CCIT Report #615 March 2007

Linear Regression with Gaussian Model Uncertainty: Algorithms and Bounds

Ami Wiesel, Yonina C. Eldar and Arie Yeredor

***Abstract*—We consider the problem of estimating an unknown deterministic parameter vector in a linear regression model with random Gaussian uncertainty in the mixing matrix. We prove that the maximum likelihood (ML) estimator is a regularized least squares estimator and develop three alternative approaches for finding the regularization parameter which maximizes the likelihood. We analyze the performance using the Cramér Rao bound (CRB) on the mean squared error, and show that the degradation in performance due the uncertainty is not as severe as may be expected. Next, we address the problem again assuming that the variances of the noise and the elements in the model matrix are unknown and derive the associated CRB and ML estimator.**

We compare our methods to known results on linear regression in the error in variables (EIV) model. We discuss the similarity between these two competing approaches, and provide a thorough comparison which sheds light on their theoretical and practical differences.

***Index Terms*—Maximum likelihood estimation, Total least squares, Errors in Variables, Linear models, Random model matrix.**

I. Introduction

One of the most classical problems in statistical signal processing is that of estimating an unknown, deterministic vector parameter $\mathbf{x}$ in the linear regression model $\mathbf{y} = \mathbf{G}\mathbf{x}+\mathbf{w}$ where $\mathbf{G}$ is a linear transformation and $\mathbf{w}$ is a Gaussian noise vector. The importance of this problem stems from the fact that a wide range of problems in communications, array processing, and many other areas can be cast in this form.

Most of the literature concentrates on the simplest case, in which it is assumed that the model matrix $\mathbf{G}$ is completely specified. In this setting, the celebrated least squares (LS) estimator coincides with the maximum likelihood (ML) solution and is known to minimize the mean-squared-error (MSE) among all unbiased estimators of $\mathbf{x}$ [1], [2]. Nonetheless, it may be outperformed in terms of MSE by biased methods such as the regularized LS estimator due to Tikhonov [3], the James-Stein method [4], and the minimax MSE approach [5].

The linear regression problem for cases where $\mathbf{G}$ is not completely specified received much less attention. In this case, there are many mathematical models for describing the uncertainty in $\mathbf{G}$. Each of these models leads to different optimization criteria and accordingly to different estimation algorithms. Most of the literature can be divided into two main categories, in which the uncertainty is expressed using either deterministic or random models. A standard deterministic approach is the "robust LS" which is designed to cope with the worst-case $\mathbf{G}$ within a known deterministic set [6], [7]. Recently, the minimax MSE criterion was also considered in this problem formulation [5]. In the stochastic uncertainty models, $\mathbf{G}$ is usually known up to some Gaussian distortion. Typically, there are two approaches in this setting. First, one can use a random variables (RV) model and assume that $\mathbf{G}$ is a random Gaussian matrix with known statistics. Based on this model, different estimation methods have been considered. The ML estimator was derived in our recent letter [8]. An alternative strategy is to minimize the expected LS criterion with respect to $\mathbf{G}$ [9], [10]. The minimax MSE estimator was also generalized to this setting in [10]. The second approach is the standard Errors-in-Variables (EIV) model, where $\mathbf{G}$ is considered a deterministic unknown matrix, and an additional noisy observation on this matrix is available [11]. The ML solution for $\mathbf{x}$ in this case was addressed in [11], and coincides with the well known total LS (TLS) estimator [12] (when the additive Gaussian noise $\mathbf{w}$ is independent and identically distributed).

Evidently, there are different models and optimization criteria for estimating $\mathbf{x}$ in a linear model with uncertainty in the model matrix. The main objective of this paper is to compare the Gaussian uncertainty approaches and to shed light on their advantages and disadvantages. In particular, we consider the two classical Gaussian uncertainty formulations: the RV and EIV models. We explain the practical and theoretical differences between them, and discuss the scenarios in which each is appropriate.

The main part of this paper considers ML estimation of $\mathbf{x}$ in the RV linear regression model. We prove that the ML estimate (MLE) is a regularized (or deregularized) LS estimator, and that its regularization parameter and squared norm can be characterized as a saddle point of a concave-quasiconvex objective function. Thus, we can efficiently find the optimal parameters numerically. In fact, our previous solution in [8] can be interpreted as a minimax search for this saddle point. Using this new characterization, we present a more efficient maximin search. Furthermore, an appealing approach for finding the ML estimate in this setting is to resort to the classical expectation maximization (EM) algorithm which is known to converge to a stationary point of the ML objective (see [13], [14], [15] and references within). Due to the non-convexity of the log-likelihood function, there is no guarantee that this point will indeed be the global maximum. Fortunately, our saddle point interpretation provides a simple method for checking the global optimality of the convergence

A. Wiesel and Y. C. Eldar were supported by the EU 6th framework programme, via the NEWCOM network of excellence. Some of the results in this paper were presented in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2006). A. Wiesel and Y. C. Eldar are with the Technion, Israel Institute of Technology and A. Yeredor is with Tel Aviv University, Israel. Emails: amiw@tx.technion.ac.il, yonina@ee.technion.ac.il, and arie@eng.tau.ac.il.

point. We conclude this part of the paper with a comparison to the ML in the EIV model and show that our ML estimator is a regularized version of the latter.

In the second part of the paper, we analyze the performance in the RV model using the Cramér Rao bound (CRB) on the MSE of unbiased estimators [1], [2]. We derive the CRB associated with our model and quantify the degradation in performance due to the randomness of $\mathbf{G}$. Interestingly, the degradation in performance is not as severe as one may suspect. Actually, as we will show and quantify, randomness in $\mathbf{G}$ may even improve the performance in terms of MSE.

However, the potential improvement is contingent upon the assumption that the variances of the random variables are all known. In practice, this knowledge is not always available, and therefore we also consider the case in which these variances are unknown deterministic nuisance parameters. As before, we begin with the ML estimator which reduces to the standard LS. Then, we derive the associated CRB and analyze the degradation in performance inflicted by the lack of knowledge regarding the variances. We conclude this section with a comparison to the EIV model. Interestingly, under these assumptions the ML estimate does not exist in the EIV model ([16] and references within), and the CRB has a similar structure to our random model bound.

The paper is organized as follows. In Section II we introduce the problem formulation. ML estimation of $\mathbf{x}$ when the variances are known is discussed in Section III. CRB analysis under this setting is analyzed in Section IV. Next, we dedicate Section V to the estimation of $\mathbf{x}$ when the variances are unknown nuisance parameters. A few numerical examples are demonstrated in Section VI. Finally, in Section VII we provide concluding remarks.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscripts $(\cdot)^T$, $(\cdot)^{-1}$, $(\cdot)^\dagger$, $(\cdot)'$ and $(\cdot)''$ denote the transpose, matrix inverse, Moore-Penrose pseudoinverse and first and second derivatives, respectively. The operators $\otimes$, $\mathrm{vec}\left(\cdot\right)$, $\mathrm{Tr}\left\{\cdot\right\}$, $\|\cdot\|$ and $\|\cdot\|_F$ denote the Kronecker matrix multiplication, the vector obtained by stacking the columns of a matrix one over the other, the trace operator, the standard Euclidean norm and the Frobenius matrix norm, respectively. The matrix $\mathbf{I}$ denotes the identity, $\mathcal{R}\left(\mathbf{A}\right)$ is the range space of the columns of $\mathbf{A}$, $\lambda_{\min}\left(\mathbf{A}\right)$ is the minimum eigenvalue of $\mathbf{A}$, and $\mathbf{A}\succeq\mathbf{0}$ means that $\mathbf{A}$ is positive semidefinite.

II. Problem formulation

We consider the classical linear regression model in which

$$\mathbf{y}=\mathbf{G}\mathbf{x}+\mathbf{w}, \tag{1}$$

where $\mathbf{y}$ is a length N observed vector, $\mathbf{G}$ is an $N\times K$ linear model matrix, $\mathbf{x}$ is a length K deterministic unknown vector, and $\mathbf{w}$ is a zero-mean Gaussian random vector with mutually independent elements, each with variance $\sigma^2>0$.

Methods for estimating $\mathbf{x}$ in (1) when $\mathbf{G}$ is completely specified have been intensively studied. The problem becomes more interesting and challenging when $\mathbf{G}$ is not exactly known. In this case, there are different mathematical models for describing the uncertainty in $\mathbf{G}$. In this paper, we model $\mathbf{G}$ as a random matrix with known mean given by $\mathbf{H}$:

$$\mathbf{G} \;=\; \mathbf{H}+\mathbf{W}, \tag{2}$$

where $\mathbf{W}$ is a random matrix of mutually independent, zero-mean Gaussian elements with variance $\sigma_h^2>0$, independent of $\mathbf{w}$. For simplicity, we assume that $\mathbf{H}$ is full column rank.

Our problem is to estimate $\mathbf{x}$ from the observations $\mathbf{y}$ in the model (1)-(2). We consider two problem formulations. First, we assume that the variances σ^2 and σ_h^2 are known. Under this assumption, we focus on ML estimation and CRB analysis. Second, we discuss the problem when the variances are unknown nuisance parameters which must be estimated as well.

A. Comparison to EIV

Throughout the paper, we will compare our results with previous works on a very similar uncertainty model, namely, the EIV model. We conclude each section with a comparison to this standard model. For simplicity, we slightly abuse the notations so as to use the same notations for both formulations.

In the EIV formulation, we model $\mathbf{G}$ as a deterministic unknown matrix, and assume that we are given a noisy observation on $\mathbf{G}$ in addition to the vector $\mathbf{y}$:

$$\mathbf{H} \;=\; \mathbf{G}+\mathbf{W}, \tag{3}$$

where $\mathbf{W}$ is a random matrix of mutually independent, zero-mean Gaussian elements with variance $\sigma_h^2>0$, independent of $\mathbf{w}$. For simplicity, we assume that $\mathbf{G}$ is full column rank.

