Fundamental Performance Analysis in Image Registration Problems: Cramér-Rao Bound and its Variations

Dan Ruan and Jeffrey A. Fessler

COMMUNICATIONS & SIGNAL PROCESSING LABORATORY
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, Michigan 48109-2122

March 17, 2008

Technical Report No. 386
Approved for public release; distribution unlimited.

Fundamental Performance Analysis in Image Registration Problems - Cramér-Rao Bound and its Variations

Dan Ruan and Jeffrey A. Fessler
4125 EECS Bldg., 1301 Beal Ave.
University of Michigan
Ann Arbor, MI 48109-2122
FAX: 734-763-8041
email: druan@eecs.umich.edu
version: Mar. 2008

Technical Report # 386
Communications and Signal Processing Laboratory
Dept. of Electrical Engineering and Computer Science
The University of Michigan

Image registration, as a special form of signal warping, is an important task in image processing. Given the many current developments in algorithms and techniques in image registration, it is desirable to have fundamental performance criteria to compare the overall optimality of different estimators. This report presents an observation model for image registration that accounts for image noise more realistically than most formulations, and describes performance analysis based on Cramér-Rao Bound and its related variant MCRB.

I. Model - the Ideal v.s. Commonly Used

In a general setting, image registration methods aim to find the motion in an image sequence. Let z_i denote the ith observation (frame) of an underlying image. In reality, only sampled observations are available, with spatial sample spacing Δ. Therefore, it is natural to use a discrete spatial index to refer to the sampled location. Without loss of generality, we take $z_i[n] = z_i^c(n\Delta)$ where z^c notates the underlying continuous intensity map. Accounting for additive observation noise, we formulate the generative model as:

$$z_i[n] = f(n + \tau_i(n)) + \epsilon_i[n], \tag{1}$$

where it is standard to assume ϵ_i are normally distributed I.I.D noise. In principle, the task of registratering the observation sequence is to find the deformation sequence of continuous maps $\{\tau_i\}$ for all i. We adopt the parametric setting, and represent the underlying continuous image intensity as a linear combination of a finite number of basis functions b with coefficients $\boldsymbol{c} = \{c_k\}$, *i.e.*, $f(x) = \sum_{k=1}^K c_k b(x,k)$. For simplicity, we focus on pairwise registration which requires estimating one deformation field τ, and drop the subindex in τ_i. Furthermore, we assume the deformation field is properly (sufficiently) parameterized with α, so the estimation performance for deformation and image intensity may be characterized by that of the parameter set $(\boldsymbol{c}, \alpha)$. For simplicity, we formulate our problem in 1-D format, but the analysis generalizes to higher dimension (2-D and/or 3-D). The two observed images are modeled as:

$$\begin{aligned} z_1[n] &= \sum_{k=1}^K c_k b(n,k) + \epsilon_1[n], \\ z_2[n] &= \sum_{k=1}^K c_k b(n + \tau_\alpha(n), k) + \epsilon_2[n] \quad n = 1, 2, \ldots, N, \end{aligned} \tag{2}$$

where $\{b(\cdot, k)\}$ are common intensity bases, and τ parameterized by α captures the pointwise deformation. The components of additive noise ϵ_i are zero mean I.I.D Gaussian with variance σ^2.

The formulation in (2) captures the spatial sampling of the observation, the finite representation of the underlying "true" intensity $\{c_k\}_{k=1}^K$ and τ denotes the point-wise deformation.

For comparison purposes, in traditional registration setup, the estimator is often designed to find the transformation $\hat{\Gamma}$ such that [1]

$$\hat{\Gamma} = \arg\min_{\Gamma} D(z_2, z_1 \circ \Gamma), \tag{3}$$

where D is some difference measure, *e.g.*, sum-of-squared-difference (SSD) or mutual information (MI), and Γ indicates the transformation. In this setting, it is implicitly assumed that the z_1 (sometimes called "reference") is a noise-free version of the true intensity image f, and z_2 (also called "homologous") is a deformed image with statistical noise corresponding to the form of the difference metric. Clearly there is a lack of symmetry regarding the presence of noise in this formulation.

For simplicity, we use sum-of-squared-difference (SSD) as our default choice of the error metric D for (3) hereafter, to reveal the parallel structure with Gaussian noise assumptions, which is made in many practical cases.

II. Cramér-Rao Bound and its Asymptotic Behavior

We first reformulate (2) in a compact vector form as follows.

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A_0 \\ A_\tau \end{bmatrix} \boldsymbol{c} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} = \mathcal{A}_\tau \boldsymbol{c} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{z} \stackrel{\triangle}{=} [z_1(1), \ldots, z_1(N), z_2(1), \ldots, z_2(N)]^T \in \Re^{2N}$ and $\boldsymbol{c} \stackrel{\triangle}{=} [c_1, \ldots, c_K]^T \in \Re^K_{\geq 0}$ are column vectors by stacking the corresponding elements. The concatenated random noise vector $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma = \sigma^2 I_{2N})$. $A_0, A_\tau \in \boldsymbol{M}^{N\times K}$ have elements $A_0(i,j) = b(i,j)$ and $A_\tau(i,j) = b(i+\tau(i), j)$ for $i = 1, 2, \ldots, N, j = 1, 2, \ldots, K$. The overall system matrix $\mathcal{A} = [A_0^T, A_\tau^T]^T$. The Cramér-Rao Bound (CRB) is a fundamental lower bound on the variance of any unbiased estimator [1] and serves as a benchmark for estimator performance. When maximum-likelihood (ML) estimators are applied, which are known to be asymptotically unbiased, it is interesting use CRB to bound their variance. In [2], it is suggested that when inverting the Fisher information matrix (FIM) corresponding to the parameter of interest only is not straight-forward, it is feasible to use "complete-parameter" Fisher information matrices. Following a similar logic, we can write (4) in a more general form,

$$\begin{aligned} \boldsymbol{z} &= h(\tau_\alpha, \boldsymbol{c}) + \boldsymbol{\epsilon} \\ &= h(\theta) + \boldsymbol{\epsilon}, \end{aligned} \tag{5}$$

where $h(\tau_\alpha, \boldsymbol{c}) \stackrel{\triangle}{=} \mathcal{A}\boldsymbol{c}$ and $\theta = [\alpha, \boldsymbol{c}]$ denotes the "complete-parameter" vector. It follows immediately from the i.i.d Gaussian assumption of noise $\boldsymbol{\epsilon}$ that the ML estimator $\hat{\theta}_{ML}$ minimizes the L_2 distance between observation $\boldsymbol{z}$ and system response $h(\theta)$ as follows:

$$\hat{\theta}_{ML} = \arg\min_{\theta} \|\boldsymbol{z} - h(\theta)\|_2 .$$

Before we delve into the detailed computation, we clarify our goal and the structure of FIM here. We are ultimately interested in the performance of estimators for the deformation parameter α, and the image intensity parameter $\boldsymbol{c}$ is chosen to augment the data to simplify expression. With $\theta \in \Re^{N+K}$, the FIM corresponding to θ takes on the form:

$$F(\theta^*) = E_{\boldsymbol{z}|\theta=\theta*} \left\{ -\frac{\partial^2}{\partial \theta^2} \Lambda(\boldsymbol{z}|\theta)|_{\theta=\theta^*} \right\},$$

where Λ is the log-likelihood function $\Lambda(\boldsymbol{z}|\theta) \stackrel{\triangle}{=} \log f(\boldsymbol{z}|\theta)$.

Moreover, if we define $J_{x,y} = E\left\{ [\frac{\partial}{\partial x}\Lambda(\boldsymbol{z})]^T [\frac{\partial}{\partial y}\Lambda(\boldsymbol{z})] \right\}$, then the complete-data FIM can be decomposed into block form as:

$$F_\theta = \begin{bmatrix} J_{\alpha,\alpha} & J_{\alpha,\boldsymbol{c}} \\ J_{\boldsymbol{c},\alpha} & J_{\boldsymbol{c},\boldsymbol{c}} \end{bmatrix}. \tag{6}$$

[1] There is a slight abuse of notation here. The more precise formulation would be: $\hat{\Gamma} = \arg\min_\Gamma D(z_2, P(z_1^c \circ \Gamma))$, where z_1^c the underlying intensity map that agrees with z_1 on sampling grids, and P is the sampling function such that $P(z^c)(n) = z^c(n\Delta)$. Even so, the cost function is still incomplete, as only z_1 is observed and the interpolator $I : z_1 \rightarrow z_1^c$ needs to be specified. The de facto objective function is thus $D(z_2, P(I(z_1) \circ \Gamma))$.

The sub-block $J_{\tau,\tau}$ is the FIM with respect to the quantity of interest - the deformation parameters. As CRB is the inverse of the FIM, we can invoke the formula for partitioned-matrix inverse [3] to obtain:

$$\begin{aligned} \mathrm{CRB}(\alpha) &= [J_{\alpha,\alpha} - J_{\alpha,\boldsymbol{c}} J_{\boldsymbol{c},\boldsymbol{c}}^{-1} J_{\boldsymbol{c},\alpha}]^{-1} \\ \mathrm{CRB}(\boldsymbol{c}) &= [J_{\boldsymbol{c},\boldsymbol{c}} - J_{\boldsymbol{c},\alpha} J_{\alpha,\alpha}^{-1} J_{\alpha,\boldsymbol{c}}]^{-1}. \end{aligned} \tag{7}$$

This form can be further simplified using its symmetry - a fact that we will utilize later in our computation.

