Model Mismatch Trade-offs in LMMSE Estimation

Martin Hellkvist, Ayça Özçelikkale,
Dept. of Electrical Engineering, Uppsala University, Sweden
{Martin.Hellkvist, Ayca.Ozcelikkale}@angstrom.uu.se

***Abstract*—We consider a linear minimum mean squared error (LMMSE) estimation framework with model mismatch where the assumed model order is smaller than that of the underlying linear system which generates the data used in the estimation process. By modelling the regressors of the underlying system as random variables, we analyze the average behaviour of the mean squared error (MSE). Our results quantify how the MSE depends on the interplay between the number of samples and the number of parameters in the underlying system and in the assumed model. In particular, if the number of samples is not sufficiently large, neither increasing the number of samples nor the assumed model complexity is sufficient to guarantee a performance improvement.**

***Index Terms*—Model uncertainty, missing features, robustness.**

I. Introduction

Described in several text books and used in a wide range of applications, the linear minimum mean square error (LMMSE) estimator [1], [2] is one of the fundamental estimation methods of signal processing. Being a Bayesian estimation approach, the parameters of interest are modeled as random variables with some joint probability density function (pdf), based on some background knowledge. The LMMSE estimator is the optimal estimator out of all the possible linear (more precisely affine) estimators in terms of minimizing the mean squared error (MSE), and it only depends on the mean and covariances. If the assumed covariance matrices are inaccurate, which is generally the case for real-world problems, then the performance of the computed LMMSE estimator can be suboptimal. In this article, we focus on characterizing such performance degradation.

We consider the underlying system $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{y}$ is the observed output vector, $\boldsymbol{A}$ is the matrix of regressors, $\boldsymbol{v}$ is some unknown noise vector and the vector $\boldsymbol{x}$ denotes the unknown model parameters which we want to estimate. We model $\boldsymbol{x}$ and $\boldsymbol{v}$ as random vectors, and propose an LMMSE estimation framework which allows us to systematically study the MSE when only a subset $\boldsymbol{A}_S$ of the columns in $\boldsymbol{A}$ are available for estimation. In particular, the mismatched estimator is based on the assumed system $\boldsymbol{y} = \boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{z}$, where the assumed number of unknowns (length of $\boldsymbol{x}_S$) is smaller than the number of unknowns in the underlying system (length of $\boldsymbol{x}$). We model the regressors in $\boldsymbol{A}$ as random variables and derive an analytical expression for the expected MSE of the low order LMMSE estimator, over the distribution of $\boldsymbol{A}$.

M. Hellkvist and A. Özçelikkale acknowledges the support from Swedish Research Council under grant 2015-04011.

A range of methods have been proposed for robustness against uncertainties or model flaws in the LMMSE estimation. Methods to deal with covariance matrix uncertainties have been presented in [3]–[5], and in [6] the effect of having missing features, i.e., unknowns, in the underlying model was investigated. Robustness have been also investigated under a classical estimation framework where the unknown is modelled as deterministic, such as for uncertainties in the regressors [7]. Further model mismatch trade-offs in classical estimation settings have been studied, focusing on the relationship between model size and number of observations [8], [9]. In our setup, only parts of the regressors are available for estimation, and the respective models on $\boldsymbol{x}_S$ and $\boldsymbol{w}$ do not match the underlying models $\boldsymbol{x}$ and $\boldsymbol{v}$, constituting a hybrid setting with uncertain regressors and a model mismatch in the unknowns and the noise.

We study the average MSE performance under a model mismatch and with an isotropic Gaussian model on the regressors. Our contributions can be summarized as follows: i) Our analytical results show that the MSE depends on the respective signal powers of $\boldsymbol{x}$, $\boldsymbol{x}_S$ and $\boldsymbol{v}$, but not on the general covariance structure of the unknowns $\boldsymbol{x}$. ii) These results quantify how the MSE heavily depends on the relation between the number of samples, the underlying and the assumed model orders: If the number of samples is not sufficiently large, then the performance is not guaranteed to improve by increasing the number of samples or the assumed model complexity. In particular, lowering the assumed model order can improve the performance even when the number of samples is larger than the number of unknowns in the underlying system.

The rest of the paper is organized as follows: In Section II, we provide the problem formulation. In Section III, we present and discuss our main analytical results, which are numerically verified in Section IV. Conclusions are summarized in Section V.

Notation: We denote the Moore-Penrose pseudoinverse and the transpose of a matrix $\boldsymbol{A}$ as $\boldsymbol{A}^+$ and $\boldsymbol{A}^{\mathrm{T}}$, respectively. The $p \times p$ identity matrix is denoted as $\boldsymbol{I}_p$. The Euclidean norm and trace operator are denoted by $\|\cdot\|$ and $\mathrm{tr}(\cdot)$, respectively. We use the notation $\underset{\boldsymbol{x}}{\mathbb{E}}$ or $\mathbb{E}_{\boldsymbol{x}}$ to emphasize that the expectation is taken with respect to the random variable $\boldsymbol{x}$. For two column vectors $\boldsymbol{z}$, $\boldsymbol{w}$, we denote their covariance matrix by $\boldsymbol{K}_{\boldsymbol{z}\boldsymbol{w}} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{w}}[(\boldsymbol{z}-\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}])(\boldsymbol{w}-\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}])^{\mathrm{T}}]$. For auto-covariance matrices, we write the subscript only once: $\boldsymbol{K}_{\boldsymbol{z}} = \boldsymbol{K}_{\boldsymbol{z}\boldsymbol{z}}$.