Models (2) and (3) are closely related. In fact, one can easily pass $\mathbf{W}$ from the right hand side of (3) to its left hand side and since the distribution of $\mathbf{W}$ is invariant to a sign change, it may initially seem that the two models are identical. One of the main contributions of our work is to elucidate the mathematical-statistical differences between these two models. First, let us propose a practical example for the use of each. Consider a communication system over a multiple-input-multiple-output (MIMO) channel denoted by $\mathbf{G}$. The RV model is appropriate when the channel itself is random and time varying, with a known distribution around some “nominal” $\mathbf{H}$. The cause of the uncertainty is the randomness of the channel. On the other hand, the EIV model is appropriate when the channel is unknown but fixed (with no prior distribution) but some noisy estimate $\mathbf{H}$ of the channel is available (for example, in communication systems using a training phase). The uncertainty is the result of the imperfect noisy estimation. Despite these differences, the models are very similar. In most applications it is not obvious which reason is the cause of the uncertainty and it can often be a combination of both. Moreover, the models basically provide the same information: we have access to $\mathbf{y}$ and $\mathbf{H}$, and the true channel $\mathbf{G}$ is equal to $\mathbf{H}$ plus some Gaussian noise. Yet when we consider statistical properties of the estimate, a key question is whether in each realization of the data $\mathbf{y}$, $\mathbf{G}$ remains constant and $\mathbf{H}$ varies (EIV) or $\mathbf{H}$ remains constant and $\mathbf{G}$ varies (RV).

III. Maximum Likelihood estimation

In this section we discuss the ML estimation of $\mathbf{x}$ in (1)-(2) from $\mathbf{y}$ when $\mathbf{H}$, σ^2 and σ_h^2 are known parameters. We characterize the structure of the MLE and suggest efficient numerical methods for the associated optimization problem.

The ML method is one of the most common approaches in estimation theory whereby the estimates are chosen as the parameters that maximize the likelihood of the observations:

$$\widehat{\mathbf{x}}_{\mathrm{ML}} = \arg\max_{\mathbf{x}} \log f(\mathbf{y};\mathbf{x}), \tag{4}$$

where $f(\mathbf{y};\mathbf{x})$ is the probability density function of $\mathbf{y}$ parameterized by $\mathbf{x}$. In our model the vector $\mathbf{y}$ is a Gaussian vector with mean $\mathbf{Hx}$ and covariance $(\sigma_h^2\|\mathbf{x}\|^2+\sigma^2)\mathbf{I}$. Therefore, the ML estimator can be found by solving

$$\widehat{\mathbf{x}}_{\mathrm{ML}} = \arg\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{y}-\mathbf{Hx}\|^2}{\sigma_h^2\|\mathbf{x}\|^2+\sigma^2} + N\log(\sigma_h^2\|\mathbf{x}\|^2+\sigma^2) \right\}. \tag{5}$$

Problem (5) is a K dimensional, nonlinear, and nonconvex optimization problem and is therefore considered difficult. In the following theorem, we characterize its solution using a simple change of variables.

Theorem 1: Consider the estimation of $\mathbf{x}$ in (5) under the assumption that the norm of $\mathbf{x}$ is upper bounded, say $\|\mathbf{x}\|^2 \leq U$ for some sufficiently large U. Then, the ML estimator is a (de)regularized LS, i.e.,

$$\widehat{\mathbf{x}}_{\mathrm{ML}} = \left(\mathbf{H}^T\mathbf{H}+\alpha_{\mathrm{ML}}\mathbf{I}\right)^{\dagger}\mathbf{H}^T\mathbf{y}, \tag{6}$$

with squared norm $t_{\mathrm{ML}} = \|\widehat{\mathbf{x}}_{\mathrm{ML}}\|^2$. The parameters $\alpha_{\mathrm{ML}} \in \mathcal{A}$ and $t_{\mathrm{ML}} \in \mathcal{T}$ are a saddle point of the following concave-quasiconvex optimization problem

$$\min_{t\in\mathcal{T}}\max_{\alpha\in\mathcal{A}} f(\alpha,t) = \max_{\alpha\in\mathcal{A}}\min_{t\in\mathcal{T}} f(\alpha,t), \tag{7}$$

where

$$\begin{aligned} f(\alpha,t) &= \frac{c(\alpha)-\alpha t}{\sigma_h^2 t+\sigma^2} + N\log(\sigma_h^2 t+\sigma^2) \\ c(\alpha) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}\left[\mathbf{H}^T\mathbf{H}+\alpha\mathbf{I}\right]^{\dagger}\mathbf{H}^T\mathbf{y} \\ \mathcal{A} &= \left\{\alpha : \begin{array}{c} \alpha \geq -\lambda_{\min}\left(\mathbf{H}^T\mathbf{H}\right) \\ \mathbf{H}^T\mathbf{y} \in \mathcal{R}\left(\mathbf{H}^T\mathbf{H}+\alpha\mathbf{I}\right) \end{array}\right\} \\ \mathcal{T} &= \{t : 0 \leq t \leq U\}. \end{aligned} \tag{8}$$

Before proving the theorem, we note that the bounded norm assumption is a technical mathematical condition needed for the proof, and in practice the saddle point does not depend on the specific choice of U. Thus, the estimator is not assumed to have knowledge of U. Second, the set $\mathcal{A}$ is convex and for all practical purposes is equivalent to $\alpha \geq -\lambda_{\min}\left(\mathbf{H}^T\mathbf{H}\right)$. The range constraint is again a technical condition which is almost always satisfied.

Proof: Introducing an auxiliary variable $t = \|\mathbf{x}\|^2$ into (5) yields

$$\min_{t\in\mathcal{T}} \frac{g(t)}{\sigma_h^2 t+\sigma^2} + N\log\left(\sigma_h^2 t+\sigma^2\right), \tag{9}$$

where

$$g(t) = \left\{ \begin{array}{ll} \min_{\mathbf{x}} & \|\mathbf{y}-\mathbf{Hx}\|^2 \\ \text{s.t.} & \|\mathbf{x}\|^2 = t \end{array} \right. . \tag{10}$$

In [17], [18] it was shown that due to hidden convexity, (10) can be solved via its Lagrange dual program

$$g(t) = \max_{\alpha\in\mathcal{A}}\{c(\alpha)-\alpha t\}, \tag{11}$$

where $c(\alpha)$ and $\mathcal{A}$ are defined in (8). Moreover, the optimal $\mathbf{x}$ and α satisfy

$$\mathbf{x} = \left(\mathbf{H}^T\mathbf{H}+\alpha\mathbf{I}\right)^{\dagger}\mathbf{H}^T\mathbf{y}. \tag{12}$$

Thus, we can rewrite the problem as

$$\min_{t\in\mathcal{T}}\max_{\alpha\in\mathcal{A}} f(\alpha,t), \tag{13}$$

where $f(\alpha,t)$ is defined in (8). The objective $f(\alpha,t)$ is concave in any $\alpha\in\mathcal{A}$ for fixed $t\in\mathcal{T}$ since it originates in a Lagrange dual program. In Appendix A, we prove that it is also quasi-convex in $t\in\mathcal{T}$ for fixed $\alpha\in\mathcal{A}$. The feasible sets are both convex and due to our assumption on the norm of $\mathbf{x}$, $\mathcal{T}$ is compact. Therefore, according to Sion's quasi-concave-convex minimax theorem there exists a saddle point as expressed in (7) [19]. ■

The theorem characterizes the structure of the MLE and relates it to the class of (de)regularized LS solutions. Moreover, as we will show, the concave-quasiconvex property allows for efficient numerical methods for finding the optimal regularization parameter. It is important to emphasize that Theorem 1 does not claim that the original ML estimation problem in (5) is convex in $\mathbf{x}$. The convexity is a result of our change in variables and refers to α and t alone. In the following subsections, we explain how this property may be used to find the MLE numerically.

A. Minimax - numerical method

The first approach is to solve the minimax problem, i.e., to minimize

$$\widetilde{f}(t) = \max_{\alpha\in\mathcal{A}} f(\alpha,t) \tag{14}$$

with respect to $t\in\mathcal{T}$. This requires two nested line searches. We have an outer minimization with respect to t and for each fixed t we need to solve (14) via (10). More details on this approach can be found in our earlier letter [8].

In practice, the main computational complexity in this approach is evaluating $c(\alpha)$ for different values of α. Fortunately, these could be easily implemented by utilizing the eigenvalue decomposition of $\mathbf{H}^T\mathbf{H}$.

B. Maximin - numerical method

The second approach is to solve the maximin problem, i.e., to maximize

$$\overline{f}(\alpha) = \min_{t\in\mathcal{T}} f(\alpha,t) \tag{15}$$

with respect to $\alpha\in\mathcal{A}$. This approach leads to a single line search as the inner minimization can be solved in closed form. It is a convex minimization over a closed interval, and its solution is either a stationary point or one of the extreme points. Setting the derivative to zero results in

$$\frac{\partial f(\alpha,t)}{\partial t} = \frac{N\sigma_h^2-\alpha}{\sigma_h^2 t+\sigma^2} - \sigma_h^2\frac{c(\alpha)-\alpha t}{\left(\sigma_h^2 t+\sigma^2\right)^2} = 0, \tag{16}$$

and solving for t yields

$$\bar{t}(\alpha) = \frac{\alpha\sigma^2}{N\sigma_h^4} + \frac{c(\alpha)}{N\sigma_h^2} - \frac{\sigma^2}{\sigma_h^2}. \tag{17}$$

Therefore,

$$\overline{f}(\alpha) = \begin{cases} f(\alpha, \bar{t}(\alpha)) & 0 \le \bar{t}(\alpha) \le U \\ \min\{f(\alpha, 0), f(\alpha, U)\} & \text{else.} \end{cases} \tag{18}$$

The function $\overline{f}(\alpha)$ is concave in α since $f(\alpha, t)$ is concave in α. Therefore, any standard line search can easily find its maximum which corresponds to a saddle point of $f(\alpha, t)$.