The likelihood function with respect to θ is :

$$f(\boldsymbol{z};\theta) = \frac{1}{(2\pi)^{2N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{e}^T\Sigma^{-1}\boldsymbol{e}\right),$$

where $\boldsymbol{e} = \boldsymbol{z} - h(\theta) = \boldsymbol{z} - \mathcal{A}(\tau_\alpha)\boldsymbol{c}$.

The log-likelihood turns out to be:

$$\begin{aligned} \Lambda &= \log f(\boldsymbol{z};\theta) \\ &= -N\log(2\pi) - 2N\log\sigma - \frac{1}{2\sigma^2}\left\|\boldsymbol{z} - \mathcal{A}(\tau_\alpha)\boldsymbol{c}\right\|^2. \end{aligned} \tag{8}$$

Now we compute each term of the FIM.

$$\begin{aligned} \nabla_\tau\Lambda &= -\frac{1}{2\sigma^2}\nabla_\tau\left\|\boldsymbol{z} - \mathcal{A}(\tau_\alpha)\boldsymbol{c}\right\|^2 \\ &= \frac{1}{\sigma^2}(\boldsymbol{z} - \mathcal{A}(\tau_\alpha)\boldsymbol{c})^T\nabla_\tau(\mathcal{A}(\tau_\alpha)\boldsymbol{c}) \end{aligned} \tag{9}$$

Notice that

$$\begin{aligned} \frac{\partial}{\partial\tau(l)}\{A_\tau[n,:]\boldsymbol{c}\} &= \frac{\partial}{\partial\tau(l)}\sum_{k=1}^{K} c(k)A_\tau(n,k) \\ &= \frac{\partial}{\partial\tau(l)}\sum_{k=1}^{K} c(k)b(n+\tau(n),k) \\ &= \begin{cases} \sum_{k=1}^{K} c(k)\dot{b}(n+\tau(n),k), & l = n; \\ 0 & \text{else}, \end{cases} \end{aligned} \tag{10}$$

where $\dot{b}(\cdot,\cdot)$ denotes the derivative of $b(\cdot,\cdot)$ with respect to the first variable.

Plugging (10) into (9), we obtain

$$\frac{\partial}{\partial\tau(l)}\Lambda = \frac{1}{\sigma^2}(z_2(l) - A_\tau[l,:]\boldsymbol{c})\sum_{k=1}^{K}\dot{b}(l+\tau(l),k)c(k). \tag{11}$$

Therefore, the gradient of Λ with respect to τ is:

$$\frac{\partial}{\partial\tau}\Lambda = \frac{1}{\sigma^2}[(z_2 - A_\tau\boldsymbol{c})\odot(D\boldsymbol{c})]^T = \frac{1}{\sigma^2}[\mathsf{diag}\{D\boldsymbol{c}\}(z_2 - A_\tau(\boldsymbol{c}))]^T, \tag{12}$$

where D is the matrix whose elements are $D(i,j) = \dot{b}(i+\tau(i),j), 1 \le i \le N, 1 \le j \le K$, and "$\odot$" denotes the Schur/Hadamard product.

By chain rule, the gradient of Λ with respect to α is given by:

$$\begin{aligned} \nabla_\alpha\Lambda &= \frac{\partial}{\partial\tau}\Lambda\frac{\partial}{\partial\alpha}\tau \\ &= \frac{1}{\sigma^2}[(z_2 - A_\tau\boldsymbol{c})\odot(D\boldsymbol{c})]^T[\frac{\partial}{\partial\alpha}\tau], \end{aligned} \tag{13}$$

where $\frac{\partial}{\partial\alpha}\tau \in \boldsymbol{M}^{N\times L}$ is the derivative matrix with element $[\frac{\partial}{\partial\alpha}\tau](i,j) = \frac{\partial}{\partial\alpha(j)}\tau(i)$, and L corresponds to the length of the deformation parameter α.

Now we compute the FIM $J_{\alpha,\alpha}$ with

$$\begin{aligned} E\left\{\frac{\partial^2}{\partial\alpha^2}\Lambda\right\} &= -E\left\{[\frac{\partial}{\partial\alpha}\Lambda]^T[\frac{\partial}{\partial\alpha}\Lambda]\right\} \\ &= -\frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T E\left\{[\frac{\partial}{\partial\tau}\Lambda]^T[\frac{\partial}{\partial\tau}\Lambda]\right\}\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} \qquad (14) \\ &= -\frac{1}{\sigma^2}\frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T E\left\{\mathsf{diag}\{D\boldsymbol{c}\}(z_2 - A_\tau\boldsymbol{c})[\frac{1}{\sigma^2}(z_2 - A_\tau\boldsymbol{c})^T\,\mathsf{diag}\{D\boldsymbol{c}\}]^T\right\}\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} \\ &= -\frac{1}{\sigma^2}\frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T\,\mathsf{diag}^2\{D\boldsymbol{c}\}\,\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} \qquad (15) \end{aligned}$$

To calculate $J_{\boldsymbol{c},\tau}$ and $J_{\boldsymbol{c},\boldsymbol{c}}$, we take the derivative of Λ with respect to $\boldsymbol{c}$:

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{c}}\Lambda &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial\boldsymbol{c}}||\boldsymbol{z} - \mathcal{A}\boldsymbol{c}||^2 \\ &= \frac{1}{\sigma^2}(\boldsymbol{z} - \mathcal{A}\boldsymbol{c})^T\mathcal{A}. \qquad (16) \end{aligned}$$

It is now straight forward to compute the entries for the complete FIM:

$$E\left\{\frac{\partial^2}{\partial\boldsymbol{c}\partial\boldsymbol{c}}\Lambda\right\} = -\frac{1}{\sigma^2}\mathcal{A}^T\mathcal{A} \qquad (17)$$

$$\begin{aligned} E\left\{\frac{\partial^2}{\partial\tau(l)\partial\boldsymbol{c}(m)}\Lambda\right\} &= \frac{1}{\sigma^2}E\left\{-A_\tau[l,m]D[l,:]\boldsymbol{c} + \epsilon_2(l)D[l,m]\right\} \\ &= -\frac{1}{\sigma^2}A_\tau[l,m]D[l,:]\boldsymbol{c}. \qquad (18) \end{aligned}$$

The matrix $J_{\alpha,\boldsymbol{c}}$ can be represented in compact form as:

$$E\left\{\frac{\partial^2}{\partial\alpha\partial\boldsymbol{c}}\Lambda\right\} = -\frac{1}{\sigma^2}\frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T\,\mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau. \qquad (19)$$

With symmetry, the complete FIM is obtained:

$$F_\theta = \frac{1}{\sigma^2}\begin{bmatrix} \frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T\,\mathsf{diag}^2\{D\boldsymbol{c}\}\,\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} & \frac{\mathrm{d}\tau}{\mathrm{d}\alpha}^T\,\mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau \\ A_\tau^T\,\mathsf{diag}\{D\boldsymbol{c}\}\,\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} & \mathcal{A}^T\mathcal{A} \end{bmatrix}. \qquad (20)$$

As a special case, when τ is parameterized with rect functions, *i.e.*, $\tau(n) = \alpha[n]$, we have $\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} = I$. The FIM for $(\tau, \boldsymbol{c})$ is then given by:

$$F_{(\tau,\mathbf{c})} = \frac{1}{\sigma^2}\begin{bmatrix} \mathsf{diag}^2\{D\boldsymbol{c}\} & \mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau \\ A_\tau^T\,\mathsf{diag}\{D\boldsymbol{c}\} & \mathcal{A}^T\mathcal{A} \end{bmatrix}. \qquad (21)$$

At this point, we make the following observations:

1. With the commonly used model (3), it is assumed that the observed reference image z_1 corresponds to the ground truth $\boldsymbol{c}$. In other words, most existing methods solve for the ML estimator τ with the generative model:

$$z_2 = \sum_{k=1}^{K} c_k b(n + \tau(n), k) + \epsilon_2(n), \qquad (22)$$

by plugging in the c_k's that best fits z_1. It is easy to derive the CRB for the log-likelihood function $\Lambda^{\text{com}}(z_2;\tau) = -N/2\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}||z_2 - A_\tau c||^2$. The FIM matrix $F_\tau^{\text{com}} = J_{\tau,\tau}$ as we derived in (15). Therefore, $\text{CRB}^{\text{com}}(\tau) = J_{\tau,\tau}^{-1}$. Notice that[2] as $J_{\tau,\boldsymbol{c}} J_{\boldsymbol{c},\boldsymbol{c}}^{-1} J_{\boldsymbol{c},\tau} \geq 0$, $\text{CRB}^{\text{com}}(\tau) \leq \text{CRB}(\tau)$ as extra information (known $\{c_k\}$) is assumed in the case of (22). In other words, the plug-in operation provides a "looser" bound for the variance than the "true" CRB corresponding to model (2).