II. Problem Statement

A. The Underlying System

The observations $\boldsymbol{y}$ come from the following linear system

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$ denotes the unknowns, $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$ denotes the vector of observations, $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \cdots \mid \boldsymbol{a}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ denotes the known matrix of regressors, and $\boldsymbol{v} = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times 1}$ denotes the unknown noise. Here, $\boldsymbol{x}$ and $\boldsymbol{v}$ are modeled as zero-mean uncorrelated random vectors. Note that $\boldsymbol{a}_i^{\mathrm{T}}$ denotes the ith row of $\boldsymbol{A}$, i.e., the regressors corresponding to the observation $y_i = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} + v_i$.

Consider the class of linear estimators, i.e., the estimators such that the estimate $\hat{\boldsymbol{x}}$ is a linear function of the vector of observations $\boldsymbol{y}$, with $\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{y}$ where $\boldsymbol{W} \in \mathbb{R}^{p\times n}$. The mean squared error (MSE) associated with $\boldsymbol{W}$ is given by

$$J(\boldsymbol{W}) = \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{x} - \boldsymbol{W}\boldsymbol{y}\|^2\right]. \tag{2}$$

Under the linear model in (1), $J(\boldsymbol{W})$ is found as

$$\begin{aligned} J(\boldsymbol{W}) &= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}}\left[\|\boldsymbol{x} - \boldsymbol{W}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{v})\|^2\right] && (3)\\ &= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{v}}\left[\|(\boldsymbol{I}_p - \boldsymbol{W}\boldsymbol{A})\boldsymbol{x} - \boldsymbol{W}\boldsymbol{v}\|^2\right] && (4)\\ &= \mathrm{tr}((\boldsymbol{I}_p - \boldsymbol{W}\boldsymbol{A})\boldsymbol{K}_{\boldsymbol{x}}(\boldsymbol{I}_p - \boldsymbol{W}\boldsymbol{A})^{\mathrm{T}} + \boldsymbol{W}\boldsymbol{K}_{\boldsymbol{v}}\boldsymbol{W}^{\mathrm{T}}). && (5) \end{aligned}$$

The linear minimum MSE (LMMSE) estimator, i.e., the matrix $\boldsymbol{W}$ that minimizes the MSE $J(\boldsymbol{W})$ over all $\boldsymbol{W} \in \mathbb{R}^{p\times n}$, is given by [1]

$$\boldsymbol{W}_O = \boldsymbol{K}_{\boldsymbol{x}\boldsymbol{y}}\boldsymbol{K}_{\boldsymbol{y}}^{+} = \boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{K}_{\boldsymbol{v}})^{+}, \tag{6}$$

where we have used the fact that under (1), we have $\boldsymbol{K}_{\boldsymbol{x}\boldsymbol{y}} = \boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{K}_{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{K}_{\boldsymbol{v}}$. In (6) we have used the Moore-Penrose pseudoinverse, rather than the ordinary inverse, which, as discussed in [1, Theorem 3.2.3] will minimize the MSE regardless of whether $\boldsymbol{K}_{\boldsymbol{y}}$ is singular or not.

B. Model Mismatch and Assumed Model

In this paper, our focus is on estimation under a model mismatch. In particular, we consider the case that the LMMSE estimator relies on an incorrect signal model such that i) only a subset of the unknowns $\boldsymbol{x}_i$ are assumed to be present in the system equation; ii) the assumed covariances are possibly inconsistent with the underlying system in (1). Let this subset of $\boldsymbol{x}$ be denoted by $\boldsymbol{x}_S \in \mathbb{R}^{p_S\times 1}$ and its complement (i.e., the elements of $\boldsymbol{x}$ that are not in $\boldsymbol{x}_S$) by $\boldsymbol{x}_C \in \mathbb{R}^{p_C\times 1}$, where $p = p_S + p_C$. Let $\boldsymbol{A}_S \in \mathbb{R}^{n\times p_S}$ and $\boldsymbol{A}_C \in \mathbb{R}^{n\times p_C}$ denote the submatrices of $\boldsymbol{A}$ consisting of the columns corresponding to the indices that are in $\boldsymbol{x}_S$ and in $\boldsymbol{x}_C$, respectively.

The estimator uses the following partial model

$$\boldsymbol{y} = \boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{z}, \tag{7}$$

where $\boldsymbol{x}_S$ and the noise $\boldsymbol{z}$ are assumed to be uncorrelated and zero-mean, and $\boldsymbol{A}_S$ is known. Here, the respective *assumed* covariance matrices for $\boldsymbol{x}_S$ and the noise $\boldsymbol{z}$ are given by $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S}$ and $\hat{\boldsymbol{K}}_z$. We have used the notation $\hat{\boldsymbol{K}}$ to emphasize that these covariance matrices are not necessarily the same as the ones that can be derived from (1). Hence, there is a model mismatch between (7) and (1). According to (7), other covariance matrices of interest are given by

$$\hat{\boldsymbol{K}}_{\boldsymbol{x}_S\boldsymbol{y}} = \mathop{\mathbb{E}}_{\boldsymbol{x}_S,\boldsymbol{y}}\left[\boldsymbol{x}_S\boldsymbol{y}^{\mathrm{T}}\right] = \mathop{\mathbb{E}}_{\boldsymbol{x}_S,\boldsymbol{z}}\left[\boldsymbol{x}_S(\boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{z})^{\mathrm{T}}\right] = \hat{\boldsymbol{K}}_{\boldsymbol{x}_S}\boldsymbol{A}_S^{\mathrm{T}}, \tag{8}$$

$$\hat{\boldsymbol{K}}_{\boldsymbol{y}} = \mathop{\mathbb{E}}_{\boldsymbol{x}_S,\boldsymbol{z}}\left[(\boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{z})(\boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{z})^{\mathrm{T}}\right] = \boldsymbol{A}_S\hat{\boldsymbol{K}}_{\boldsymbol{x}_S}\boldsymbol{A}_S^{\mathrm{T}} + \hat{\boldsymbol{K}}_{\boldsymbol{z}}. \tag{9}$$