Here too, the main computational complexity is evaluating $c(\alpha)$ for different values of α which may be implemented by utilizing the eigenvalue decomposition of $\mathbf{H}^T\mathbf{H}$. Nevertheless, the maximin approach is more appealing than the minimax version since it involves only one line search instead of two nested searches.

C. EM - numerical method

We now provide an alternative numerical solution for the ML problem in (4) based on the classical EM algorithm [13]. The EM method is an iterative approach for solving ML problems with missing data. It is known to converge to a stationary point of the likelihood. We will apply the EM technique to our problem and will show how Theorem 1 can be used to verify whether a stationary point obtained by EM is indeed the correct ML estimate.

At each iteration, the EM algorithm maximizes the expected log likelihood of $\{\mathbf{y}, \mathbf{G}\}$ with respect to the missing $\mathbf{G}$ (instead of the log likelihood of $\mathbf{y}$ itself). The result is the following updating formula (See Appendix B):

$$\mathbf{x}_{n+1} \quad = \quad \left[\overline{\mathbf{G}_n^T\mathbf{G}_n}\right]^{-1} \overline{\mathbf{G}_n}^T \mathbf{y}, \tag{19}$$

where

$$\begin{aligned} \overline{\mathbf{G}_n} \quad &= \quad \mathbf{H} + \frac{\sigma_h^2}{\sigma_h^2\|\mathbf{x}_n\|^2 + \sigma^2}\left[\mathbf{y} - \mathbf{H}\mathbf{x}_n\right]\mathbf{x}_n^T \\ \overline{\mathbf{G}_n^T\mathbf{G}_n} \quad &= \quad \overline{\mathbf{G}_n}^T\overline{\mathbf{G}_n} + N\left[\sigma_h^2\mathbf{I} - \frac{\sigma_h^4\mathbf{x}_n\mathbf{x}_n^T}{\sigma_h^2\|\mathbf{x}_n\|^2 + \sigma^2}\right]. \end{aligned} \tag{20}$$

The iterations are very simple to implement in practical fixed point signal processors. In practice, there is no need to invert the matrix $\overline{\mathbf{G}_n^T\mathbf{G}_n}$, as $\mathbf{x}_{n+1}$ can be found as a solution of a set of linear equations (the matrix $\overline{\mathbf{G}_n^T\mathbf{G}_n}$ is always invertible).

The EM algorithm will converge to a stationary point of the likelihood function [13]. However, since our problem is non-convex, this may be a local maximum depending on the initial conditions $\mathbf{x}_0$. Our experience with arbitrary parameters and

$$\mathbf{x}_0 \quad = \quad \left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{y}, \tag{21}$$

shows that the above iterations usually converge to the correct ML estimate, but it is easy to find initial conditions which converge to spurious local maximum. Nonetheless, using Theorem 1 we can determine whether or not a given stationary point of the EM algorithm is indeed the global maximum:

Theorem 2: Let $\mathbf{x}_{\text{EM}}$ be a stationary point of the EM algorithm. Then, $\mathbf{x}_{\text{EM}}$ is a (de)regularized LS estimate:

$$\mathbf{x}_{\text{EM}} = \left[\mathbf{H}^T\mathbf{H} + \alpha_{\text{EM}}\mathbf{I}\right]^{-1}\mathbf{H}^T\mathbf{y}, \tag{22}$$

where

$$\alpha_{\text{EM}} = \sigma_h^2\left(N - \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}_{\text{EM}}\|^2}{\sigma_h^2\|\mathbf{x}_{\text{EM}}\|^2 + \sigma^2}\right). \tag{23}$$

Moreover, let $t_{\text{EM}} = \|\mathbf{x}_{\text{EM}}\|^2$. Then a necessary and sufficient condition for $\mathbf{x}_{\text{EM}} = \mathbf{x}_{\text{ML}}$ is $\alpha_{\text{EM}} \in \mathcal{A}$, $t_{\text{EM}} \in \mathcal{T}$ where $\mathcal{A}$ and $\mathcal{T}$ are defined in (8).

Proof: Any stationary point will satisfy the EM iteration with $\mathbf{x}_{\text{EM}} = \mathbf{x}_n = \mathbf{x}_{n+1}$. Rearranging the terms in the iterations using simple algebraic manipulations yields

$$c\left[\mathbf{H}^T\mathbf{H} + \sigma_h^2\left(N - \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma_h^2\|\mathbf{x}_{\text{EM}}\|^2 + \sigma^2}\right)\mathbf{I}\right]\mathbf{x} = c\mathbf{H}^T\mathbf{y}, \tag{24}$$

where

$$c = \frac{\sigma^2}{\sigma_h^2\|\mathbf{x}_{\text{EM}}\|^2 + \sigma^2}. \tag{25}$$

Due to $\sigma^2 > 0$, we can divide both sides by $c > 0$ and obtain the required result. Therefore, $\mathbf{x}_{\text{EM}}$ is a (de)regularized LS with parameters α_{EM} and t_{EM}.

It will be equal to $\mathbf{x}_{\text{ML}}$ if and only if α_{EM} and t_{EM} satisfy the conditions in Theorem 1. It remains to show that if $t_{\text{EM}} \in \mathcal{T}$ then it is the solution to $\min_{t\in\mathcal{T}} f(\alpha_{\text{EM}}, t)$, and that if $\alpha_{\text{EM}} \in \mathcal{A}$ then it is the solution to $\max_{\alpha\in\mathcal{A}} f(\alpha, t_{\text{EM}})$. The first property holds since for any (de)regularized LS with parameters $\alpha \in \mathcal{A}$ and $t = \|\mathbf{x}\|^2$ we have

$$\begin{aligned} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \quad &= \quad \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\mathbf{x} + \mathbf{x}^T\mathbf{H}^T\mathbf{H}\mathbf{x} \\ &= \quad \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\mathbf{x} + \mathbf{x}^T\left(\mathbf{H}^T\mathbf{H} + \alpha\mathbf{I}\right)\mathbf{x} - \alpha\mathbf{x}^T\mathbf{x} \\ &= \quad \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\mathbf{x} + \mathbf{y}^T\mathbf{H}\mathbf{x} - \alpha\mathbf{x}^T\mathbf{x} \\ &= \quad c(\alpha) - \alpha t, \end{aligned} \tag{26}$$

where we have used $\mathbf{x} = \left(\mathbf{H}^T\mathbf{H} + \alpha\mathbf{I}\right)^\dagger\mathbf{H}^T\mathbf{y}$ and $\mathbf{H}^T\mathbf{y} \in \mathcal{R}\left(\mathbf{H}^T\mathbf{H} + \alpha\mathbf{I}\right)$ in the third equality. Combining (23) with (26) yields

$$\alpha_{\text{EM}} = \sigma_h^2\left(N - \frac{c(\alpha_{\text{EM}}) - \alpha_{\text{EM}}t_{\text{EM}}}{\sigma_h^2 t_{\text{EM}} + \sigma^2}\right), \tag{27}$$

which can be easily seen to satisfy the condition in (16). The second property holds since for any (de)regularized LS, the equation

$$\|\left(\mathbf{H}^T\mathbf{H} + \alpha\mathbf{I}\right)^\dagger\mathbf{H}^T\mathbf{y}\|^2 = t, \tag{28}$$

has a unique root in $\alpha \in \mathcal{A}$ which is the optimal solution to $\max_{\alpha\in\mathcal{A}} f(\alpha, t_{\text{EM}})$ (see [18] for more details). ■

D. Comparison to EIV

We now compare the above results with the corresponding results in the EIV model. The ML estimator in model (1) and (3) estimates both $\mathbf{x}$ and $\mathbf{G}$ by solving

$$\max_{\mathbf{x},\mathbf{G}} \log f\left(\mathbf{y}, \mathbf{H}; \mathbf{x}, \mathbf{G}\right), \tag{29}$$

where $f\left(\mathbf{y},\mathbf{H};\mathbf{x},\mathbf{G}\right)$ is the joint probability density function of $\mathbf{y}$ and $\mathbf{H}$ parameterized by $\mathbf{x}$ and $\mathbf{G}$. Due to the Gaussian assumption, (29) is equivalent to

$$\widehat{\mathbf{x}}_{\text{TLS}} = \arg\min_{\mathbf{x}}\left\{\min_{\mathbf{G}}\left\{\frac{\|\mathbf{y}-\mathbf{G}\mathbf{x}\|^2}{\sigma^2}+\frac{\|\mathbf{H}-\mathbf{G}\|_F^2}{\sigma_h^2}\right\}\right\}, \tag{30}$$

where we intentionally insert the minimization with respect to $\mathbf{G}$ inside the objective in order to emphasize that $\mathbf{G}$ is a nuisance parameter. In the signal processing literature (30) is usually known as the (column-wise weighted) TLS estimator [12]. It is a generalization of the LS solution to the problem $\mathbf{y}\approx\mathbf{H}\mathbf{x}$ when both $\mathbf{y}$ and $\mathbf{H}$ are subject to measurement errors. It tries to find $\mathbf{x}$ and $\mathbf{G}$ which minimize the squared errors in $\mathbf{y}$ and in $\mathbf{H}$ as expressed in (30).

Under a simple condition which is usually satisfied, it can be shown that the TLS estimator in (30) is a deregularized LS solution [12]

$$\widehat{\mathbf{x}}_{\text{TLS}} = \left(\mathbf{H}^T\mathbf{H}+\alpha_{\text{TLS}}\mathbf{I}\right)^{-1}\mathbf{H}^T\mathbf{y}, \tag{31}$$

where

$$\alpha_{\text{TLS}} = -\lambda_{\min}\left(\left[\begin{array}{cc}\mathbf{H}^T\mathbf{H} & -\frac{\sigma_h}{\sigma}\mathbf{H}^T\mathbf{y}\\ -\frac{\sigma_h}{\sigma}\mathbf{y}^T\mathbf{H} & \frac{\sigma_h^2}{\sigma^2}\mathbf{y}^T\mathbf{y}\end{array}\right]\right). \tag{32}$$

Since our ML estimator is also a (de)regularized LS it is interesting to compare the two.