2. For asymptotically large SNR, *i.e.*, $\sigma^2 \to 0$, we do expect a decent estimate of $\boldsymbol{c}$ directly from the reference image only, assuming no model mismatch in the generative basis. In this case, the plug-in estimator as used in the traditional model, even though not a true ML estimator, is expected to perform similarly to the real ML estimator. Indeed, [4] shows that the "fake" bound approximates the true CRB [3].
3. The above points may be interpreted better with a slight modification of the model in (2). Instead of i.i.d noise, we assume that noise level in the two images are not symmetric, more specifically, we assume $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2 I_N)$ and $\epsilon_2 \sim \mathcal{N}(0, \sigma_2^2 I_N)$.
   The log-likelihood is given by:
$$\Lambda = -\frac{1}{2\sigma_1^2}\|z_1 - A_0\boldsymbol{c}\|_2 - \frac{1}{2\sigma_2^2}\|z_2 - A_\tau\boldsymbol{c}\|_2 + \text{some constant}. \tag{23}$$
   The partial derivatives of the log-likelihood with respect to τ (thus α) is not affected by target image model, and the second-order derivative the log-likelihood with respect to $\boldsymbol{c}$ is given by:
$$E\left\{\frac{\partial^2}{\partial\boldsymbol{c}\partial\boldsymbol{c}}\Lambda\right\} = -\frac{1}{\sigma_1^2}A_0^T A_0 - \frac{1}{\sigma_2^2}A_\tau^T A_\tau.$$
   We thus obtain the complete FIM with respect to $(\tau, \boldsymbol{c})$ as:
$$F_{(\tau,\boldsymbol{c})} = \begin{bmatrix} \frac{1}{\sigma_2^2}\,\mathsf{diag}^2\{D\boldsymbol{c}\} & \frac{1}{\sigma_2^2}\,\mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau \\ \frac{1}{\sigma_2^2}A_\tau^T\,\mathsf{diag}\{D\boldsymbol{c}\} & \frac{1}{\sigma_1^2}A_0^T A_0 + \frac{1}{\sigma_2^2}A_\tau^T A_\tau. \end{bmatrix}. \tag{24}$$
   When $\sigma_1 \to 0$, corresponding to high SNR in the template image, then $J_{\boldsymbol{c},\boldsymbol{c}} \to \infty$ and
$$\text{CRB}(\tau) = [J_{\tau,\tau} - J_{\tau,\boldsymbol{c}} J_{\boldsymbol{c},\boldsymbol{c}}^{-1} J_{\boldsymbol{c},\tau}]^{-1} \to J_{\tau,\tau}^{-1},$$
   which reduces to the CRB^{com}.
4. To compute $\text{CRB}(\tau)$ exactly could be challenging, as $\mathcal{A}^T\mathcal{A}$ may not be easy to invert for arbitrary τ. Notice that the sub-matrix A_0 of $\mathcal{A}$ has nice shift-invariant structure, yet A_τ depends on the deformation. In special cases, such as when the whole image (signal) experience uniform transformation $\tau(i) = const$ for $i = 1, 2, \ldots, N$, then $J_{\boldsymbol{c},\boldsymbol{c}}$ is block-shift-invariant, and efficient inversion is possible.
5. As a special case, we consider when the whole image experiences uniform transformation, where a natural parameterization is to use α to describe the global transformation, *i.e.*, $\tau_\alpha(i) = \alpha$ for $\forall i$.
   Under the uniform transformation assumption, we have
$$\frac{\mathrm{d}\tau}{\mathrm{d}\alpha} = \mathbf{1},$$
   where $\mathbf{1}$ indicates a column vector (of length N in our case) with all unity elements.
   Substituting this relation into (15), (19) respectively and we obtain:
$$\begin{aligned} F_\theta &= \frac{1}{\sigma^2}\begin{bmatrix} \mathbf{1}^T\,\mathsf{diag}^2\{D\boldsymbol{c}\}\,\mathbf{1} & \mathbf{1}^T\,\mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau \\ A_\tau^T\,\mathsf{diag}\{D\boldsymbol{c}\}\,\mathbf{1} & \mathcal{A}^T\mathcal{A} \end{bmatrix} \\ &= \frac{1}{\sigma^2}\begin{bmatrix} \boldsymbol{c}^T D D^T \boldsymbol{c} & [D\boldsymbol{c}]^T A_\tau \\ A_\tau^T [D\boldsymbol{c}] & \mathcal{A}^T\mathcal{A} \end{bmatrix}. \end{aligned} \tag{25}$$

[2] In most cases, we assume it is nonsingular, so it is in fact positive definite.

[3] In particular, the parameters to be estimated τ is not coupled with the nuisance parameters $\boldsymbol{c}$ in our case, and the asymptotic behavior of the bound can be shown with ease.

III. Relating to MCRB

The modified Cramér-Rao Bound (MCRB) was first introduced [5] to resolve the synchronization issues in decoding systems. Rather than seeking the variance around the estimator for the "true" augmented data ("complete data") which includes both the quantity of interest and the nuisance parameters $\boldsymbol{c}$, MCRB choose to look on the other parameters as "unwanted". Instead of using the true CRB, the MCRB may be regarded as an approximation via "marginalizing" over the nuisance parameters. In fact, MCRB is always lower than CRB, thus a looser bound. In some cases, MCRB approaches the true CRB [4].

The central idea is the following. Instead of computing the true FIM

$$F = E_z \left\{ [\frac{\partial}{\partial \tau} \log f(z;\tau)]^2 \right\},$$

it uses

$$E_{z,\boldsymbol{c}} \left\{ [\frac{\partial}{\partial \tau} \log f(z;\tau,\boldsymbol{c})] \right\}. \tag{26}$$

The rationale for MCRB is the following:

$$\begin{aligned} E_{z,\boldsymbol{c}} \left\{ [\hat{\tau(z)} - \tau]^2 \right\} &= E_{\boldsymbol{c}} \left\{ E_{z|\boldsymbol{c}}[(\hat{\tau(z)} - \tau)^2] \right\} \\ &\geq E_{\boldsymbol{c}} \left\{ \frac{1}{E_{z|\boldsymbol{c}}[(\frac{\partial}{\partial\tau}\log f(z;\tau,\boldsymbol{c}))^2]} \right\} \\ &\geq \frac{1}{E_{\boldsymbol{c}}\left\{E_{z|\boldsymbol{c}}[(\frac{\partial}{\partial\tau}\log f(z;\tau,\boldsymbol{c}))^2]\right\}} \\ &= \frac{1}{E_{z,\boldsymbol{c}}\left\{[\frac{\partial}{\partial\tau}\log f(z;\tau,\boldsymbol{c})]^2\right\}}. \end{aligned} \tag{27}$$

The first inequality comes from the application of CRB to the estimator $\hat{\tau}(z)$ for a fixed $\boldsymbol{c}$ and second is Jensen's inequality.

IV. An Alternating Minimization Algorithm

For registration purposes, we want to minimize the negative log-likelihood in (23). We assume the underlying image intensity f (and thus $\boldsymbol{c}$ are fixed unknown) and adopt the frequentist point of view. It is natural to ask for the solution of the augmented problem:

$$(\hat{\tau}, \hat{\boldsymbol{c}}) = \arg\min_{\tau,\boldsymbol{c}} -\Lambda.$$

Here, we describe an alternating minimization algorithm to solve this problem.

Algorithm 1 Alternating minimizing the nagetive log-likelihood in (23).

1: Initialize $\hat{\boldsymbol{c}}$
2: **repeat**
3: For given $\boldsymbol{c} = \hat{\boldsymbol{c}}$, minimize $\|z_2 - A_\tau \boldsymbol{c}\|_2$ over τ. This step coincides with conventional registration methods by assuming $\boldsymbol{c}$ known. Obtain $\hat{\tau}$.
4: For given $\tau = \hat{\tau}$, minimize $\frac{1}{2\sigma_1^2}\|z_1 - A_0\boldsymbol{c}\|_2^2 + \frac{1}{2\sigma_2^2}\|z_2 - A_\tau\boldsymbol{c}\|_2^2$. This is a typical quadratic minimization problem, and the solution is given by:

$$\hat{\boldsymbol{c}} = [\frac{1}{\sigma_1^2}A_0^T A_0 + \frac{1}{\sigma_2^2}A_\tau^T A_\tau]^\dagger (\frac{1}{\sigma_1^2}A_0^T z_1 + \frac{1}{\sigma_2^2}A_\tau^T z_2), \tag{28}$$

where $(\cdot)^\dagger$ indicates the pseudo-inverse operator for the Gram matrix.
5: **until** Some convergence condition is satisfied.

We make the following remarks:

- As $\sigma_1 \to 0$, the contribution of A_0 and z_1 dominates (28), and the solution reduces to

$$\hat{\boldsymbol{c}} = \left[A_0^T A_0\right]^\dagger A_0^T z_1, \tag{29}$$

which corresponds to the conventional method where z_1 is considered to be a highly reliable "template" and the image intensity is solely obtained by fitting z_1.
- More generally, alternating descent may be used instead of requiring the achieving minimizer at each iteration. This could be particularly beneficial for the step in updating τ conditioned on $\hat{\boldsymbol{c}}$, as the quadratic form in the other step makes the minimization over $\boldsymbol{c}$ trivial. Relaxing conditional maximization to increase in log-likelihood may has potential computational advantage as well as better behavior to local maxima.
- As $\sigma_1 \to 0$, the alternating descent algorithm reduces to exactly any conventional descent algorithm in solving (3) with l_2 difference metric. In the asymptotic case, the conditional minimization of $\boldsymbol{c}$ given by (29) is independent of τ and the whole alternating descent algorithm reduces to using the plug-in estimator (29) and descend $-\Lambda$ with respect to τ.