Let $\hat{\boldsymbol{x}}_S = \boldsymbol{W}_S\boldsymbol{y}$ be an estimate of $\boldsymbol{x}_S$, where $\boldsymbol{W}_S \in \mathbb{R}^{p_S\times n}$. Then the corresponding MSE for $\boldsymbol{x}_S$ is given by

$$J_S(\boldsymbol{W}_S) = \mathop{\mathbb{E}}_{\boldsymbol{x}_S,\boldsymbol{y}}\left[\|\boldsymbol{x}_S - \hat{\boldsymbol{x}}_S\|^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{x}_S,\boldsymbol{y}}\left[\|\boldsymbol{x}_S - \boldsymbol{W}_S\boldsymbol{y}\|^2\right]. \tag{10}$$

An explicit expression for $J_S(\boldsymbol{W}_S)$ is provided in (17). The corresponding LMMSE estimator, assuming (7), is given by

$$\hat{\boldsymbol{x}}_S = \boldsymbol{W}_S\boldsymbol{y} = \hat{\boldsymbol{K}}_{\boldsymbol{x}_S\boldsymbol{y}}\hat{\boldsymbol{K}}_{\boldsymbol{y}}^{+}\boldsymbol{y} = \hat{\boldsymbol{K}}_{\boldsymbol{x}_S}\boldsymbol{A}_S^{\mathrm{T}}(\boldsymbol{A}_S\hat{\boldsymbol{K}}_{\boldsymbol{x}_S}\boldsymbol{A}_S^{\mathrm{T}} + \hat{\boldsymbol{K}}_{\boldsymbol{z}})^{+}\boldsymbol{y}, \tag{11}$$

We note that the estimator in (11) would be the true LMMSE estimator if the observations $\boldsymbol{y}$ were in fact generated by the model in (7). However, this is not the case. Here, $\boldsymbol{y}$ actually comes from the underlying system in (1), hence the true LMMSE estimate of $\boldsymbol{x}_S$, minimizing the MSE in (10), is

$$\hat{\boldsymbol{x}}_S = \boldsymbol{K}_{\boldsymbol{x}_S}\boldsymbol{A}_S^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{x}}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{K}_{\boldsymbol{v}})^{+}\boldsymbol{y}. \tag{12}$$

To summarize our setting, $\boldsymbol{y}$ is generated by the system in (1), while the estimation is performed under the assumption that $\boldsymbol{y}$ is generated by (7). Hence, the LMMSE estimator in (11) is used instead of the correct estimator in (12). In other words, we consider LMMSE estimation under a model mismatch.

In order to take into account the part of $\boldsymbol{x}$ that is not estimated in this partial setting, i.e., $\boldsymbol{x}_C$, we also define the MSE associated with the whole vector $\boldsymbol{x}$ under $\boldsymbol{W}_S$ as

$$J(\boldsymbol{W}_S) = J_S(\boldsymbol{W}_S) + \mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C}). \tag{13}$$

Note that the subscript S in $J_S(\cdot)$ emphasizes that the error is over $\boldsymbol{x}_S$ whereas $J(\cdot)$ refers to the error in the whole vector $\boldsymbol{x}$. Here, $J(\boldsymbol{W}_S)$ corresponds to the error associated with estimating $\boldsymbol{x}_S$ with $\boldsymbol{W}_S$ while setting the estimate of $\boldsymbol{x}_C$ to $E[\boldsymbol{x}_C] = 0$.

C. Expected MSE over Regressors

We are interested in the average behaviour of the MSE of the partial LMMSE estimator in (11) over regressor matrices $\boldsymbol{A}$. We model $\boldsymbol{a}_i$'s as independent and identically distributed (i.i.d.) Gaussian random vectors, i.e., $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{K}_{\boldsymbol{a}})$, $\forall\, i$ with $\boldsymbol{K}_{\boldsymbol{a}} = \boldsymbol{I}_p$. The *expected MSE* over the distribution of $\boldsymbol{A}$'s is given by

$$\varepsilon_S(p_S, n) = \mathop{\mathbb{E}}_{\boldsymbol{A}}\left[J_S(\boldsymbol{W}_S)\right] = \mathop{\mathbb{E}}_{\boldsymbol{A}}\left[J_S(\hat{\boldsymbol{K}}_{\boldsymbol{x}_S\boldsymbol{y}}\hat{\boldsymbol{K}}_{y}^{+})\right]. \tag{14}$$

Note that this is the expected MSE associated with $\boldsymbol{x}_S$. Here $\boldsymbol{W}_S$ is a function of $\boldsymbol{A}$ (more precisely a function of $\boldsymbol{A}_S$, a submatrix of $\boldsymbol{A}$), and $\boldsymbol{y}$ varies with $\boldsymbol{A}$. We are interested in how the MSE varies for different choices of p_S, i.e., the number of estimated parameters, and n, i.e., the number of samples in $\boldsymbol{y}$. Hence, $\varepsilon_S(p_S, n)$ is defined as a function of these variables.