Proposition 1: The regularization parameters[1] of the ML estimator in (6) and the TLS in (31) satisfy $\alpha_{\text{TLS}}\le\alpha_{\text{ML}}$.

Proof: This relation holds since the objective in (5) is equal to the objective in (30) plus an additional logarithmic regularization term. In order to see this, we begin by minimizing (30) with respect to $\mathbf{G}$ first, and find that the TLS is the solution to

$$\widehat{\mathbf{x}}_{\text{TLS}} = \arg\min_{\mathbf{x}}\frac{\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2}{\sigma_h^2\|\mathbf{x}\|^2+\sigma^2}, \tag{33}$$

which is exactly (5) without the logarithmic penalty. Now, assume in contradiction that $\alpha_{\text{TLS}}>\alpha_{\text{ML}}$. Then, $\|\widehat{\mathbf{x}}_{\text{TLS}}\|^2<\|\widehat{\mathbf{x}}_{\text{ML}}\|^2$ and

$$\begin{aligned}&\frac{\|\mathbf{y}-\mathbf{H}\widehat{\mathbf{x}}_{\text{TLS}}\|^2}{\sigma_h^2\|\widehat{\mathbf{x}}_{\text{TLS}}\|^2+\sigma^2}+N\log(\sigma_h^2\|\widehat{\mathbf{x}}_{\text{TLS}}\|^2+\sigma^2)\\ &<\frac{\|\mathbf{y}-\mathbf{H}\widehat{\mathbf{x}}_{\text{TLS}}\|^2}{\sigma_h^2\|\widehat{\mathbf{x}}_{\text{TLS}}\|^2+\sigma^2}+N\log(\sigma_h^2\|\widehat{\mathbf{x}}_{\text{ML}}\|^2+\sigma^2)\\ &\le\frac{\|\mathbf{y}-\mathbf{H}\widehat{\mathbf{x}}_{\text{ML}}\|^2}{\sigma_h^2\|\widehat{\mathbf{x}}_{\text{ML}}\|^2+\sigma^2}+N\log(\sigma_h^2\|\widehat{\mathbf{x}}_{\text{ML}}\|^2+\sigma^2)\end{aligned} \tag{34}$$

which is a contradiction to the optimality of $\widehat{\mathbf{x}}_{\text{ML}}$. ■

Thus, the MLE can also be considered a regularized TLS estimator. Interestingly, the concept of regularizing the TLS estimator is not new [20], [21]. It is well known that the TLS solution is not stable when applied to ill-posed problems. It has been shown that in many applications regularizing the TLS objective may significantly improve the performance of the TLS estimator in terms of MSE. The ML estimator in the RV model provides statistical reasoning to this phenomenon and suggests an inherent logarithmic penalty scheme. More details on this property can be found in our earlier paper [8] and in [21], [22].

There is another interpretation for the difference between (5) and (33). We obtained these two estimators by optimizing the ML criterion in two different models. In turns out that the same estimators can be obtained using the RV model but by choosing two different optimization criteria. In particular, (33) can be interpreted as the joint maximum-a-posteriori ML (JMAP-ML) estimator in the RV model [23]. It has been shown in [23] that in the general Gaussian case the difference between the ML criterion and the JMAP-ML criterion is always a logarithmic penalty. Thus, the logarithmic regularization can be interpreted as a special case of the ML and JMAP-ML relation.

IV. Cramér Rao Bound

In the previous section, we discussed the MLE and numerical algorithms for finding it. Unfortunately, analytic performance evaluation may be intractable. Instead, we now provide an indication of performance using the CRB. The CRB is a lower bound for the MSE of any unbiased estimator [1], [2]. Moreover, it is well known that under a number of regularity conditions the MSE of the MLE asymptotically attains this bound. Therefore, the CRB is a reasonable metric for shedding more light on the performance of our estimators (and on the models themselves). A closed form expression of the CRB in our problem setting is provided in the following theorem.

Theorem 3: Consider the estimation of $\mathbf{x}$ in the model (2) when σ_h^2 and σ^2 are known. Then, the MSE of any unbiased estimator is lower bounded by

$$\mathbf{CRB}_{\text{RV}} = \left(\sigma_h^2\|\mathbf{x}\|^2+\sigma^2\right)\left(\mathbf{H}^T\mathbf{H}\right)^{-1}-\mathbf{\Delta}, \tag{35}$$

where $\mathbf{\Delta}\succeq\mathbf{0}$ is given by

$$\mathbf{\Delta} \;=\; \frac{1}{\gamma}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{x}\mathbf{x}^T\left(\mathbf{H}^T\mathbf{H}\right)^{-1} \tag{36}$$

with

$$\gamma \;=\; \frac{1}{2N\sigma_h^4\left(\sigma_h^2\|\mathbf{x}\|^2+\sigma^2\right)}+\frac{\mathbf{x}^T\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{x}}{\sigma_h^2\|\mathbf{x}\|^2+\sigma^2}. \tag{37}$$

Proof: The CRB is defined as

$$\begin{aligned}\mathbf{CRB}_{\text{RV}} &= \mathbf{J}^{-1}(\mathbf{x})\\ &= \left[-E\left\{\frac{\partial^2\log f\left(\mathbf{y};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T}\right\}\right]^{-1},\end{aligned} \tag{38}$$

where $\mathbf{J}\left(\mathbf{x}\right)$ is the Fisher information matrix (FIM) for estimating $\mathbf{x}$ given $\mathbf{y}$. Fortunately, we can exploit a closed form expression for the FIM in the case of a jointly Gaussian distribution of the observations:

Lemma 1 ([2]): Let $\mathbf{z}$ be a Gaussian vector with mean $\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)$ and covariance $\mathbf{C}\left(\boldsymbol{\theta}\right)$. Then the elements of the FIM for estimating $\boldsymbol{\theta}$ from $\mathbf{z}$ are

$$\begin{aligned}\left[\mathbf{J}\left(\boldsymbol{\theta}\right)\right]_{ij} &= \left[\frac{\partial\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\right]^T\mathbf{C}^{-1}\left(\boldsymbol{\theta}\right)\left[\frac{\partial\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)}{\partial\theta_j}\right]\\ &\quad+\frac{1}{2}\text{Tr}\left\{\mathbf{C}^{-1}\left(\boldsymbol{\theta}\right)\frac{\partial\mathbf{C}\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\mathbf{C}^{-1}\left(\boldsymbol{\theta}\right)\frac{\partial\mathbf{C}\left(\boldsymbol{\theta}\right)}{\partial\theta_j}\right\}.\end{aligned} \tag{39}$$

[1] That is of course under the technical assumption in [12] required for (31).

In our setting, $\mathbf{z} = \mathbf{y}$ and $\theta = \mathbf{x}$. Thus, $\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{H}\mathbf{x}$, $\mathbf{C}(\boldsymbol{\theta}) = (\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2)\mathbf{I}$, and

$$\mathbf{J}(\mathbf{x}) = \frac{\mathbf{H}^T\mathbf{H}}{\sigma_h^2\|\mathbf{x}\|^2+\sigma^2} + \frac{2N\sigma_h^4\mathbf{x}\mathbf{x}^T}{\left(\sigma_h^2\|\mathbf{x}\|^2+\sigma^2\right)^2}. \tag{40}$$

The CRB is then obtained by applying the matrix inversion lemma. ■

Theorem 3 allows us to compare the CRB in our uncertainty model with the CRB in a model where $\mathbf{G}$ is known. In the latter case, the CRB for estimating $\mathbf{x}$ is given by [2]

$$\mathbf{CRB}_{\text{known}} = \sigma^2\left(\mathbf{G}^T\mathbf{G}\right)^{-1}. \tag{41}$$

Indeed, substituting $\sigma_h^2 = 0$ in (35) yields

$$\mathbf{CRB}_{\text{known}} = \sigma^2\left(\mathbf{H}^T\mathbf{H}\right)^{-1}, \tag{42}$$

which is consistent with (41) since $\sigma_h^2 = 0$ implies that $\mathbf{G} = \mathbf{H}$. An important difference between (35) and (42) is that, unlike the known-$\mathbf{G}$ case, the CRB under uncertainty conditions depends on the specific value of $\mathbf{x}$. Thus, some $\mathbf{x}$'s are more difficult to estimate in this model than others, depending on $\mathbf{H}$.

Surprisingly, it is not trivial to compare the RV CRB with the known-$\mathbf{G}$ CRB. Examining the bounds carefully reveals that

$$\mathbf{CRB}_{\text{known}} \not\preceq \mathbf{CRB}_{\text{RV}}, \tag{43}$$

i.e., there exist parameters such that the RV CRB is lower than the known-$\mathbf{G}$ CRB. For example, if $K = 1$ then $\mathbf{CRB}_{\text{RV}}$ approaches zero much faster than $\mathbf{CRB}_{\text{known}}$. Although this may seem like a mistake, it is actually a feature of our model. Increasing the randomness in $\mathbf{G}$ via σ_h^2 has two effects: it accounts for the uncertainty in $\mathbf{G}$, but it also affects the relation between $\mathbf{y}$ and $\mathbf{x}$. As we will now show, the increased uncertainty in $\mathbf{G}$ degrades the MSE, but random perturbation itself can be beneficial in some scenarios. Thus, the overall performance may be improved.