V. Comparison with Conventional Methods: CRB v.s. M-estimate

As we have commented briefly in the previous sections, the conventional method estimate the intensity f from the source image z_1 only. With l_2 difference metric, we can write the solution to the conventional method as:

$$\begin{aligned}\hat{\boldsymbol{c}} &= \arg\min_{\boldsymbol{c}} \|z_1 - A_0\boldsymbol{c}\|_2^2\,;\\ \hat{\tau} &= \arg\min_{\tau} \|z_2 - A_\tau\hat{\boldsymbol{c}}\|_2^2\,,\end{aligned} \tag{30}$$

where z_1, z_2 are discrete observations for the source and target image in vector form, A_0 and A_τ are defined as in (4).

The first equation in (30) can be solved in closed form given its quadratic form:

$$\hat{\boldsymbol{c}} = A_0^\dagger z_1,$$

and we can rewrite (30) as:

$$\hat{\tau} = \arg\min_{\tau} \left\| z_2 - A_\tau A_0^\dagger z_1 \right\|_2^2. \tag{31}$$

We can also stack the expression as before, and define $\mathcal{A} \triangleq [-A_\tau A_0^\dagger \quad I]$ and write the objective as:

$$\hat{\tau} = \arg\min_{\tau} \Phi(\tau, z) = \|\mathcal{A}(\tau)z\|_2^2\,. \tag{32}$$

In the following derivations, we will use the most convenenient and use the above equivalent expressions interchangably.

Our goal is to derive the covariance of the minimizer defined above and we use similar philosophy as in [6]. By implicit function theorem, the partial derivative of Φ with respect to τ are uniformly zero:

$$\frac{\partial}{\partial \tau(i)} \Phi(\tau, z)|_{\tau=\hat{\tau}} = 0, \quad \forall \text{ spatial location } i, \tag{33}$$

for any given data z.

Differentiating (33) again with respect to z and applying the chain rule yields:

$$\nabla^{20}\Phi(\hat{\tau}(z), z))\nabla_z\hat{\tau}(z) + \nabla^{11}\Phi(\hat{\tau}(z), z) = 0. \tag{34}$$

Where, the components of $\nabla^{20}\Phi(\hat{\tau}(z), z)$ are $\frac{\partial^2}{\partial\tau(i)\partial\tau(j)}\Phi(\hat{\tau}(z), z)$, and the elements of ∇^{11} are $\frac{\partial^2}{\partial\tau(i)\partial z(j)}\Phi(\hat{\tau}(z), z)$. We consider the case when $\nabla^{20}\Phi(\hat{\tau}(z), z)$ is invertible, or more precisely positive definite. Note that $\Phi(\hat{\tau}(z), z)$ is *locally* strictly convex. This assumption is true if the following regularity condition is satisfied: *There* $\exists$ *a compact neighborhood* $\boldsymbol{N}(\hat{\tau})$ *such that* $\Phi(\tau, z) > \Phi(\hat{\tau}(z), z)$ *for all* $\tau \neq \hat{\tau}$. Then we obtain:

$$\nabla_Y\hat{\tau}(z) = [-\nabla^{20}\Phi(\hat{\tau}, z)]^{-1}\nabla^{11}\Phi(\tau, z).$$

and so the covariance matrix for $\hat{\tau}$ would be $\mathsf{Cov}\{z\}$ transformed by local linearizion [7], *i.e.*,

$$\mathsf{Cov}\{\hat{\tau}\} \approx \nabla_z \hat{\tau}(z)\, \mathsf{Cov}\{z\}[\nabla_z \hat{\tau}(z)]'.$$

By substitution, we obtain

$$\mathsf{Cov}\{\hat{\tau}\} \approx [\nabla^{20}\Phi(\hat{\tau}, z)]^{-1}\nabla^{11}\Phi(\hat{\tau}, z)\, \mathsf{Cov}\{z\}[\nabla^{11}\Phi(\hat{\tau}, z)]'[\nabla^{20}\Phi(\hat{\tau}, z)]^{-1}. \tag{35}$$

By assumption, we assume the covariance of z is of the form:

$$\mathsf{Cov}\{z\} = \begin{bmatrix} \sigma_1^2 I_N & 0 \\ 0 & \sigma_2^2 I_N \end{bmatrix}. \tag{36}$$

It remains to derive the expressions for $\nabla^{20}\Phi(\hat{\tau}, z)$ and $\nabla^{11}\Phi(\hat{\tau}, z)$.

We first adopt the objective function form in (31) to take derivative with respect to $\tau(l)$.

$$\frac{\partial}{\partial \tau(l)}\Phi(\tau, z) = \sum_{n=1}^{N}(A_\tau[n, :]A_0^\dagger z_1 - z_2(n))\frac{\partial^2}{\partial \tau(l)^2}\{A_\tau(n)A_0^\dagger z_1\}. \tag{37}$$

Similar to (10),

$$\begin{aligned} \frac{\partial}{\partial \tau(l)}\left\{A_\tau[n, :]A_0^\dagger z_1\right\} &= \frac{\partial}{\partial \tau(l)}\sum_{k=1}^{K}(A_0^\dagger z_1)(k)A_\tau(n, k) \\ &= \frac{\partial}{\partial \tau(l)}\sum_{k=1}^{K}(A_0^\dagger z_1)(k)b(n + \tau(n), k) \\ &= \begin{cases} \sum_{k=1}^{K}(A_0^\dagger z_1)(k)\dot{b}(n + \tau(n), k), & l = n; \\ 0 & \text{else,} \end{cases} \end{aligned} \tag{38}$$

where $\dot{b}(\cdot, \cdot)$ denote the derivative of $b(\cdot, \cdot)$ with respect to the first variable.

Plugging (38) into the expression in (37) yields:

$$\frac{\partial}{\partial \tau(l)}\Phi(\tau, z) = (A_\tau[l, :]A_0^\dagger z_1 - z_2(l))\sum_{k=1}^{K}(A_0^\dagger z_1)(k)\dot{b}(l + \tau(l), k). \tag{39}$$

To obtain $\nabla^{20}\Phi$, we take derivative with respect to $\tau(n)$. Noticing that the dependence of $\frac{\partial}{\partial \tau(l)}\Phi$ on τ is only via $\tau(l)$, we obtain:

$$\frac{\partial^2}{\partial \tau(l)\partial \tau(n)}\Phi(\tau, z) = \begin{cases} \left\{\sum_{k=1}^{K}(A_0^\dagger z_1)(k)\dot{b}(l + \tau(l), k)\right\}^2 + \cdots & \\ \quad + (A_\tau[l, :]A_0^\dagger z_1 - z_2(l))\sum_{k=1}^{K}(A_0^\dagger z_1)(k)\ddot{b}(l + \tau(l), k), & l = n; \\ 0 & \text{else.} \end{cases}$$

where $\ddot{b}(\cdot, \cdot)$ denotes the second-order partial derivative with respect to the first argument in $b(\cdot, \cdot)$.

To compute $\nabla^{11}\Phi(\hat{\tau}, z)$, we need to take derivative of (39) with respect to each element of z. We perform this by distinguishing the among elements in z_1 and z_2 respectively.

Noting that $\frac{\partial}{\partial z_1(n)}[A_0^\dagger z_1](k) = A_0^\dagger[k, n]$, we obtain:

$$\begin{aligned} \frac{\partial^2}{\partial \tau(l)\partial z_1(n)}\Phi(\tau, z) &= A_\tau[l, :]A_0^\dagger[:, n]\sum_{k=1}^{K}(A_0^\dagger z_1)(k)\dot{b}(l + \tau(l), k) + \cdots \\ &\quad + (A_\tau[l, :]A_0^\dagger z_1 - z_2(l))\sum_{k=1}^{K}A_0^\dagger[k, n]\dot{b}(l + \tau(l), k). \end{aligned} \tag{40}$$

$$\frac{\partial^2}{\partial\tau(l)\partial z_2(n)}\Phi(\tau,z)=\begin{cases} -\sum_{k=1}^{K}(A_0^\dagger z_1)(k)\dot{b}(l+\tau(l),k), & l=n;\\ 0 & \text{else.}\end{cases}$$

We assume that at the point of evaluation $(\check{\tau},\bar{z})$, the samples of the warped z_1^c approximates the observation z_2, more specifically:

$$A_\tau A_0^\dagger \bar{z}_1 \approx \bar{z}_2.$$

This is a reasonable assumption for most registration results. For simplicity, we denote $\bar{\boldsymbol{c}} \triangleq A_0^\dagger \bar{z}_1$, $\check{D}(i,j) \triangleq \dot{b}(i+\check{\tau}(i),j)$, and the warping map $\mathcal{W} \triangleq A_{\check{\tau}}A_0^\dagger$, then we can rewrite in matrix form:

$$\begin{aligned}\nabla^{20}\Phi(\check{\tau},\bar{z}) &= \mathsf{diag}^2\{\check{D}\bar{\boldsymbol{c}}\}\\ \nabla^{11}\Phi(\check{\tau},\bar{z}) &= \left[\ \mathsf{diag}\{\check{D}\bar{\boldsymbol{c}}\}\,\mathcal{W} \quad -\mathsf{diag}\{\check{D}\bar{\boldsymbol{c}}\}\ \right].\end{aligned} \tag{41}$$

