5] J. R. Guerci, J. S. Goldstein, and I. S. Reed, "Optimal and adaptive reduced-rank STAP," *IEEE Trans. Aerosp. Elect. Syst.*, vol. 36, no. 2, pp. 647–663, Apr. 2000.

[6] R. Klemm, *Space-Time Adaptive Processing: Principles and Applications*. London, U.K.: Inst. Elect. Eng., 1998.

[7] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Ann. Math. Stat.*, vol. 34, pp. 152–177, Mar. 1963.

[8] R. O. Schmidt, "A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1981.

[9] R. Yang and J. Berger, "Estimation of a covariance matrix using the reference prior," *Ann. Statist.*, vol. 22, no. 3, pp. 1195–1211, 1994.

[10] A. Edelman, "Eigenvalues and Condition Numbers of Random Matrices," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1989.

[11] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[12] E. J. Kelly, "An adaptive detection algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-22, pp. 115–127, Mar. 1986.

[13] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, no. 9, pp. 963–974, Sep. 1982.

[14] H. Li, P. Stoica, and J. Li, "Computationally efficient maximum likelihood estimation of structured covariance matrices," *IEEE Trans. Signal Process.*, vol. 47, no. 5, pp. 1314–1323, May 1999.

[15] J. Berger, B. Liseo, and R. Wolpert, "Integrated likelihood methods for eliminating nuisance parameters (with discussion)," *Stat. Sci.*, vol. 14, pp. 1–28, 1999.

[16] R. E. Kass and L. Wasserman, "Formal rules of selecting prior distributions: A review and annotated bibliography," *J. Amer. Statist. Assoc.*, vol. 91, pp. 1343–1370, 1996.

[17] H. Jeffreys, *Theory of Probability*. Oxford, U.K.: Oxford Univ. Press, 1961.

[18] K. M. Wong, J. P. Reilly, Q. Wu, and S. Qiao, "Estimation of the directions of arrival of signals in unknown correlated noise. i. The map approach and its implementation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 2007–2017, Aug. 1992.

[19] J. Berger and J. M. Bernardo, "On the development of reference priors," in *Bayesian Statistics*. Oxford, U.K.: Oxford Univ. Press, 1992, vol. 4.

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

[21] C. D. Richmond, "Derived pdf of maximum likelihood signal estimator which employs an estimated noise covariance," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 305–315, Feb. 1996.

[22] T. F. Ayoub and A. M. Haimovich, "Modified GLRTsignal detection algorithm'," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 3, pp. 810–818, Jul. 2000.

[23] J. Hiemstra, M. Weippert, S. Goldstein, and T. Pratt, "Application of the l-curve technique to loading level determination in adaptive beamforming," in *Proc. 36th Asilomar Conf., Signals, Syst., Comput.*, Monterey, CA, Nov. 2002.

[24] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria'," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.

[25] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noise sinusoids via reversible jump mcmc," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, Oct. 1999.

[26] C. P. Robert, *The Bayesian Choice*. New York: Springer-Verlag, 2001.

[27] J. O. Berger and T. Sellke, "Testing a point null hypothesis: The irreconcilability of p values and evidence," *J. Amer. Stat. Assoc.*, vol. 82, pp. 112–122.

[28] L. Svensson and M. Lundberg, "Analytical expression for the posterior distribution of signals in colored Gaussian noise," in *Proc. 36th Asilomar Conf., Signals Syst., Comput.*, Monterey, CA, Nov. 2002.

[29] J. A. Tague and C. I. Caldwell, "Expectations of useful complex Wishart forms," *Multidim. Syst. Signal Process.*, vol. 5, pp. 263–279, 1994.
[30] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. London, U.K.: Chapman and Hall, 1997.
[31] J. M. Bernardo, "Reference posterior distributions for bayes inference (with discussion)," *J. R. Stat. Soc. B*, vol. 41, pp. 113–147, 1979.
[32] L. Svensson and M. Lundberg, "The reference prior for complex covariance matrices with efficient implementation strategies," *IEEE Trans. Signal Process.*, to be published.
[33] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.
[34] I. J. Good, "Rational decisions," *J. R. Statist. Soc. Ser.B*, vol. 14, pp. 107–114, 1952.

Lennart Svensson was born in Älvängen, Sweden, on August 12, 1976. He received the M.S. degree in electrical engineering in 1999 and the Ph.D. degree in 2004, both from Chalmers University of Technology, Gothenburg, Sweden.

His research interests concern Bayesian estimation and detection in different signal processing applications.

Magnus Lundberg received the M.Sc. degree in computer science and engineering in 1998 from Luleå University of Technology (LTU), Luleå, Sweden, and the Ph.D degree in signal processing in 2003 from the School of Electrical and Computer Engineering, Chalmers University of Technology, Gothenburg, Sweden.

Since September 2003, he has been an Assistant Professor with the Department of Computer Science and Electrical Engineering, LTU. Since October 2003, he has been on a leave of absence to visit the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins. His research interests lie in statistical signal processing and how it applies to digital communications, radar, sonar, land-mine detection, and high-level power estimation in CMOS architectures.

Dr Lundberg has received several awards and grants for his research. These include the 1998 "MD110 User-group award" for the best Masters Thesis in the telecommunication area in Sweden that year and a postdoctoral scholarship award from the Swedish Research Council.