To better understand these effects and decouple the contribution of each, let us first derive the CRB when $\mathbf{G}$ is random but known. Recall that when $\mathbf{G}$ is random, the bound is given by (38). Therefore, when $\mathbf{G}$ is known all we need to do is add $\mathbf{G}$ as an additional observation:

$$\mathbf{CRB}_{\text{known-RV}} = \left[-E\left\{\frac{\partial^2 \log f\left(\mathbf{y},\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T}\right\}\right]^{-1}. \tag{44}$$

The variables $\mathbf{y}$ and $\mathbf{G}$ are jointly Gaussian, and the bound can be derived in a straightforward manner using Lemma 1. Alternatively, the derivations can be simplified by conditioning on $\mathbf{G}$:

$$\begin{aligned}\frac{\partial^2\log f\left(\mathbf{y},\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T} &= \frac{\partial^2\log f\left(\mathbf{y}|\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T} + \frac{\partial^2\log f\left(\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T}\\ &= \frac{\partial^2\log f\left(\mathbf{y}|\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T},\end{aligned} \tag{45}$$

since the distribution of $\mathbf{G}$ does not depend on $\mathbf{x}$. Thus,

$$\begin{aligned}\mathbf{CRB}_{\text{known-RV}} &= \left[-E\left\{\frac{\partial^2\log f\left(\mathbf{y}|\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T}\right\}\right]^{-1}\\ &= \left[-E\left\{E\left\{\left.\frac{\partial^2\log f\left(\mathbf{y}|\mathbf{G};\mathbf{x}\right)}{\partial\mathbf{x}\mathbf{x}^T}\right|\mathbf{G}\right\}\right\}\right]^{-1}\\ &= \left[E\left\{\frac{\mathbf{G}^T\mathbf{G}}{\sigma^2}\right\}\right]^{-1}\\ &= \sigma^2\left(\mathbf{H}^T\mathbf{H} + N\sigma_h^2\mathbf{I}\right)^{-1},\end{aligned} \tag{46}$$

where the first equality is due to (45), the third equality is given by the inverse of (41), and the last equality is obtained by taking the expectation with respect to $\mathbf{G}$. From (46) we see that if $\mathbf{G}$ were known, its randomness would improve the performance:

$$\mathbf{CRB}_{\text{known-RV}} \preceq \mathbf{CRB}_{\text{known}}, \tag{47}$$

since the additional $N\sigma_h^2\mathbf{I}$ term in (46) decreases its inverse. An intuitive explanation is that since the performance depends on $\mathbf{G}^T\mathbf{G}$, the squared effect of the random perturbations in $\mathbf{G}$ implies that realizations of $\mathbf{G}$ with increased $\mathbf{G}^T\mathbf{G}$ are more dominant, on the average, than realizations with decreased $\mathbf{G}^T\mathbf{G}$. Thus, the effect of known perturbations is equivalent to an increased signal ($\mathbf{G}\mathbf{x}$) to noise ($\mathbf{w}$) ratio.

On the other hand, when the perturbation in $\mathbf{G}$ is random, the uncertainty degrades the performance:

$$\mathbf{CRB}_{\text{known-RV}} \preceq \mathbf{CRB}_{\text{RV}}. \tag{48}$$

The proof of (48) is trivial by comparing the closed form expressions for the bounds in (35) and (46), or by comparing the FIMs in (38) and (44) and noting that additional observations provide more information.

Thus, known random perturbations always improve the performance in our model, as can be intuitively explained. Surprisingly, unknown random perturbations may also be beneficial as expressed in (43). Nonetheless, here the relation is unclear and there is no simple ordering between the two bounds. Moreover, there is an important difference: unlike known randomness which improves the performance uniformly in $\mathbf{x}$, the improvement due to unknown randomness depends on the specific value of $\mathbf{x}$. Roughly speaking, when $\mathbf{G}$ is known, the additional information on $\mathbf{x}$ is available through the mean of $\mathbf{y}$, whereas when $\mathbf{G}$ is unknown, additional information originates from the covariance matrix of $\mathbf{y}$ ($\mathbf{W}$ takes the role of multiplicative noise). This covariance contains information only on $\|\mathbf{x}\|^2$ and not on $\mathbf{x}$ itself. Practically, we can only use the covariance information for estimating $\|\mathbf{x}\|^2$, or (if we are fortunate enough) use the combined (mean and covariance) information to improve upon the mean-only based estimate of other specific functions of $\mathbf{x}$ with gradient components in the direction of $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{x}$. Other functions will not gain much from the unknown randomness. Furthermore, in Section V we will discuss another reservation that distinguishes the known perturbations case from the case of unknown perturbations.

The above discussion demonstrates that using the CRBs in the presence of random nuisance parameters is non trivial.

For completeness, we now provide a brief review on the literature in this context. First, the expression in (46) can also be derived as an approximation of the CRB with unknown random nuisance parameters [24]. This approximation is useful when the CRB is too complicated to compute. This is not the case here. Indeed, we have already derived the CRB and it is given by $\mathbf{CRB}_{\text{RV}}$ in (35). Furthermore, it is tempting to compare $\mathbf{CRB}_{\text{RV}}$ with $E\{\mathbf{CRB}_{\text{known}}\}$ where $\mathbf{CRB}_{\text{known}}$ is given in (41) and the expectation is taken with respect to $\mathbf{G}$. The idea is that for each realization of $\mathbf{G}$ $\mathbf{CRB}_{\text{known}}$ bounds the MSE (with respect to the additive noise $\mathbf{w}$) conditioned on $\mathbf{G}$, and the averaged MSE is bounded by its expected value. However, here too, the ordering is non-trivial and it is easy to find specific parameters in which $E\{\mathbf{CRB}_{\text{known}}\} \not\preceq \mathbf{CRB}_{\text{RV}}$. A possible explanation is that the left-hand side is a bound on the MSE of estimators which are unbiased for each $\mathbf{G}$, whereas the right-hand side bounds estimators which are only unbiased "on the average" with respect to $\mathbf{G}$. Clearly, the latter constraint is less strict and allows for a larger set of estimators. Again, the same approach is discussed in the context of bounding the MSE in the presence of unknown random nuisance parameters [25]. A comprehensive discussion on these bounds and their relations may be found in [26], [27] and references within.

A. Comparison with EIV

We now turn to compare the previous results with the CRB in the EIV model. Here too the derivation of the CRB is rather straightforward using Lemma 1. The only difference is that we need to derive the CRB for estimating both $\mathbf{x}$ and $\mathbf{G}$ and then quantify the degradation in performance due to the uncertainty in $\mathbf{G}$ using the matrix inversion lemma (See Appendix C). The final result is:

$$\mathbf{CRB}_{\text{EIV}} = \left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)\left(\mathbf{G}^T\mathbf{G}\right)^{-1}. \tag{49}$$

This bound has already appeared in [28], [29]. Moreover, small error analysis of the TLS estimator proved that it is tight when σ_h^2 is sufficiently small.

Particularizing the bound to the known-$\mathbf{G}$ case ($\sigma_h^2 = 0$) yields (41). Unlike our previous results in the RV model, it is easy to see that the uncertainty in $\mathbf{G}$ always degrades the performance in the EIV model:

$$\mathbf{CRB}_{\text{known}} \preceq \mathbf{CRB}_{\text{EIV}}. \tag{50}$$

This is expected as when we increase σ_h^2 we add uncertainty but do not change the relation between $\mathbf{y}$ and $\mathbf{x}$ ($\mathbf{y}$ does not depend on σ_h^2 in the EIV model). In fact, the bound for the EIV model can be interpreted as the bound for a standard known-$\mathbf{G}$ model suffering from an additional independent noise term with variance $\sigma_h^2\|\mathbf{x}\|^2$. It is interesting to note a duality with the RV model in this context: In the RV model this additional noise term is indeed part of the data generation model, hence its information content (on $\|\mathbf{x}\|^2$) can be exploited to improve the mean-based estimate of $\mathbf{x}$, as explained above. However, in the EIV model this additional noise term is not actually part of the measurement - it merely serves to decrease the bound, as if it did not contain any information on $\mathbf{x}$.

Another difference between the RV and EIV models is that the performance in the RV model is a function of $\mathbf{H}$ (the mean channel) whereas the performance in the EIV model is a function of $\mathbf{G}$ (the true channel). Thus, it is not fair to compare them directly. Nonetheless, if we ignore this fact for the moment and let $\mathbf{H}$ and $\mathbf{G}$ play the same role, it is easy to see that (49) is similar to (35) except for the $\boldsymbol{\Delta}$ term. Thus, in some way the estimation of $\mathbf{x}$ is easier in the RV model than in the EIV model. This is easy to explain in view of the previous discussion since both models introduce uncertainty in $\mathbf{G}$ but the randomness in the RV model can also be beneficial.

V. Estimation with unknown variances

In the previous sections, we discussed the estimation of $\mathbf{x}$ when the variances σ_h^2 and σ^2 are known deterministic parameters, and derived the associated ML estimator and CRB. In practice, it is not clear whether this information is always available. Therefore, we now focus on the more difficult case in which the variances are unknown nuisance parameters. As before, we begin with ML estimation and then analyze the inherent performance limitations using the CRB.

The main result is summarized in the following theorem:

Theorem 4: Consider the estimation of $\mathbf{x}$ in (2) when σ_h^2 and σ^2 are unknown deterministic nuisance parameters. Then, the ML estimator of $\mathbf{x}$ is the standard LS solution to

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2, \tag{51}$$

and the CRB reduces to

$$\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)\left(\mathbf{H}^T\mathbf{H}\right)^{-1}. \tag{52}$$

Proof: See Appendices C and D. ■

Thus, when the variances are unknown, the ML estimator in the RV model coincides with the standard LS estimator, whereas the CRB is simplified to the expression of the known-$\mathbf{G}$ CRB with an effective noise variance of $\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)$. The lack of knowledge of the variances causes the correction term $\boldsymbol{\Delta}$ in (35) to disappear. There is a simple intuitive explanation for these properties. As we already explained, the randomness of $\mathbf{G}$ has a negative effect but also a positive effect as it provides more information on $\|\mathbf{x}\|^2$ through the covariance of $\mathbf{y}$ given by $\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)\mathbf{I}$. However, the positive effect can only be realized if we know σ_h^2 and σ^2 and can somehow estimate the contribution of $\|\mathbf{x}\|^2$ to this covariance. This is another difference between the case of known perturbations and unknown perturbations, as in order to utilize the known perturbations we do not need any knowledge of the variances (the information is completely contained in the mean of $\mathbf{y}$ rather than in its covariance).