Plugging (41) and (36) into the expression for $\mathsf{Cov}\{\hat{\tau}\}$ in (35), we obtain:

$$\mathsf{Cov}\{\hat{\tau}\}\,|_{\hat{\tau}=\check{\tau}} \approx \mathsf{diag}\{\check{D}\bar{\boldsymbol{c}}\}^{-1}[\sigma_1^2\mathcal{W}\mathcal{W}^T+\sigma_2^2 I]\,\mathsf{diag}\{\check{D}\bar{\boldsymbol{c}}\}^{-1}. \tag{42}$$

Remark: as $\sigma_1^2\to 0$, z_1 approaches the noise-free observation of a template image f, and the conventional method should yield the same estimate as the more realistic model. In fact,

$$\mathsf{Cov}_{\sigma_1\to 0}\{\hat{\tau}\}=\sigma_2^2\,\mathsf{diag}^2\{\check{D}\bar{\boldsymbol{c}}\},$$

which agrees with our previous analyis for $CRB(\tau)\to J_{\tau,\tau}^{-1}$ for as z_1 asymptotically becomes noise-free in (24).

It makes sense to compare the covariance prediction for the M-estimate of the conventional method and the Cramér-Rao Bound obtained from the more realistic model from (2). For simplicity, we assume that A_0 to be invertible so that $A_0^{-1}=A_0^\dagger$ and consequently the warping map $\mathcal{W}=A_\tau A_0^{-1}$ is invertible.

To study $CRB(\tau)$, we plug $J_{\tau,\boldsymbol{c}}$, $J_{\boldsymbol{c},\boldsymbol{c}}$ from (24) and obtain:

$$\begin{aligned} CRB(\tau) &= [J_{\tau,\tau}-J_{\tau,\boldsymbol{c}}J_{\boldsymbol{c},\boldsymbol{c}}^{-1}J_{\boldsymbol{c},\tau}]^{-1}\\ &= \Big\{\frac{1}{\sigma_2^2}\mathsf{diag}^2\{D\boldsymbol{c}\}-\frac{1}{\sigma_2^2}\mathsf{diag}\{D\boldsymbol{c}\}\,A_\tau[\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_\tau^TA_\tau]^{-1}\frac{1}{\sigma_2^2}A_\tau^T\,\mathsf{diag}\{D\boldsymbol{c}\}\Big\}^{-1}\\ &= \sigma_2^2\,\mathsf{diag}\{D\boldsymbol{c}\}^{-1}\Big\{I-\frac{1}{\sigma_2^2}A_\tau[\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_\tau^TA_\tau]^{-1}A_\tau^T\Big\}^{-1}\mathsf{diag}\{D\boldsymbol{c}\}^{-1}. \end{aligned} \tag{43}$$

With the $\mathcal{W}=A_\tau A_0^{-1}$, $A_\tau=\mathcal{W}A_0$ and we can write:

$$\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_\tau^TA_\tau=\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_0^T\mathcal{W}^T\mathcal{W}A_0.$$

The middle part of (43) can be rewritten as:

$$\begin{aligned} &\Big\{I-\frac{1}{\sigma_2^2}A_\tau[\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_\tau^TA_\tau]^{-1}A_\tau^T\Big\}^{-1}\\ &\quad= \Big\{I-\sigma_1^2A_\tau[\sigma_2^2A_0^TA_0+\sigma_1^2A_0^T\mathcal{W}^T\mathcal{W}A_0]^{-1}A_\tau^T\Big\}^{-1}\\ &\quad= \Big\{I-\sigma_1^2A_\tau A_0^{-1}[\sigma_2^2I+\sigma_1^2\mathcal{W}^T\mathcal{W}]^{-1}A_0^{-T}A_\tau^T\Big\}^{-1}\\ &\quad= \Big\{I-\sigma_1^2\mathcal{W}[\sigma_2^2I+\sigma_1^2\mathcal{W}^T\mathcal{W}]^{-1}\mathcal{W}^T\Big\}^{-1}. \end{aligned} \tag{44}$$

By Woodbury-Sherman-Morissey identity:

$$[\sigma_2^2I+(\sigma_1\mathcal{W})\sigma_1\mathcal{W}^T]^{-1}=\frac{1}{\sigma_2^2}I-\frac{1}{\sigma_2^4}\sigma_1^2\mathcal{W}[I+\frac{\sigma_1^2}{\sigma_2^2}\mathcal{W}\mathcal{W}^T]^{-1}\mathcal{W}^T,$$

thus $\sigma_2\big\{I-\frac{1}{\sigma_2^2}A_\tau[\frac{1}{\sigma_1^2}A_0^TA_0+\frac{1}{\sigma_2^2}A_\tau^TA_\tau]^{-1}A_\tau^T\big\}^{-1}=\sigma_2^2I+\sigma_1^2\mathcal{W}\mathcal{W}^T$.

Substituting into (43), we obtain:

$$\mathrm{CRB}(\tau)=\mathsf{diag}\{D\boldsymbol{c}\}(\sigma_2^2I+\sigma_1^2\mathcal{W}\mathcal{W}^T)\,\mathsf{diag}\{D\boldsymbol{c}\}. \tag{45}$$

This result coincides with the covariance estimate for the M-estimate evaluated at $(\check{D},\bar{c})$ in (42).

VI. A BABY EXAMPLE

This section uses a simple example to illustrate the results from previous sections and also motivate discussion about performance comparison. In particular, it is expected that the proposed model in (4) has advantage over the traditional model in (3) as the estimation for $\boldsymbol{c}$ which parameterizes the underlying image intensity should be more reliable as it combines the information from both the source and the target observations. Consider the model

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} I \\ \alpha I \end{bmatrix} \boldsymbol{c} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \tag{46}$$

where we assume both z_1 and z_2 are vectors of the same size as the underlying (unknown) $\boldsymbol{c}$. The scaling parameter α which relates z_1 and z_2 in the noise-free case is the quantity of interest. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_1^2 I)$ and $\boldsymbol{\epsilon}_2 \sim \mathcal{N}(0, \sigma_2^2 I)$ are independent Gaussian additive noise.

.1 M-estimator for the Conventional Method

In the conventionally method, the parameter $\boldsymbol{c}$ is estimated solely from observation z_1:

$$\hat{\boldsymbol{c}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{c}} \|z_1 - \boldsymbol{c}\|_2^2 = z_1. \tag{47}$$

Since $z_1 \sim \mathcal{N}(\boldsymbol{c}, \sigma_1^2 I)$, $\hat{\boldsymbol{c}}$ is an unbiased estimator for $\boldsymbol{c}$ with covariance $\sigma_1^2 I$.

The objective function that $\hat{\alpha}$ minimizes is

$$\Phi(\alpha, \boldsymbol{z}) \stackrel{\triangle}{=} \|[\alpha I \quad -I]\boldsymbol{z}\|_2^2 = \|z_2 - \alpha z_1\|_2^2 \,. \tag{48}$$

$$\begin{aligned} \hat{\alpha}(\boldsymbol{z}) &= \arg\min_{\alpha} \Phi(\alpha, \boldsymbol{z}) \\ &= \arg\min_{\alpha} \|z_2 - \alpha \hat{\boldsymbol{c}}\|_2^2 \\ &= \arg\min_{\alpha} \|z_2 - \alpha z_1\|_2^2 \\ &= \frac{{z_1}^T z_2}{\|z_1\|_2^2}. \end{aligned} \tag{49}$$

Hereafter, we discuss two approaches in approximating the mean and variance of $\hat{\alpha}$: a direct method based on the explicit solution in (49); and an indirect approach that relies on implicit function theorem and M-estimate. The explicit method is straightforward, requires less manipulation, and should be reasonably accurate. On the other hand, explicit solutions are not available in general (as we will see for the ML estimator), so the implicit method is more universally applicable. In this study, the direct method serves as a good baseline reference for approximation performance, and the derivation based on indirect approach is of didactic value.

.2 Direct Approximation of Mean and Variance for the M-estimate

First, we directly approximate the mean and covariance of $\hat{\alpha}$ based on the explicit solution in (49).

The expected value of $\hat{\alpha}$ from (49) is given by:

$$E[\hat{\alpha}] = E\left\{ \frac{(\bar{\boldsymbol{c}} + \epsilon_1)^T (\bar{\alpha}\bar{\boldsymbol{c}} + \epsilon_2)}{(\bar{\boldsymbol{c}} + \epsilon_1)^T (\bar{\boldsymbol{c}} + \epsilon_1)} \right\},$$

where $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2 I)$ and $\epsilon_2 \sim \mathcal{N}(0, \sigma_2^2 I)$. We compute the above expression using conditional expectation:

$$\begin{aligned} E[\hat{\alpha}] &= E_{\epsilon_1}\left\{ E_{\epsilon_2}[\hat{\alpha}] | \epsilon_1 \right\} \\ &= \bar{\alpha} E_{\epsilon_1}\left\{ \frac{(\bar{\boldsymbol{c}} + \epsilon_1)^T \bar{\boldsymbol{c}}}{(\bar{\boldsymbol{c}} + \epsilon_1)^T (\bar{\boldsymbol{c}} + \epsilon_1)} \right\}. \end{aligned} \tag{50}$$

where the second line follows from the independence between ϵ_1 and ϵ_2.