Here, we analyze the MSE from the perspective of repeated experiments using different matrices $\boldsymbol{A}$, hence we here model $\boldsymbol{A}$ as a random matrix. Nevertheless, note that while doing the LMMSE estimation, $\boldsymbol{A}$ and $\boldsymbol{A}_S$ are known in (6) and (11), respectively.

We similarly define the expected MSE associated with the whole vector $\boldsymbol{x}$ as

$$\varepsilon(p_S, n) = \mathop{\mathbb{E}}_{\boldsymbol{A}}[J(\boldsymbol{W}_S)] = \varepsilon_S(p_S, n) + \mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C}), \tag{15}$$

which in addition to (14) takes into account the power of the signal $\boldsymbol{x}_C$ that is disregarded by the assumed model (7).

As a part of our analysis of $\varepsilon(p_S, n)$, we also compare it to the expected MSE associated with the full LMMSE estimator

$$\bar{\varepsilon} = \mathop{\mathbb{E}}_{\boldsymbol{A}}[J(\boldsymbol{W}_O)] = \mathop{\mathbb{E}}_{\boldsymbol{A}}[J(\boldsymbol{K}_{\boldsymbol{xy}}\boldsymbol{K}_{\boldsymbol{y}}^{+})]. \tag{16}$$

where $\boldsymbol{W}_O$ is the estimator in (6).

III. Expected MSE under a Model Mismatch

We now provide an explicit expression $J_S(W_S)$ of (10). Plugging in y from the underlying model in (1),

$$\begin{aligned} J_S(\boldsymbol{W}_S) &= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{v}}\left[\|\boldsymbol{x}_S - \boldsymbol{W}_S(\boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{A}_C\boldsymbol{x}_C + \boldsymbol{v})\|^2\right] && (17)\\ &= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{v}}\left[\|(\boldsymbol{I}_{p_S} - \boldsymbol{W}_S\boldsymbol{A}_S)\boldsymbol{x}_S - \boldsymbol{W}_S\boldsymbol{A}_C\boldsymbol{x}_C - \boldsymbol{W}_S\boldsymbol{v}\|^2\right] && (18)\\ &= \mathrm{tr}\big((\boldsymbol{I}_{p_S} - \boldsymbol{W}_S\boldsymbol{A}_S)\boldsymbol{K}_{\boldsymbol{x}_S}(\boldsymbol{I}_{p_S} - \boldsymbol{W}_S\boldsymbol{A}_S)^{\mathrm{T}} \\ &\quad + \boldsymbol{W}_S\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C}\boldsymbol{A}_C^{\mathrm{T}}\boldsymbol{W}_S^{\mathrm{T}} + \boldsymbol{W}_S\boldsymbol{K}_{\boldsymbol{v}}\boldsymbol{W}_S^{\mathrm{T}} && (19)\\ &\quad - 2\boldsymbol{W}_S\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C\boldsymbol{x}_S}(\boldsymbol{I}_{p_S} - \boldsymbol{W}_S\boldsymbol{A}_S)^{\mathrm{T}}\big). \end{aligned}$$

The following result describes the generalization error associated with the partial LMMSE estimator in (11):

Theorem 1. *With $\boldsymbol{K}_{\boldsymbol{a}} = \boldsymbol{I}_p$, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S} = \boldsymbol{I}_{p_S}$ and $\hat{\boldsymbol{K}}_{\boldsymbol{z}} = 0$, i.e., the noise $\boldsymbol{z}$ is assumed to be zero, the partial LMMSE estimator in* (11) *has the expected MSE*

$$\begin{aligned} \varepsilon_S(p_S, n) = \mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_S})\left(1 - \frac{\min\{p_S, n\}}{p_S}\right) \\ + \gamma\left(\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C}) + \frac{1}{n}\mathrm{tr}(\boldsymbol{K}_v)\right), \end{aligned} \tag{20}$$

with γ defined as

$$\gamma = \begin{cases} \frac{p_S}{n - p_S - 1} & \text{for } p_S < n - 1, & (21a)\\ \frac{n}{p_S - n - 1} & \text{for } p_S > n + 1, & (21b)\\ +\infty & \text{otherwise,} & (21c) \end{cases}$$

where $p = p_S + p_C$.

Proof: See Section VI-A.

Theorem 1 quantifies the dependence of the expected MSE ε_S on the individual powers of $\boldsymbol{x}_S$, $\boldsymbol{x}_C$ and the noise $\boldsymbol{v}$, i.e., $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_S})$, $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C})$ and $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{v}})$. It also reveals that the error does not depend on the covariance between $\boldsymbol{x}_S$ and $\boldsymbol{x}_C$, or on the general structure of $\boldsymbol{K}_{\boldsymbol{v}}$.

Effect of γ: The factor γ, hence ε_S, can take extremely large values if the number of samples n is too close to the number of estimated parameters p_S. We observe that if both $\boldsymbol{x}_C$ and $\boldsymbol{v}$ are identically zero, then γ does not affect the MSE, however this is generally not the case.