A. Comparison with EIV

There are many results in the literature on the EIV model with unknown noise variances. This is a very difficult estimation problem and there are still many open questions regarding it. Interestingly, it was shown that, under this setting, the log-likelihood function does not have a maximum and the ML estimator does not exist ([16] and references within). In fact, to our knowledge there is no consistent estimator of $\mathbf{x}$

in this model unless σ_h^2, σ_h^2/σ^2 or some other instrumental variable is known [11]. In our view, this means that our ML estimator which reduces to the standard LS method is a practical approach for avoiding the problematic (and non existing) EIV ML estimator using a different mathematical model.

In Appendix C, we obtain the following CRB for estimating $\mathbf{x}$ in the EIV model when σ^2 and σ_h^2 are unknown nuisance parameters:

$$\mathbf{CRB}_{\text{EIV}} = \left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)\left(\mathbf{G}^T\mathbf{G}\right)^{-1}. \tag{53}$$

As before, in the EIV model the performance is a function of $\mathbf{G}$ rather than $\mathbf{H}$. Yet, bearing this in mind, it is tempting to compare (52) with (53) and conclude that when the variances are unknown nuisance parameters the two bounds are similar.

VI. Numerical results

We now provide a few numerical examples illustrating the behavior of the proposed ML estimator.

Example I: In the first example, we consider the classical linear regression problem of fitting a line to noisy measurements [30]. Let $\mathbf{a}$ be a vector with N uniformly spaced samples on the interval $[-1, 1]$, and assume that $\mathbf{y} \approx \mathbf{a}x$ are noisy observations of a line with an unknown slope x. We are interested in estimating x given $\mathbf{y}$. If $\mathbf{a}$ is exactly known and only $\mathbf{y}$ is noisy, then the ML method coincides with the standard LS estimator

$$\widehat{\mathbf{x}}_{\text{LS}} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{y} \tag{54}$$

with $\mathbf{G} = \mathbf{a}$. In the following examples, we consider this problem under uncertainty conditions on $\mathbf{a}$.

First, we concentrate on the RV model. We assume that $\mathbf{a}$ consists of the mean values of the true samples, rather than the samples themselves. Thus, $\mathbf{H} = \mathbf{a}$ and $\mathbf{G}$ is a random matrix. The parameters are $N = 10$, $x = 1$, $\sigma_h^2 = \sigma^2$. We provide the empirical MSEs over 200 trials of four estimators: the clairvoyant LS in (54), the mismatched LS with $\mathbf{G}$ replaced by $\mathbf{H}$, the ML in (4) solved by the minimax approach in Section III-A, and the EIV-ML (TLS) in (29) implemented using the EVD as expressed in (31). The MSEs are then compared to $\mathbf{CRB}_{\text{RV}}$. The results are plotted in Fig. 1. The first observation is that the CRB is non informative in this case. As expected, the RV-ML estimator performs better than the mismatched LS estimator. The performance of the EIV-ML estimator is very poor and it is clear that it is not appropriate for estimation in the RV model in this example.

Next, we focus on the EIV model and assume that $\mathbf{G} = \mathbf{a}$ consists of the unknown samples, and $\mathbf{H}$ is a noisy observation on $\mathbf{G}$. We consider the same parameters and estimators as before. The estimated MSEs along with $\mathbf{CRB}_{\text{EIV}}$ are plotted in Fig. 2. Unlike the previous case, the CRB in the EIV model appears to be reasonably tight. The two LS estimators perform as expected but the performance of the ML estimators is surprising. The mismatched RV-ML performs considerably better than the more appropriate EIV-ML estimator. In fact, this example demonstrates the instability of the EIV-ML estimator in low signal-to-noise-ratios.

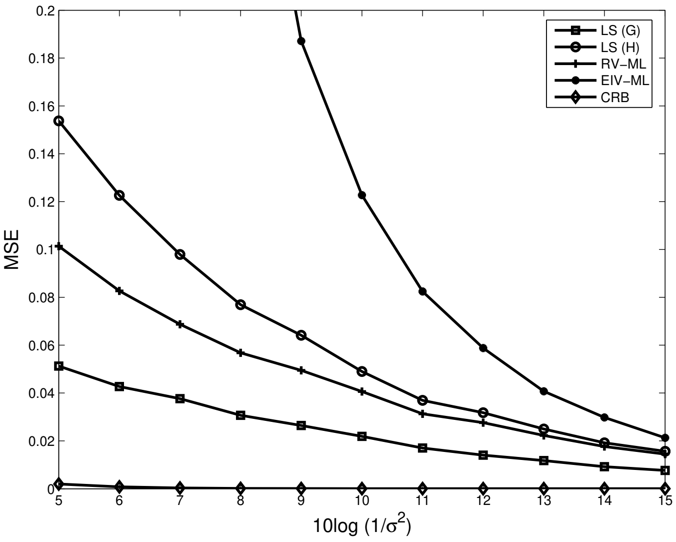

Fig. 1. MSE in estimating the slope of a straight line in the RV model.

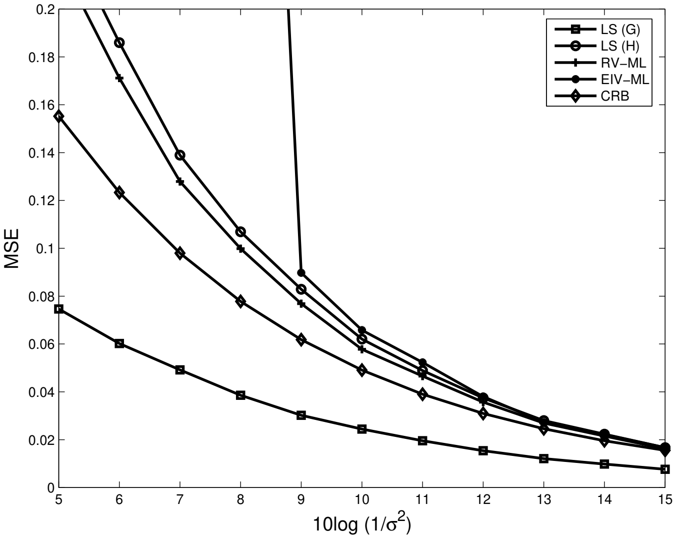

Fig. 2. MSE in estimating the slope of a straight line in the EIV model.

Example II: We examine the asymptotic performance of the estimators in the RV model. The parameters in our simulation are as follows. The matrix $\mathbf{H}$ is chosen as a concatenation of T matrices of size 5×5 with unit diagonal elements and 0.5 off-diagonal elements. We expect the ML estimator to attain its asymptotic performance as T increases, therefore we choose $T = 50$. The vector $\mathbf{x}$ was chosen as the normalized eigenvector of $\mathbf{H}^T\mathbf{H}$ associated with its minimal eigenvalue. We present the empirical MSEs of the four estimators defined above in Fig. 3 for $\sigma_h^2 = 0.1$. As before, the MSEs of the mismatched EIV-ML estimator were significantly worse than the others and were therefore omitted from the graph. It is easy to see the advantage of the RV-ML estimator over the mismatched LS estimator. As expected, the MSE of the RV-ML estimator approaches the CRB when σ^2 is sufficiently low.

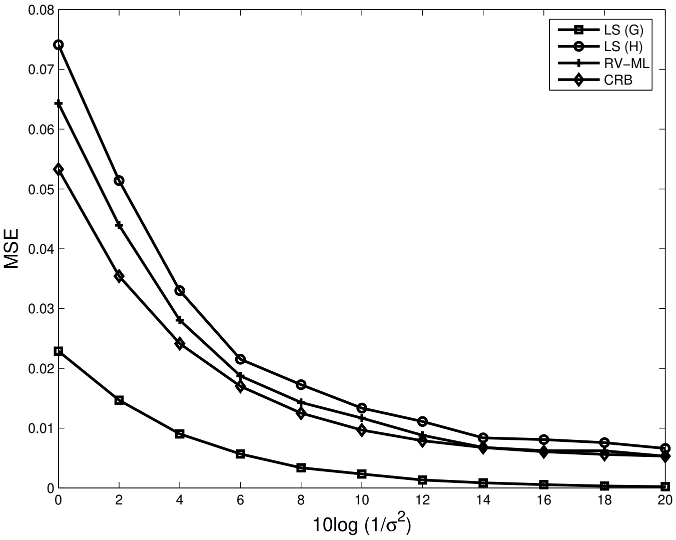

Fig. 3. Approaching the asymptotic MSE in the RV model.

VII. Conclusions

We considered the problem of linear regression in a Gaussian uncertainty model. We focused on ML estimation and proved that it is a regularized LS. We characterized the optimal regularization parameter using saddle point theory and provided efficient numerical methods for finding it. In addition, we analyzed the traditional EM solution to our problem using this new characterization. Next, we addressed the inherent performance limitations using the CRB. We quantified the degradation in performance due to the uncertainty, and showed that it is less severe than expected. We explained this property by noting that our model introduces uncertainty into $\mathbf{G}$ but also randomness. The uncertainty is clearly undesirable, but under some reservations the randomness itself may be beneficial. Next, we considered ML estimation and CRB analysis when the variances of the random variables are unknown nuisance parameters. Finally, using a few simple numerical results we demonstrated the instability of the EIV-ML estimator, and the advantage of the RV-ML estimator (in both models).

We note in addition, that the numerical algorithms, as well as some of the CRB results, are firmly based on the assumption of independent, identically distributed perturbations in the elements of the model matrix (as well as in the additive noise term w). In realistic situations this assumption may not be valid, and adaptation of some of our results to different distributions (even within the Gaussian framework) may be non-trivial, if at all possible.