Let c_i denote the ith element of $\bar{\boldsymbol{c}}$ and e_i denote the ith element of ϵ_1. Then c_i are constants and e_i are scalar i.i.d Gaussian variables $e_i \sim \mathcal{N}(0, \sigma_1^2)$.

We can rewrite (50) as:

$$E[\hat{\alpha}]/\bar{\alpha} = E\left\{\frac{\sum_{i=1}^n (c_i + e_i)c_i}{\sum_{i=1}^n (c_i + e_i)^2}\right\}. \tag{51}$$

Define function $f : \Re^n \to \Re$ via $f(\boldsymbol{x}) = \frac{\boldsymbol{x}^T \boldsymbol{c}}{\boldsymbol{x}^T \boldsymbol{x}}$. We perform second-order Taylor expansion of f around the point $\boldsymbol{x} = \bar{\boldsymbol{c}}$ and then take expectation with respect to $\boldsymbol{x} = \bar{\boldsymbol{c}} + \epsilon_1$:

$$\begin{aligned} E[\hat{\alpha}]/\bar{\alpha} &= E[f(\bar{\boldsymbol{c}}) + \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{c}})^T \nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})(\boldsymbol{x} - \bar{\boldsymbol{c}})] \\ &= 1 + \frac{1}{2}E[(\boldsymbol{x} - \bar{\boldsymbol{c}})^T \nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})(\boldsymbol{x} - \bar{\boldsymbol{c}})] \\ &= 1 + \frac{1}{2}E[\epsilon_1^T \nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})\epsilon_1]. \end{aligned} \tag{52}$$

Now we focus on the term $E[\epsilon_1^T \nabla^2 f_{\boldsymbol{x}}(\bar{\boldsymbol{c}})\epsilon_1]$ whose sign determines the bias. The gradient $\nabla_{\boldsymbol{x}} f$ and the Hessian $\nabla^2_{\boldsymbol{x}} f$ of f are derived as follows:

$$\nabla_{\boldsymbol{x}} f = \|\boldsymbol{x}\|_2^{-2} \bar{\boldsymbol{c}}^T - 2\|\boldsymbol{x}\|_2^{-4} (\boldsymbol{x}^T \bar{\boldsymbol{c}})\boldsymbol{x}^T.$$

The ith element of $\nabla_{\boldsymbol{x}} f$ is

$$[\nabla_x f]_i = \|\boldsymbol{x}\|_2^{-2} c_i - 2\|\boldsymbol{x}\|_2^{-4} (\boldsymbol{x}^T \bar{\boldsymbol{c}})x_i.$$

Taking derivative with respect to x_j yields:

$$\frac{\partial}{\partial x_j}[\nabla_x f]_i = -2\|\boldsymbol{x}\|_2^{-4} c_i x_j - 2\left\{-4\|\boldsymbol{x}\|_2^{-6} \boldsymbol{x}^T \bar{\boldsymbol{c}} x_i x_j + \|\boldsymbol{x}\|_2^{-4} (x_i c_j + \boldsymbol{x}^T \bar{\boldsymbol{c}}\delta[i-j]\right\},$$

where δ is the kronecker impulse function defined as

$$\delta[x] = \begin{cases} 1 & x = 0; \\ 0 & \text{otherwise.} \end{cases}$$

The equivalent matrix representation of the Hessian is given by:

$$\nabla^2_{\boldsymbol{x}} f = 8\|\boldsymbol{x}\|_2^{-6} \boldsymbol{x}^T \bar{\boldsymbol{c}}\boldsymbol{x}\boldsymbol{x}^T - 2\|\boldsymbol{x}\|_2^{-4} (\boldsymbol{x}\bar{\boldsymbol{c}}^T + \bar{\boldsymbol{c}}\boldsymbol{x}^T) - 2(\boldsymbol{x}^T \bar{\boldsymbol{c}})\|\boldsymbol{x}\|_2^{-4} I. \tag{53}$$

Evaluating the Hessian at $\bar{\boldsymbol{c}}$ and noting that the i.i.d structure of the noise ϵ_1 made the computation of $E[\epsilon_1 \nabla^2 f(\bar{\boldsymbol{c}})\epsilon_1] = \sigma_1^2 \,\mathsf{trace}\{\nabla^2 f(\bar{\boldsymbol{c}})\}$ depend only on the diagonal elements of the Hessian, we obtain:

$$[\nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})]_{ii} = 2\|\bar{\boldsymbol{c}}\|_2^{-4} (2c_i^2 - \sum_{j=1}^n c_j^2).$$

Subsequently:

$$\begin{aligned} E[\epsilon_1^T \nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})\epsilon_1] &= \sigma_1^2 \sum_{i=1}^n [\nabla^2_{\boldsymbol{x}} f(\bar{\boldsymbol{c}})]_{ii} \\ &= 2\sigma_1^2 \|\bar{\boldsymbol{c}}\|_2^{-2} (2-n), \end{aligned} \tag{54}$$

which is negative for all $n > 2$.

Subsequently,

$$E[\hat{\alpha}]/\bar{\alpha} \approx 1 - (n-2)\sigma_1^2 \|\bar{\boldsymbol{c}}\|_2^{-2}. \tag{55}$$

As (54) describes the difference between $E[\hat{\alpha}/\bar{\alpha}]$ and unity, this indicates that for $n > 2$, $\hat{\alpha}$ is an estimate of $\bar{\alpha}$ that biases towards smaller magnitude.

Similarly, we compute $\mathsf{Var}\{\hat{\alpha}\}$ via $E[\hat{\alpha}^2]-E[\hat{\alpha}]^2$. The correlation reads:

$$E\left\{\frac{(\bar{\boldsymbol{c}}+\epsilon_1)^T(\bar{\alpha}\bar{\boldsymbol{c}}+\epsilon_2)(\bar{\alpha}\bar{\boldsymbol{c}}+\epsilon_2)^T(\bar{\boldsymbol{c}}+\epsilon_1)}{\|\bar{\boldsymbol{c}}+\epsilon_1\|_2^4}\right\}.$$

As before, we first use conditional expectation to separate out the uncertainty in ϵ_2 via:

$$E[\hat{\alpha}^2]=E_{\epsilon_1}E_{\epsilon_2}[\hat{\alpha}^2|\epsilon_1]=E\left\{\frac{(\bar{\boldsymbol{c}}+\epsilon_1)^T(\bar{\alpha}^2\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T+\sigma_2^2I)(\bar{\boldsymbol{c}}+\epsilon_1)}{\|\bar{\boldsymbol{c}}+\epsilon_1\|_2^4}\right\}.$$

Define a deterministic symmetric matrix $H\triangleq(\bar{\alpha}^2\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T+\sigma_2^2I)$ and a function $f(\boldsymbol{x})=\frac{\boldsymbol{x}^TH\boldsymbol{x}}{\|\boldsymbol{x}\|_2^4}$, and we aim to find $E[f(\boldsymbol{x})]$ for $\boldsymbol{x}=\bar{\boldsymbol{c}}+\epsilon_1$. We expand the function $f(\boldsymbol{x})$ around $\boldsymbol{x}=\bar{\boldsymbol{c}}$ and approximate $E[\hat{\alpha}^2]$ via:

$$\begin{aligned}E[\hat{\alpha}^2] &\approx \frac{\bar{\boldsymbol{c}}^TH\bar{\boldsymbol{c}}}{\|\bar{\boldsymbol{c}}\|_2^4}+\frac{1}{2}E[(\boldsymbol{x}-\bar{\boldsymbol{c}})^T\nabla_{\boldsymbol{x}}^2f(\bar{\boldsymbol{c}})(\boldsymbol{x}-\bar{\boldsymbol{c}})]\\ &= \frac{\bar{\boldsymbol{c}}^TH\bar{\boldsymbol{c}}}{\|\bar{\boldsymbol{c}}\|_2^4}+\frac{1}{2}E[\epsilon_1^T\nabla_{\boldsymbol{x}}^2f(\bar{\boldsymbol{c}})\epsilon_1]. \end{aligned}\tag{56}$$

The deterministic term $f(\bar{\boldsymbol{c}})$ simplifies to:

$$f(\bar{\boldsymbol{c}})=\frac{\bar{\boldsymbol{c}}^TH\bar{\boldsymbol{c}}}{\|\bar{\boldsymbol{c}}\|_2^4}=\bar{\alpha}^2+\frac{\sigma_2^2}{\|\bar{\boldsymbol{c}}\|_2^2}.$$

Since ϵ_1 is componentwise independent, $E[\hat{\alpha}^2]$ only depends on the diagonal element of $\nabla_{\boldsymbol{x}}^2f(\bar{\boldsymbol{c}})$, which we derive as follows.