We continue the discussion of the behaviour of ε_S by considering the following scenarios of n versus p_S: **i)** $n > p_S$, and **ii)** $n < p_S$.

i) $n > p_S$: Here, the MSE component from $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_S})$ is constantly zero, and γ in (21a) decreases monotonically with an increasing n. Hence, if the noise level per sample does not increase with the number of samples, i.e., if $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{v}})/n$ doesn't increase with n, then the MSE monotonically decreases with increasing n. Regarding the MSE's dependency on p_S, we will show in Corollary 2 that if n is not large enough, then the expected MSE is not guaranteed to improve with p_S, under some additional constraints.

ii) $n < p_S$: The result in Theorem 1 shows that the performance is not guaranteed to improve by having more samples. In (21b), we see that for $n < p_S$, γ increases with n, hence the expected MSE can also increase with n. In particular, as we will illustrate with numerical examples in Section IV, the power in $\boldsymbol{x}_S$ must be significantly larger than the combined powers of $\boldsymbol{x}_C$ and $\boldsymbol{v}$, in order for the MSE to decrease as n increases. For such small n, it is also not immediately apparent which choice of p_S gives the lowest MSE. This insight is illustrated in the numerical examples in Section IV.

The following corollary is a special case of Theorem 1 where the powers in $\boldsymbol{x}_S$ and $\boldsymbol{x}_C$ are directly proportional to p_S and p_C and the noise level per sample is constant:

Corollary 1. *Consider the setting of Theorem 1, with $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_S}) = \sigma_x^2 p_S > 0$, $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C}) = \sigma_x^2 p_C \geq 0$ and $\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{v}}) = n\sigma_v^2 > 0$, then the partial LMMSE estimator in* (11) *has the following expected MSE:*

$$\varepsilon_S(p_S, n) = \sigma_x^2(p_S - \min\{p_S, n\}) + \gamma(\sigma_x^2 p_C + \sigma_v^2). \tag{22}$$

Proof: This result is readily obtained by plugging in the respective assumptions on $\boldsymbol{K}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{x}_S}, \boldsymbol{K}_{\boldsymbol{x}_C}$ and $\boldsymbol{K}_{\boldsymbol{v}}$ into (20).

Under the given additional assumptions, Corollary 1 gives a clear characterization of the dependence of the MSE on the respective dimensions of $\boldsymbol{x}_S$ and $\boldsymbol{x}_C$, the number of samples n, and the power levels σ_x^2 and σ_v^2.

While our Bayesian problem formulation is different than the classical estimation setting of [9], the result in this corollary describes the same phenomenon as studied in [9, Theorem 2.1]. In [9], according to least-squares estimation setting [2, Ch.8], the performance is measured by the residuals $y_i - \boldsymbol{a}_i^{\mathrm{T}}\hat{\boldsymbol{x}}$, i.e., the error made when predicting $\boldsymbol{y}$ with $\hat{\boldsymbol{x}}$. On the other hand, in this paper we focus on the error associated with the estimate $\hat{\boldsymbol{x}}$. Nevertheless, the expected error for the estimate of a new y_j satisfies $\mathbb{E}_{y_j,\boldsymbol{x},\hat{\boldsymbol{x}},\boldsymbol{a}_j,\boldsymbol{v}}[(y_j - \boldsymbol{a}_j^{\mathrm{T}}\hat{\boldsymbol{x}})^2] = \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}}}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] + \sigma_v^2$, with $\boldsymbol{a}_j \sim \mathcal{N}(0, I_p)$. Moreover, the least-squares (LS) estimator in [9] matches our estimator under the assumptions of Theorem 1, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{A}_S^{+}\boldsymbol{y}$.

Corollary 2. *Consider the setting of Corollary 1. Let $n > p + 1$. Then the expected MSE $\varepsilon(p_S, n)$ decreases monotonically with p_S if*

$$n > p + \sigma_v^2/\sigma_x^2 + 1. \tag{23}$$

Furthermore, $\varepsilon(p_S, n)$ increases monotonically with p_S if

$$n < p + \sigma_v^2/\sigma_x^2 + 1. \tag{24}$$

Proof: This result can be found by treating p_S as a continuous variable and taking the derivative of $\varepsilon(p_S, n) = \varepsilon_S(p_S, n) + \mathrm{tr}(K_{x_C})$ w.r.t. p_S, and solving the inequalities $\partial\varepsilon/\partial p_S < 0$ and $\partial\varepsilon/\partial p_S > 0$, for n.

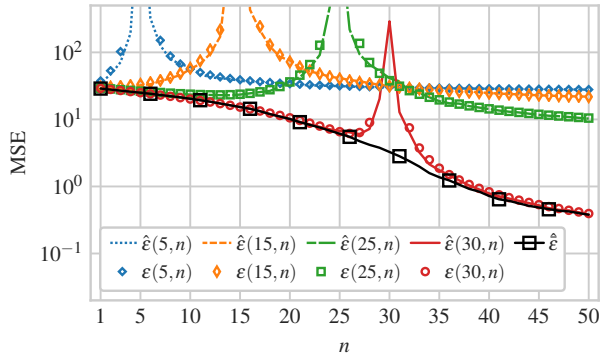

Fig. 1: Empirical and analytical MSE versus the number of samples, for the partial LMMSE estimator in the high SNR case *(S1)*.

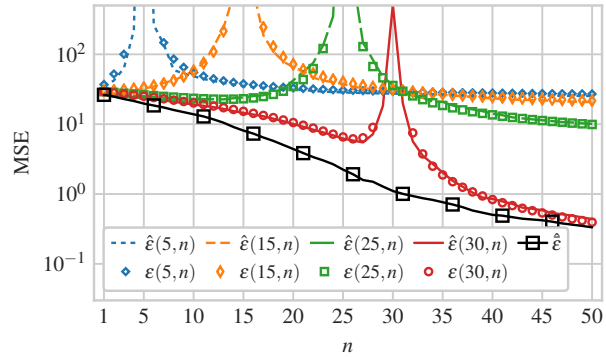

Fig. 3: The MSE in setting *(S3)*. The SNR is high, but the covariance matrix for $\boldsymbol{x}$ is randomized, rather than being identity.