In our view, the main contribution of this paper is in providing more insight into the different uncertainty models. Estimation under uncertainty conditions is an important problem in modern statistical signal processing. Our results show that the first step in such problems is to properly define the setting. Different uncertainty models give rise to different algorithms, and different performance measures. We believe that in order to decide which model fits a specific application, one must fully understand the theoretical differences between these models and the advantages and disadvantages of each.

Acknowledgments

The first author would like to thank Dr. Amir Beck for many fruitful discussions and insightful suggestions during the preparation of this paper.

Appendix A
Proof of quasi-convexity of $f(\alpha, t)$ in t

The proof is based on the following result:

Lemma 2 ([17]): If $r'(t) = 0$ implies $r''(t) > 0$ for any $t \geq 0$, then $r(t)$ is unimodal in $t \geq 0$.

The condition $\frac{\partial f(\alpha,t)}{\partial t} = 0$ is given in (16). Multiplying it by $\frac{\sigma_h^2}{\sigma_h^2 t+\sigma^2}$ yields:

$$\frac{\sigma_h^4 c(\alpha) - \sigma_h^4 \alpha t}{\left(\sigma_h^2 t + \sigma^2\right)^3} = \frac{N\sigma_h^4 - \sigma_h^2 \alpha}{\left(\sigma_h^2 t + \sigma^2\right)^2}. \tag{55}$$

The second derivative is

$$\begin{aligned}\frac{\partial^2 f(\alpha,t)}{\partial t^2} \quad = \quad & \frac{\sigma_h^2 \alpha}{\left(\sigma_h^2 t + \sigma^2\right)^2} + \frac{2\sigma_h^4 c(\alpha)}{\left(\sigma_h^2 t + \sigma^2\right)^3} + \frac{\sigma_h^2 \alpha}{\left(\sigma_h^2 t + \sigma^2\right)^2} \\ & - \frac{2\sigma_h^4 \alpha t}{\left(\sigma_h^2 t + \sigma^2\right)^3} - \frac{N\sigma_h^4}{\left(\sigma_h^2 t + \sigma^2\right)^2}. \end{aligned} \tag{56}$$

Plugging in the left hand side of (55) yields

$$\frac{\partial^2 f(\alpha,t)}{\partial t^2} \quad = \frac{N\sigma_h^4}{\left(\sigma_h^2 t+\sigma^2\right)^2} > 0, \tag{57}$$

which concludes the proof.

Appendix B
EM solution of the ML problem

In this appendix, we provide the derivation of the EM algorithm in (19)-(20). At each iteration, the algorithm maximizes the expected log likelihood (with respect to $\mathbf{G}$ given $\mathbf{y}$)

$$\begin{aligned}\mathbf{x}_{n+1} \quad &= \quad \arg\max_{\mathbf{x}} E\left\{ \log f(\mathbf{y}, \mathbf{G}; \mathbf{x}) \middle| \, \mathbf{y}; \mathbf{x}_n \right\} \\ &= \quad \arg\max_{\mathbf{x}} E\left\{ \log f(\mathbf{y}|\mathbf{G}; \mathbf{x}) + \log f(\mathbf{G}; \mathbf{x}) \middle| \, \mathbf{y}; \mathbf{x}_n \right\} \\ &= \quad \arg\min_{\mathbf{x}} E\left\{ \|\mathbf{y} - \mathbf{G}\mathbf{x}\|^2 \middle| \, \mathbf{y}; \mathbf{x}_n \right\} \\ &= \quad \left[\overline{\mathbf{G}_n^T \mathbf{G}_n}\right]^{-1} \overline{\mathbf{G}_n}^T \mathbf{y}, \end{aligned} \tag{58}$$

where

$$\begin{aligned}\overline{\mathbf{G}_n} \quad &= \quad E\left\{ \mathbf{G} \middle| \, \mathbf{y}; \mathbf{x}_n, \mathbf{H}_n \right\} \\ \overline{\mathbf{G}_n^T \mathbf{G}_n} \quad &= \quad E\left\{ \mathbf{G}^T \mathbf{G} \middle| \, \mathbf{y}; \mathbf{x}_n, \mathbf{H}_n \right\}. \end{aligned} \tag{59}$$

Fortunately enough, the expectations in (59) can be easily evaluated based on jointly Gaussian optimal MMSE estimation theory [2]. Using the Kronecker product we have

$$\mathbf{y} = \left(\mathbf{x}^T \otimes \mathbf{I}\right) \mathbf{g} + \mathbf{w}, \tag{60}$$

where $\mathbf{g} = \operatorname{vec}(\mathbf{G})$. The Bayesian MMSE estimator of $\mathbf{g}$ given $\mathbf{y}$ in (60) satisfies

$$E\{\mathbf{g}|\mathbf{y}; \mathbf{x}_n, \mathbf{H}_n\} = \operatorname{vec}(\mathbf{H}_n) + \frac{\sigma_h^2 \left(\mathbf{x}_n \otimes \mathbf{I}\right)\left[\mathbf{y} - \mathbf{H}_n \mathbf{x}_n\right]}{\sigma_h^2 \|\mathbf{x}_n\|^2 + \sigma^2}, \tag{61}$$

and

$$\mathrm{COV}\{\mathbf{g}|\mathbf{y}; \mathbf{x}_n, \mathbf{H}_n\} = \sigma_h^2 \mathbf{I} - \frac{\sigma_h^4 \left(\mathbf{x}_n \mathbf{x}_n^T \otimes \mathbf{I}\right)}{\sigma_h^2 \|\mathbf{x}_n\|^2 + \sigma^2}, \tag{62}$$

where COV$\{\cdot\}$ is the corresponding covariance matrix. Using (61)-(62) and straightforward algebraic manipulations yields the first and second moments of $\mathbf{G}$ given in (19)-(20).

Appendix C
Derivations of the CRBs

A. Proof of (49)

In the EIV model, $\mathbf{z} = \left[\mathbf{y}^T\ \mathbf{h}^T\right]^T$ and $\boldsymbol{\theta} = \left[\mathbf{x}^T\ \mathbf{g}^T\right]^T$ where $\mathbf{h} = \mathrm{vec}\left(\mathbf{H}\right)$ and $\mathbf{g} = \mathrm{vec}\left(\mathbf{G}\right)$. Therefore,

$$\begin{aligned} \boldsymbol{\eta}\left(\boldsymbol{\theta}\right) &= \begin{bmatrix} \left(\mathbf{x}^T \otimes \mathbf{I}\right)\mathbf{g} \\ \mathbf{g} \end{bmatrix} \\ \mathbf{C}\left(\boldsymbol{\theta}\right) &= \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_h^2\mathbf{I} \end{bmatrix}, \end{aligned} \tag{63}$$

the CRB is given by the top left sub-block of $\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)$ in Lemma 1. Using a well-known matrix inversion relation we obtain

$$\mathbf{CRB} = \left(\mathbf{J}_{\mathbf{xx}} - \mathbf{J}_{\mathbf{xg}}\mathbf{J}_{\mathbf{gg}}^{-1}\mathbf{J}_{\mathbf{gx}}\right)^{-1}, \tag{64}$$

where

$$\begin{aligned} \mathbf{J}_{\mathbf{xx}} &= \frac{\mathbf{G}^T\mathbf{G}}{\sigma^2} \\ \mathbf{J}_{\mathbf{xg}} &= \mathbf{J}_{\mathbf{gx}}^T = \frac{\mathbf{G}^T\left(\mathbf{x}^T \otimes \mathbf{I}\right)}{\sigma^2} \\ \mathbf{J}_{\mathbf{gg}} &= \frac{\left(\mathbf{x}\mathbf{x}^T \otimes \mathbf{I}\right)}{\sigma^2} + \frac{1}{\sigma_h^2}\mathbf{I}. \end{aligned} \tag{65}$$

Using the properties of the Kronecker product, we obtain

$$\begin{aligned} &\mathbf{J}_{\mathbf{xg}}\mathbf{J}_{\mathbf{gg}}^{-1}\mathbf{J}_{\mathbf{gx}} \\ &= \frac{\mathbf{G}^T\left(\mathbf{x}^T \otimes \mathbf{I}\right)}{\sigma^2}\left[\left(\frac{\mathbf{x}\mathbf{x}^T}{\sigma^2} + \frac{\mathbf{I}}{\sigma_h^2}\right)^{-1} \otimes \mathbf{I}\right]\frac{\left(\mathbf{x} \otimes \mathbf{I}\right)\mathbf{G}}{\sigma^2} \\ &= \frac{\sigma_h^2\|\mathbf{x}\|^2}{\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2}\frac{\mathbf{G}^T\mathbf{G}}{\sigma^2}, \end{aligned} \tag{66}$$

and

$$\mathbf{J}_{\mathbf{xx}} - \mathbf{J}_{\mathbf{xg}}\mathbf{J}_{\mathbf{gg}}^{-1}\mathbf{J}_{\mathbf{gx}} = \frac{\mathbf{G}^T\mathbf{G}}{\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2}, \tag{67}$$

as required.