$$\nabla_{\boldsymbol{x}}f(\boldsymbol{x})=-4\,\|\boldsymbol{x}\|_2^{-6}\,\boldsymbol{x}^T(\boldsymbol{x}^TH\boldsymbol{x})+2\,\|\boldsymbol{x}\|_2^{-4}\,\boldsymbol{x}^TH.$$

The ith element of $\nabla_{\boldsymbol{x}}f(\boldsymbol{x})$ reads $-4\,\|\boldsymbol{x}\|_2^{-6}\,x_i(\boldsymbol{x}^TH\boldsymbol{x})+2\,\|\boldsymbol{x}\|_2^{-4}\,\boldsymbol{x}^TH(:,i)$., where $H(:,i)$ indicates the ith column of H. We may explicitly write $\boldsymbol{x}^TH(:,i)=\sum_j x_j[\bar{\alpha}^2c_ic_j+\sigma_2^2\delta[i-j]]$. The second-order derivative is given by:

$$\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})=-4\,\|\boldsymbol{x}\|_2^{-6}\,[\boldsymbol{x}^TH\boldsymbol{x}+2x_i\boldsymbol{x}^TH(:,i)]+24\,\|\boldsymbol{x}\|_2^{-8}\,x_i^2\boldsymbol{x}^TH\boldsymbol{x}+2\,\|\boldsymbol{x}\|_2^{-4}\,(\bar{\alpha}^2c_i^2+\sigma_2^2)-8\,\|\boldsymbol{x}\|_2^{-6}\,x_i\boldsymbol{x}^TH(:,i).\tag{57}$$

To evaluate $\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})$ at $\boldsymbol{x}=\bar{\boldsymbol{c}}$, we use the following relations:

$$\bar{\boldsymbol{c}}^TH(:,i)=c_i(\bar{\alpha}^2\,\|\bar{\boldsymbol{c}}\|_2^2+\sigma_2^2);$$

$$\bar{\boldsymbol{c}}^TH\bar{\boldsymbol{c}}=\|\bar{\boldsymbol{c}}\|_2^2\,(\bar{\alpha}^2\,\|\bar{\boldsymbol{c}}\|_2^2+\sigma_2^2).$$

Substituting these relations into the expression (57) for $\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})$, we obtain:

$$\frac{\partial^2}{\partial x_i^2}f(\boldsymbol{x})|_{\boldsymbol{x}=\bar{\boldsymbol{c}}}=8\,\|\bar{\boldsymbol{c}}\|^{-6}\,c_i^2(\bar{\alpha}^2\,\|\bar{\boldsymbol{c}}\|_2^2+\sigma_2^2)-4\,\|\bar{\boldsymbol{c}}\|^{-4}\,(\bar{\alpha}^2\,\|\bar{\boldsymbol{c}}\|_2^2+\sigma_2^2)+2\,\|\bar{\boldsymbol{c}}\|^{-4}\,(\bar{\alpha}^2c_i^2+\sigma_2^2).$$

By the independence of the elements in ϵ_1, we obtain:

$$\begin{aligned}E[\epsilon_1^T\nabla_{\boldsymbol{x}}^2f(\bar{\boldsymbol{c}})\epsilon_1] &= \sigma_1^2\sum_i\frac{\partial^2}{\partial x_i^2}f(\bar{\boldsymbol{c}})\\ &= \|\bar{\boldsymbol{c}}\|^{-2}\,(10-4n)\bar{\alpha}^2\sigma_1^2+\|\bar{\boldsymbol{c}}\|^{-4}\,(8-2n)\sigma_1^2\sigma_2^2.\end{aligned}\tag{58}$$

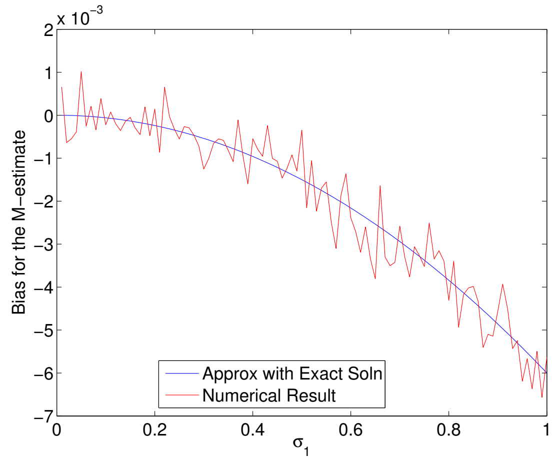
Substituting this quantity into (56) provides:

$$E[\hat{\alpha}^2]\approx\bar{\alpha}^2+\|\bar{\boldsymbol{c}}\|^{-2}\,\sigma_2^2+\|\bar{\boldsymbol{c}}\|^{-2}\,(5-2n)\bar{\alpha}^2\sigma_1^2+\|\bar{\boldsymbol{c}}\|^{-4}\,(4-n)\sigma_1^2\sigma_2^2.$$

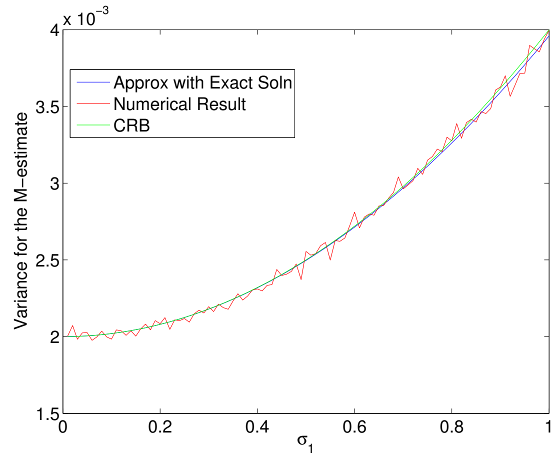
Together with the estimation for $E[\hat{\alpha}]$ obtained in (55), this equation yields an approximation for $\mathsf{Var}\{\hat{\alpha}\}$ as:

$$\begin{aligned}\mathsf{Var}\{\hat{\alpha}\} &= E[\hat{\alpha}^2] - E[\hat{\alpha}]^2 \\ &= \bar{\alpha}^2 + \|\bar{\boldsymbol{c}}\|^{-2}\sigma_2^2 + \|\bar{\boldsymbol{c}}\|^{-2}(5-2n)\bar{\alpha}^2\sigma_1^2 + \|\bar{\boldsymbol{c}}\|^{-4}(4-n)\sigma_1^2\sigma_2^2 - (1+\|\bar{\boldsymbol{c}}\|^{-2}(2-n)\sigma_1^2)^2\bar{\alpha}^2 \\ &= \|\bar{\boldsymbol{c}}\|^{-2}(\bar{\alpha}^2\sigma_1^2+\sigma_2^2) - \|\bar{\boldsymbol{c}}\|^{-4}\sigma_1^2[(n-4)\sigma_2^2 - (n-2)^2\bar{\alpha}^2\sigma_1^2]. \end{aligned} \tag{59}$$

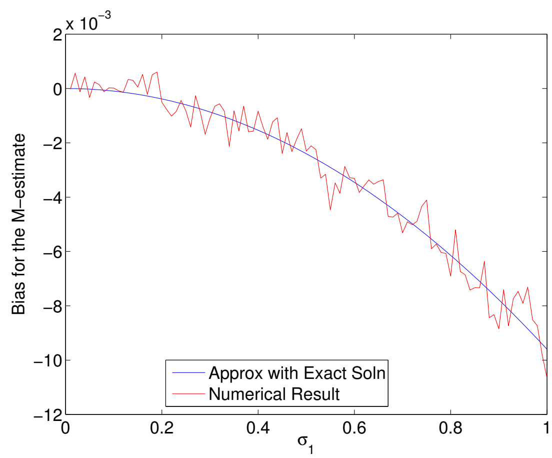
Expressions (55) and (59) reveal some interesting structure. For large enough n (in fact for $n > 6$), the variance estimate (59) becomes upper-bounded by $\|\bar{\boldsymbol{c}}\|^{-2}(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)$, which we will show later is the Cramér-Rao Bound for the statistical model. This implies that it cannot be unbiased. In fact, the bias quantity measured by $(2-n)\|\bar{\boldsymbol{c}}\|^{-2}\sigma_1^2\bar{\alpha}$ also increases accordingly.



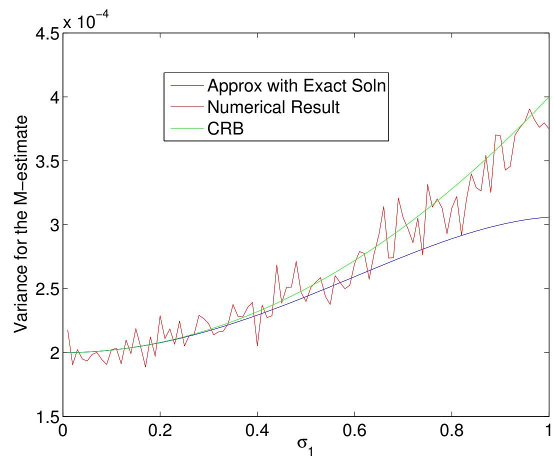
(a1) bias approximation $n = 5$



(a2) variance approximation $n = 5$



(b1) bias approximation $n = 50$



(b2) variance approximation $n = 50$

Fig. 1. Bias and variance approximation obtained from explicit solution for conventional M-estimate.