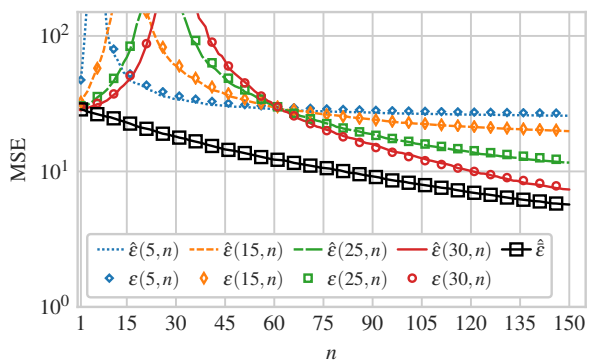

Fig. 2: The MSE in the low SNR case *(S2)*.

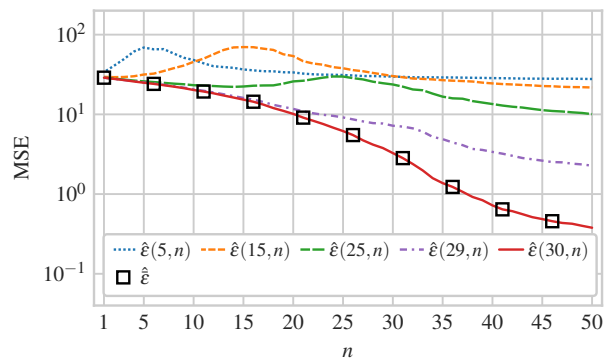

Fig. 4: The empirical MSE in setting *(S4)*. Here the partial LMMSE estimator knows the noise level, i.e., $\hat{\sigma}_z = \sigma_v$.

Corollary 2 shows that n needs to be sufficiently large to guarantee a performance gain with an increase in the assumed model's complexity p_S. It also shows that for $p+1 < n < p + \sigma_v^2/\sigma_x^2 + 1$, the MSE increases with p_S. We note that the bound is larger for worse signal-to-noise (SNR) ratios σ_x^2/σ_v^2, increasing as the SNR decreases.

IV. Numerical Results

A. Experimental Setup

The numerical results are obtained by averaging over $M = 100$ simulations. In each simulation (j), $j = 1, \ldots, M$, we draw one random vector $\boldsymbol{x}^{(j)}$ from a Gaussian distribution $\mathcal{N}(0, \boldsymbol{K}_x)$, one random vector $\boldsymbol{v}^{(j)}$ from $\mathcal{N}(0, \boldsymbol{K}_v)$ and one matrix $\boldsymbol{A}^{(j)}$ where each row is drawn from $\mathcal{N}(0, \boldsymbol{I}_p)$. The matrix $\boldsymbol{A}_S^{(j)}$ is extracted as the first p_S columns of $\boldsymbol{A}^{(j)}$. The partial estimate $\hat{\boldsymbol{x}}_S^{(j)}$ is then created using $\boldsymbol{W}_S^{(j)}$ from (11). The MSE $J^{(j)}(\boldsymbol{W}_S^{(j)})$ is then computed as

$$J^{(j)}(\boldsymbol{W}_S^{(j)}) = \|\boldsymbol{x}_S^{(j)} - \hat{\boldsymbol{x}}_S^{(j)}\|^2 + \mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C}), \tag{25}$$

and averaged over the M simulations to create the empirical average MSE, as an estimate of ε:

$$\hat{\varepsilon}(p_S, n) \triangleq \frac{1}{M} \sum_{i=1}^{M} J^{(j)}(\boldsymbol{W}_S^{(j)}). \tag{26}$$

We set $p = 30$ and vary p_S and n to illustrate how $\hat{\varepsilon}$ changes. We have $\boldsymbol{K}_{\boldsymbol{v}} = \sigma_v^2 \boldsymbol{I}_n$ and $\hat{\boldsymbol{K}}_{\boldsymbol{z}} = \hat{\sigma}_z^2 \boldsymbol{I}_n$. We consider the following experiment scenarios *(S1)* – *(S4)* in terms of assumptions on $\boldsymbol{K}_{\boldsymbol{x}}$, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S}$, σ_v and $\hat{\sigma}_z$: *(S1)* $\boldsymbol{K}_{\boldsymbol{x}} = \boldsymbol{I}_p$, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S} = \boldsymbol{I}_{p_S}$, $\sigma_v = 0.5$, $\hat{\sigma}_z = 0$; *(S2)* $\boldsymbol{K}_{\boldsymbol{x}} = \boldsymbol{I}_p$, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S} = \boldsymbol{I}_{p_S}$, $\sigma_v = \sqrt{p}$, $\hat{\sigma}_z = 0$; *(S3)* $\boldsymbol{K}_{\boldsymbol{x}} = \frac{p}{\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{C}_{\boldsymbol{x}})} \boldsymbol{C}_{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{C}_{\boldsymbol{x}}$, with the entries of $\boldsymbol{C}_{\boldsymbol{x}} \in \mathbb{R}^{p \times p}$ sampled once from $\mathcal{N}(0,1)$ and fixed throughout the M simulations, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S} = \boldsymbol{I}_{p_S}$, $\sigma_v = 0.5$, $\hat{\sigma}_z = 0$; *(S4)* $\boldsymbol{K}_{\boldsymbol{x}} = \boldsymbol{I}_p$, $\hat{\boldsymbol{K}}_{\boldsymbol{x}_S} = \boldsymbol{I}_{p_S}$, $\sigma_v = \hat{\sigma}_z = 0.5$. Note that in the settings *(S1)*, *(S3)* and *(S4)*, the SNR $10 \log_{10}(\mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}})/\sigma_v^2)$ is around 21 dB, and in setting *(S2)*, the SNR is at 0 dB.