B. Proof of (52)

In this scenario, the unknown parameters are $\boldsymbol{\theta} = \left[\mathbf{x}^T\ \sigma^2\ \sigma_h^2\right]^T$ and $\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{H}\mathbf{x}$, $\mathbf{C}(\boldsymbol{\theta}) = (\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2)\mathbf{I}$. Applying Lemma 1 yields

$$\begin{aligned} \mathbf{J}_{\mathbf{xx}} &= \frac{\mathbf{H}^T\mathbf{H}}{\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2} + \frac{2N\sigma_h^4\mathbf{x}\mathbf{x}^T}{\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ \mathbf{J}_{\sigma_h^2\sigma_h^2} &= \frac{N\|\mathbf{x}\|^4}{2\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ \mathbf{J}_{\sigma^2\sigma^2} &= \frac{N}{2\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ \mathbf{J}_{\mathbf{x}\sigma_h^2} &= \mathbf{J}_{\sigma_h^2\mathbf{x}}^T = \frac{N\sigma_h^2\|\mathbf{x}\|^2\mathbf{x}}{\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ \mathbf{J}_{\mathbf{x}\sigma^2} &= \mathbf{J}_{\sigma^2\mathbf{x}}^T = \frac{N\sigma_h^2\mathbf{x}}{\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ \mathbf{J}_{\sigma_h^2\sigma^2} &= \mathbf{J}_{\sigma^2\sigma_h^2} = \frac{N\|\mathbf{x}\|^2}{2\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2}. \end{aligned} \tag{68}$$

The CRB is given by the top left sub block of $\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)$ which is equivalent to

$$\mathbf{CRB} = \left(\mathbf{J}_{\mathbf{xx}} - \mathbf{J}_{\mathbf{x}\boldsymbol{\sigma}^2}\mathbf{J}_{\boldsymbol{\sigma}^2\boldsymbol{\sigma}^2}^{\dagger}\mathbf{J}_{\boldsymbol{\sigma}^2\mathbf{x}}\right)^{-1}, \tag{69}$$

where

$$\begin{aligned} &\mathbf{J}_{\mathbf{x}\boldsymbol{\sigma}^2}\mathbf{J}_{\boldsymbol{\sigma}^2\boldsymbol{\sigma}^2}^{\dagger}\mathbf{J}_{\boldsymbol{\sigma}^2\mathbf{x}} \\ &= \begin{bmatrix} \mathbf{J}_{\mathbf{x}\sigma_h^2} & \mathbf{J}_{\mathbf{x}\sigma^2} \end{bmatrix}\begin{bmatrix} \mathbf{J}_{\sigma_h^2\sigma_h^2} & \mathbf{J}_{\sigma_h^2\sigma^2} \\ \mathbf{J}_{\sigma^2\sigma_h^2} & \mathbf{J}_{\sigma^2\sigma^2} \end{bmatrix}^{\dagger}\begin{bmatrix} \mathbf{J}_{\sigma_h^2\mathbf{x}} \\ \mathbf{J}_{\sigma^2\mathbf{x}} \end{bmatrix} \\ &= \frac{2N\sigma_h^4\mathbf{x}\begin{bmatrix} \|\mathbf{x}\|^2 & 1 \end{bmatrix}\begin{bmatrix} \|\mathbf{x}\|^4 & \|\mathbf{x}\|^2 \\ \|\mathbf{x}\|^2 & 1 \end{bmatrix}^{\dagger}\begin{bmatrix} \|\mathbf{x}\|^2 \\ 1 \end{bmatrix}\mathbf{x}^T}{\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2} \\ &= \frac{2N\sigma_h^4}{\left(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right)^2}\mathbf{x}\mathbf{x}^T. \end{aligned} \tag{70}$$

Plugging (70) into (69) yields (52).

Note that we have applied the CRB of singular Fisher information matrices [1], [31]. The practical meaning of the singularity in (69)-(70) is that it is impossible to differentiate between σ_h^2 and σ^2 in the model. Fortunately, this is not crucial for the estimation of $\mathbf{x}$ since all we are interested in is an estimate of the effective variance $\left[\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2\right]$.

C. Proof of (53)

Here, $\mathbf{z} = \left[\mathbf{y}^T\ \mathbf{h}^T\right]^T$ and $\boldsymbol{\theta} = \left[\mathbf{x}^T\ \mathbf{g}^T\ \sigma_h^2\ \sigma^2\right]^T$ and $\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)$ and $\mathbf{C}\left(\boldsymbol{\theta}\right)$ are defined in (63). It is easy to see that the cross terms between $\{\mathbf{x}, \mathbf{g}\}$ and $\{\sigma^2, \sigma_h^2\}$ in the FIM are all zero. Therefore, the lack of knowledge of the variances does not change the CRB which is equal to (49).

Appendix D
Proof of Theorem 4

The ML estimator of $\mathbf{x}$ in the RV model when σ_h^2 and σ^2 are unknown deterministic nuisance parameters is the solution to

$$\min_{\mathbf{x}}\left\{\min_{\sigma_h^2 \geq 0, \sigma^2 \geq 0}\left\{\frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2} + N\log(\sigma_h^2\|\mathbf{x}\|^2 + \sigma^2)\right\}\right\}. \tag{71}$$

First, we use a change of variables and replace σ^2 by the effective variance $\xi = \sigma_h^2\|\mathbf{x}\|^2 + \sigma^2$. Problem (71) is then

$$\min_{\mathbf{x}} \left\{ \min_{\xi \geq \sigma_h^2\|\mathbf{x}\|^2, \sigma_h^2 \geq 0} \left\{ \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\xi} + N\log(\xi) \right\} \right\}. \quad (72)$$

Now, the optimal σ_h^2 is any non negative number satisfying $\xi \geq \sigma_h^2\|\mathbf{x}\|^2$, and does not effect the objective function. Without loss of generality, we choose $\sigma_h^2 = 0$ and obtain

$$\min_{\mathbf{x}} \left\{ \min_{\xi \geq 0} \left\{ \frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\xi} + N\log(\xi) \right\} \right\}. \quad (73)$$

Simple differentiation yields the optimal ξ

$$\xi = \frac{1}{N}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \geq 0. \quad (74)$$

Plugging (74) back into (73) results in

$$\min_{\mathbf{x}} \left\{ N + N\log\left(\frac{1}{N}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \right) \right\}. \quad (75)$$

which is equivalent to

$$\min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \right\}. \quad (76)$$

References

[1] C. R. Rao, *Linear statistical inference and its applications*. John Wiley and Sons, Inc., 1973.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall, 1993.

[3] A. N. Tikhonov and V. Y. Arsenin, "Solution of ill posed problems," in *Washington, DC: V.H. Winston*, 1997.

[4] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1961.

[5] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean-squared error estimation in the presence of model uncertainties," *IEEE Trans. Signal Process.*, vol. 53, pp. 168–181, January 2005.

[6] L. E. Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal Matrix Analysis and Applications*, pp. 1034–1064, Oct. 1997.

[7] S. Chandrasekaran, M. G. G. H. Golub, and A. H. Sayed, "Parameter estimation in the presence of bounded data uncertainties," *SIAM Journal Matrix Analysis and Applications*, pp. 235–252, Jan. 1998.

[8] A. Wiesel, Y. C. Eldar, and A. Beck, "Maximum likelihood estimation in linear models with a gaussian model matrix," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 292–295, May 2006.

[9] H. A. Hindi and S. P. Boyd, "Robust solutions to l_1, l_2, and l_∞ uncertain linear approximation problems using convex optimization," in *Proc. of the 1998 American Control Conference*, June 1998.

[10] Y. C. Eldar, "Minimax estimation of deterministic parameters in linear models with a random model matrix," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 601–612, Feb. 2006.

[11] W. A. Fuller, *Measurement error models*. Wiley Series in Probability and Mathematical Statistics, 1987.

[12] S. V. Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. Frontiers in Applied Mathematics 9, Siam, 1991.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, December 1977.

[14] Y. Li, C. N. Georghiades, and G. Huang, "Iterative maximum-likelihood sequence estimation for space-time coded systems," *IEEE Trans. Commun.*, vol. 49, no. 6, pp. 948–951, June 2001.

[15] D. K. C. So and R. S. Cheng, "Iterative EM receiver for space-time coded systems in MIMO frequency-selective fading channels with channel gain and order estimation," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 1928–1935, Nov 2004.

[16] P. Stoica and J. Li, "On nonexistence of the maximum likelihood estimate in blind multichannel identification," *IEEE Signal Processing Magazine*, vol. 22, pp. 99– 101, July 2005.

[17] S. Boyd and L. Vandenberghe, *Introduction to Convex Optimization with Engineering Applications*. Stanford, 2003.

[18] Y. C. Eldar and A. Beck, "Hidden convexity based near maximum-likelihood CDMA detection," in *Proc. of conference on Information Sciences and Systems, Princeton, (CISS-2005)*, March 2005.

[19] M. Sion, "On general minimax theorems," *Pac. J. Math*, vol. 8, p. 171176, 1958.

[20] P. C. H. G. H. Golub and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM Journal Matrix Analysis and Applications*, Oct. 1999.

[21] A. Beck and A. Ben-Tal, "On the solution of the tikhonov regularization of the total least squares," *SIAM Journal on Optimization*, vol. 17, no. 1, pp. 98–118.

[22] A. Beck, A. Ben-Tal, and M. Teboulle, "Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares," *SIAM Journal on Matrix Anal. Appl*, vol. 28, no. 2, pp. 425–445, 2006.

[23] A. Yeredor, "The joint MAP-ML and its relation to ML and to extended least-squares," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3484–3492, Dec. 2000.

[24] F. Gini, R. Reggiannini, and U. Mengali, "The modified Cramer Rao bound in vector parameter estimation," *IEEE Trans. Commun.*, vol. 46, no. 1, pp. 2120–2127, Jan. 1998.

[25] R. Miller and C. Chang, "A modified Cramer-Rao bound and its applications," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 398 – 400, May 1978.

[26] I. Reuven and H. Messer, "A barankin-type lower bound on the estimation error of a hybrid parameter vector," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 1084–1093, May 1997.

[27] F. Gini and R. Reggiannini, "On the use of Cramer Rao Like bounds in the presence of random nuisance parameters," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2120–2127, Dec. 2000.

[28] M. J. Levin, "Estimation of a system pulse transfer function in the presence of noise," *IEEE Trans. on Automatic Control*, vol. AC-9, pp. 229–235, July 1964.

[29] J. J. Fuchs and S. Maria, "A new approach to variable selection using the TLS approach," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 10–19, Jan. 2007.

[30] A. Wald, "The fitting of straight lines if both variables are subject to error," *The Annals of Mathematical Statistics*, vol. 11, no. 3, pp. 284–300, Sept. 1940.

[31] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2026–2041, Aug. 1996.