B. The MSE, Model Order and the Number of Samples

In Figures 1 – 3 we plot the empirical average MSE $\hat{\varepsilon}$ from (26) together with the analytical expected MSE $\varepsilon = \varepsilon_S + \mathrm{tr}(\boldsymbol{K}_{\boldsymbol{x}_C})$ with ε_S from (20) versus the number of samples n, for the experiments *(S1)* – *(S3)*. The empirical values are marked with lines, and the analytical with markers. We also plot the empirically averaged MSE $\hat{\hat{\varepsilon}}$ of the true LMMSE estimator in (6) as a line with sparsely placed markers.

Overview: In the plots, we observe a perfect match between the empirical and the analytical curves, confirming our analytical results. There are clear peaks in MSE when the number of samples n is close to the number of parameters p_S in the assumed model, as expected from the behaviour of γ in (20). It is clear that in this estimation setting, one should avoid having n close to p_S, and that changing the number of samples or the assumed model order can significantly improve the MSE.

Effect of n and p_s: Consistently over all figures and all p_S, the MSE decreases monotonically as n grows, for $n > p_S$. However, to have performance close to that of the true LMMSE estimator, p_S must be equal to $p = 30$. For other p_S, there is a gap in the MSE between the partial and the true estimator which does not vanish as n grows. In *(S1)*, where the model on $\boldsymbol{x}_S$ is correct and the SNR is high, the MSE with $p_S = p$ is close to that of the true estimator's for all n (except for when n is close to p_S), despite the fact that the assumed model ignores the noise. Furthermore, even in *(S2)* and *(S3)*, where the SNR is either low (and noise is still ignored in the assumed estimator), or the assumed model on $\boldsymbol{x}_S$ is incorrect, it is still possible to get error values comparable to that of the true estimator with $p_S = p$ for large n.

Effect of SNR: Figure 2 illustrates the result of Corollary 2, i.e., if n is large enough $n > p + \sigma_v^2/\sigma_x^2 + 1$, then the MSE decreases monotonically with p_S. This bound on n provides a clear marking of an operating range on which performance gain is guaranteed when increasing p_S. In this low SNR setting, there is a large range of n on which a smaller p_S gives lower MSE. In other words, although we have the correct model on x_S, choosing a smaller model size p_S can improve the performance even after interpolation threshold (i.e., $n = p$), if the estimator ignores the noise ($\hat{\sigma}_z = 0$) and there is not enough data.

Matched noise level: In Figure 4, we plot the results of experiment *(S4)*, where the assumed noise level is the same as the level of the true noise: $\hat{\sigma}_z = \sigma_v$. There are still peaks in the MSE but they are significantly dampened compared to the earlier experiments. For high values of p_S, e.g., $p_S = 29$, 30, the MSE decreases monotonically with n. For lower values of p_S, we still have the same effects on the MSE as we saw in the settings *(S1)–(S3)*. More specifically, as in previous experiments, the performance is not guaranteed to improve with increasing n or p_S.

V. Conclusions

Under an LMMSE estimation framework, we investigated the average degradation of the estimation performance due to model mismatch. Our analytical results, verified with simulations, illustrate the interplay between the SNR, the number of samples and the model orders of the underlying and the assumed models. In general, neither increasing the number of samples nor the assumed model complexity can guarantee performance improvement. Extending these results to different covariance structures on the regressors as well as to complex regressors, including non-circular scenarios, are important directions for future work.

VI. Appendix

A. Proof of Theorem 1

In the setting of Theorem 1, the partial LMMSE estimator in (11) is $\boldsymbol{W}_S = \boldsymbol{A}_S^{\mathrm{T}}(\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+ = \boldsymbol{A}_S^+$. Plugging this $\boldsymbol{W}_S$ into (17), and applying the trace operator, we have

$$\begin{aligned} J_S(\boldsymbol{A}_S^+) = &\operatorname{tr}\big((\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)^{\mathrm{T}}(\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)\boldsymbol{K}_{\boldsymbol{x}_S} \\ &+ \boldsymbol{A}_C^{\mathrm{T}}(\boldsymbol{A}_S^+)^{\mathrm{T}}\boldsymbol{A}_S^+\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C} + (\boldsymbol{A}_S^+)^{\mathrm{T}}\boldsymbol{A}_S^+\boldsymbol{K}_{\boldsymbol{v}} \\ &- 2(\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)^{\mathrm{T}}\boldsymbol{A}_S^+\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C\boldsymbol{x}_S}\big), \end{aligned} \tag{27}$$

By the definition of the pseudoinverse, the matrix $\boldsymbol{A}_S^+\boldsymbol{A}_S$ is symmetric, $\boldsymbol{A}_S^+\boldsymbol{A}_S\boldsymbol{A}_S^+ = \boldsymbol{A}_S^+$. Hence, $(\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)^{\mathrm{T}}(\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S) = (\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)$, and $(\boldsymbol{I}_{p_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S)^{\mathrm{T}}\boldsymbol{A}_S^+ = \boldsymbol{A}_S^+ - \boldsymbol{A}_S^+ = 0$. Furthermore, the pseudoinverse has the property $(\boldsymbol{A}_S^+)^{\mathrm{T}}\boldsymbol{A}_S^+ = (\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+$. We can now write

$$\begin{aligned} J_S(\boldsymbol{A}_S^+) = &\operatorname{tr}\big(\boldsymbol{K}_{\boldsymbol{x}_S} - \boldsymbol{A}_S^+\boldsymbol{A}_S\boldsymbol{K}_{\boldsymbol{x}_S} \\ &+ \boldsymbol{A}_C^{\mathrm{T}}(\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C} + (\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+\boldsymbol{K}_{\boldsymbol{v}}\big). \end{aligned} \tag{28}$$

We now take the expectation over the distribution of $\boldsymbol{A}$, noting that $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$ are uncorrelated, and use the linearity and cyclic property of the trace operator to write:

$$\begin{aligned} \varepsilon(p_S, n) = &\operatorname{tr}(\boldsymbol{K}_{\boldsymbol{x}_S}) - \operatorname{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{A}_S}\left[\boldsymbol{A}_S^+\boldsymbol{A}_S\right]\boldsymbol{K}_{\boldsymbol{x}_S}\right) \\ &+ \operatorname{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{A}_S}\left[(\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+\right]\mathop{\mathbb{E}}_{\boldsymbol{A}_C}\left[\boldsymbol{A}_C\boldsymbol{K}_{\boldsymbol{x}_C}\boldsymbol{A}_C^{\mathrm{T}}\right]\right) \\ &+ \operatorname{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{A}_S}\left[(\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+\right]\boldsymbol{K}_{\boldsymbol{v}}\right). \end{aligned} \tag{29}$$

We continue the proof by noting that

$$\operatorname{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{A}_S}[\boldsymbol{A}_S^+\boldsymbol{A}_S]\boldsymbol{K}_{\boldsymbol{x}_S}\right) = \mathop{\mathbb{E}}_{\boldsymbol{x}_S}\left[\boldsymbol{x}_S^{\mathrm{T}}\mathop{\mathbb{E}}_{\boldsymbol{A}_S}[\boldsymbol{A}_S^+\boldsymbol{A}_S]\boldsymbol{x}_S\right] \tag{30}$$

$$= \tfrac{\min\{p_S,n\}}{p_S}\operatorname{tr}(\boldsymbol{K}_{\boldsymbol{x}_S}), \tag{31}$$

where in the last step we used Lemma 3 of [10], as well as $E_{\boldsymbol{x}_S}[\|\boldsymbol{x}_S\|^2] = \operatorname{tr}(\boldsymbol{K}_{\boldsymbol{x}_S})$. By [11], we have that $\mathbb{E}_{\boldsymbol{A}_S}[(\boldsymbol{A}_S\boldsymbol{A}_S^{\mathrm{T}})^+] = \frac{1}{n}\gamma\boldsymbol{I}_n$, with γ as in (21). We can now write

$$\begin{aligned} \varepsilon_S(p_S, n) = &\operatorname{tr}(\boldsymbol{K}_{\boldsymbol{x}_S}) - \tfrac{\min\{p_S,n\}}{p_S}\operatorname{tr}(\boldsymbol{K}_{\boldsymbol{x}_S}) \\ &+ \frac{1}{n}\gamma\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{x}_C}\mathop{\mathbb{E}}_{\boldsymbol{A}_C}\left[\boldsymbol{A}_C^{\mathrm{T}}\boldsymbol{A}_C\right]\right) + \frac{1}{n}\gamma\operatorname{tr}(\boldsymbol{K}_{\boldsymbol{v}}), \end{aligned} \tag{32}$$

into which we plug in that $\mathbb{E}_{\boldsymbol{A}_C}[\boldsymbol{A}_C^{\mathrm{T}}\boldsymbol{A}_C] = n\boldsymbol{I}_{p_C}$, to yield the final expression for $\varepsilon_S(p_S, n)$.

References

[1] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[2] S. M. Kay, *Fundamentals of Stat. Signal Process.* Prentice Hall, 1993.

[3] D. Lederman and J. Tabrikian, "Constrained MMSE estimator for distribution mismatch compensation," in *Fourth IEEE Workshop on Sensor Array and Multichannel Processing*, 2006, pp. 439–443.

[4] R. Mittelman and E. L. Miller, "Robust estimation of a random parameter in a Gaussian linear model with joint eigenvalue and elementwise covariance uncertainties," *IEEE Trans. on Signal Process.*, vol. 58, no. 3, pp. 1001–1011, 2010.

[5] D. Zachariah, N. Shariati, M. Bengtsson, M. Jansson *et al.*, "Estimation for the linear model with uncertain covariance matrices," *IEEE Trans. on Signal Process.*, vol. 62, no. 6, pp. 1525–1535, 2014.

[6] X. Liu, D. Zachariah, and P. Stoica, "Robust prediction when features are missing," *IEEE Signal Process. Letters*, vol. 27, p. 720–724, 2020.

[7] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Trans. on Signal Process.*, vol. 53, no. 1, pp. 168–181, 2005.

[8] L. Breiman and D. Freedman, "How many variables should be entered in a regression equation?" *J. Amer. Stat. Assoc.*, vol. 78, no. 381, pp. 131–136, 1983.

[9] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020.

[10] M. Hellkvist, A. Özçelikkale, and A. Ahlén, "Generalization error for linear regression under distributed learning," *IEEE Int. Workshop on Signal Process. Advances in Wireless Commun.*, May 2020.

[11] R. D. Cook and L. Forzani, "On the mean and variance of the generalized inverse of a singular Wishart matrix," *Electron. J. Statist.*, vol. 5, pp. 146–158, 2011.