Alternatively, we can follow [6], use implicit function theorem and Taylor expansion to approximate the bias and variance of $\hat{\alpha}$ as the minimizer of (48). The data point $\check{\boldsymbol{z}}$ at which to perform Taylor expansion about is mainly a choice of convenience rather than considerations of asymptotic behavior. One natural choice of the expansion point would be the noiseless data. Let $\bar{\boldsymbol{z}}$ denote the noiseless observation and $\bar{\boldsymbol{c}}$ and $\bar{\alpha}$ denote the true parameter values, with $\check{\boldsymbol{c}}$ and $\check{\alpha}$ denoting the resulting estimation in (47) and (49) when $\bar{\boldsymbol{z}}$ is observed. Then $\bar{\boldsymbol{z}} = [\bar{\boldsymbol{c}}; \bar{\alpha}\bar{\boldsymbol{c}}]$, and

$$\begin{aligned}\check{\boldsymbol{c}} &= \hat{\boldsymbol{c}}(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{c}}; \\ \check{\alpha} &= \hat{\alpha}(\bar{\boldsymbol{z}}) = \frac{\bar{\alpha}\bar{\boldsymbol{c}}^T\bar{\boldsymbol{c}}}{\|\bar{\boldsymbol{c}}\|_2^2} = \bar{\alpha}. \end{aligned} \tag{60}$$

As the minimizer for (48), $\hat{\alpha}$ satisfies:

$$\frac{\partial}{\partial \alpha}\Phi(\alpha, \boldsymbol{z})|_{\alpha=\hat{\alpha}} = 2\boldsymbol{z}^T \begin{bmatrix} I \\ 0 \end{bmatrix} \begin{bmatrix} \alpha I & -I \end{bmatrix} \boldsymbol{z} = 0 \quad \forall \boldsymbol{z}.$$

Taking derivative with respect to $\boldsymbol{z}$ and invoking the chain rule, we obtain:

$$\frac{\partial^2}{\partial \alpha^2}\Phi \frac{\partial}{\partial \boldsymbol{z}}\alpha + \frac{\partial^2}{\partial \alpha \partial \boldsymbol{z}}\Phi = 0,$$

where

$$\frac{\partial^2}{\partial \alpha^2}\Phi = 2\left\|z_1\right\|_2^2 = 2\boldsymbol{z}^T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{z}, \tag{61}$$

and

$$\frac{\partial^2}{\partial \alpha \partial \boldsymbol{z}}\Phi = 2\boldsymbol{z}^T \left\{ \begin{bmatrix} \alpha I & -I \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha I & 0 \\ -I & 0 \end{bmatrix} \right\} = 2\boldsymbol{z}^T \begin{bmatrix} 2\alpha I & -I \\ -I & 0 \end{bmatrix}. \tag{62}$$

Therefore,

$$\frac{\partial}{\partial \boldsymbol{z}}\hat{\alpha}(\boldsymbol{z}) = -\frac{\partial^2}{\partial \alpha^2}\Phi^{-1}\frac{\partial^2}{\partial \alpha \partial \boldsymbol{z}}\Phi = -\left\|z_1\right\|_2^{-2} \boldsymbol{z}^T \begin{bmatrix} 2\alpha I & -I \\ -I & 0 \end{bmatrix}. \tag{63}$$

Evaluating (63) at $\boldsymbol{z} = \bar{\boldsymbol{z}}$, we obtain an estimate of covariance $\mathsf{Cov}\{\alpha\}$ at $\check{\alpha} = \hat{\alpha}\boldsymbol{z}$ as

$$\begin{aligned} \mathsf{Cov}\{\hat{\alpha}(\boldsymbol{z})\} &\approx \frac{\partial}{\partial \boldsymbol{z}}\alpha(\bar{\boldsymbol{z}})\,\mathsf{Cov}\{z\}\,\frac{\partial}{\partial \boldsymbol{z}}\alpha^T(\bar{\boldsymbol{z}}) \\ &= \frac{-1}{\|\bar{\boldsymbol{c}}\|_2^2}\bar{\boldsymbol{c}}^T \begin{bmatrix} \bar{\alpha} I & -I \end{bmatrix} \begin{bmatrix} \sigma_1^2 I & \\ & \sigma_2^2 I \end{bmatrix} \frac{-1}{\|\bar{\boldsymbol{c}}\|_2^2} \begin{bmatrix} \bar{\alpha} I \\ -I \end{bmatrix} \bar{\boldsymbol{c}} \\ &= \frac{\bar{\alpha}^2\sigma_1^2 + \sigma_2^2}{\|\bar{\boldsymbol{c}}\|_2^2}. \end{aligned} \tag{64}$$

This quantity (64) coincides with the Cramér-Rao Bound obtained from the statistical model as we will show later.

To estimate the bias for $\hat{\alpha}$, we present the first and second-order Taylor expansion for $E[\hat{\alpha}]$ as:

$$\begin{aligned} E^{(1)}[\hat{\alpha}] &= E[h(\boldsymbol{z})] \\ &\approx E\left\{h(\check{\boldsymbol{z}}) + \nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} \\ &= h(\check{\boldsymbol{z}}) + E\left\{\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\}. \end{aligned} \tag{65}$$

$$\begin{aligned} E^{(2)}[\hat{\alpha}] &\approx E\left\{h(\check{\boldsymbol{z}}) + \nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}}) + \frac{1}{2}(\boldsymbol{z} - \check{\boldsymbol{z}})^T \nabla_{\boldsymbol{z}}^2 h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} \\ &= h(\check{\boldsymbol{z}}) + E\left\{\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} + \frac{1}{2}E\left\{(\boldsymbol{z} - \check{\boldsymbol{z}})^T \nabla_{\boldsymbol{z}}^2 h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\}. \end{aligned} \tag{66}$$

Notice that when $\check{\boldsymbol{z}}$ is chosen to be $\bar{\boldsymbol{z}}$, $\boldsymbol{z} - \check{\boldsymbol{z}}$ is zero mean Gaussian. It follows that the first order term $E\left\{\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} = 0$ in (65) and (66). Therefore, the first order Taylor approximation yields:

$$E^{(1)}[\hat{\alpha}] = h(\check{\boldsymbol{z}}) = h(\bar{\boldsymbol{z}}) = \bar{\alpha}, \tag{67}$$

corresponding to zero bias.

The second-order approximation (66) requires computing $\nabla_{\boldsymbol{z}}^2 h(\check{\boldsymbol{z}})$, which can be be obtained up to second order [6] via:

$$\nabla_{\boldsymbol{z}}^2 h = [-\frac{\partial^2}{\partial \alpha^2}\Phi]^{-1} \left\{ \frac{\partial^3}{\partial \alpha^3}\Phi \nabla_{\boldsymbol{z}} h^T \nabla_{\boldsymbol{z}} h + \frac{\partial^3}{\partial \alpha^2 \partial \boldsymbol{z}}\Phi^T \nabla_{\boldsymbol{z}} h + \nabla_{\boldsymbol{z}} h^T \frac{\partial^3}{\partial \alpha^2 \partial \boldsymbol{z}}\Phi + \frac{\partial}{\partial \alpha}\nabla_{\boldsymbol{z}}^2 \Phi \right\}. \tag{68}$$

Terms involved in the above expression are computed as follows:

$$\frac{\partial^3}{\partial\alpha^3}\Phi = 0.$$

Taking derivative of (61) with respect to $\boldsymbol{z}$ yields

$$\frac{\partial^3}{\partial\alpha^2\partial\boldsymbol{z}}\Phi = 2\boldsymbol{z}^T \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = 2 \begin{bmatrix} z_1^T & 0 \end{bmatrix}.$$

Taking derivative of (62) with respect to $\boldsymbol{z}$ yields

$$\frac{\partial^3}{\partial\alpha\partial\boldsymbol{z}^2}\Phi = 2 \begin{bmatrix} 2\alpha I & -I \\ -I & 0 \end{bmatrix}.$$

Evaluating at $\boldsymbol{z} = \check{\boldsymbol{z}} = \bar{\boldsymbol{z}}$ and substituting into (68) yields:

$$\nabla^2_{\boldsymbol{z}} h(\bar{\boldsymbol{z}}) = -\frac{1}{2\,\|\bar{\boldsymbol{c}}\|_2^2}\left\{\frac{-2}{\|\bar{\boldsymbol{c}}\|_2^2}\begin{bmatrix} \bar{\boldsymbol{c}} \\ 0 \end{bmatrix}\begin{bmatrix} \bar{\alpha}\bar{\boldsymbol{c}}^T & -\bar{\boldsymbol{c}}^T \end{bmatrix} + \frac{-2}{\|\bar{\boldsymbol{c}}\|_2^2}\begin{bmatrix} \bar{\alpha}\bar{\boldsymbol{c}} \\ -\bar{\boldsymbol{c}} \end{bmatrix}\begin{bmatrix} \bar{\boldsymbol{c}}^T & 0 \end{bmatrix} + 2\begin{bmatrix} 2\bar{\alpha} I & -I \\ -I & 0 \end{bmatrix}\right\}. \tag{69}$$

Since $\boldsymbol{z} - \check{\boldsymbol{z}} \sim \mathcal{N}(0, \begin{bmatrix} \sigma_1^2 I & 0 \\ 0 & \sigma_2^2 I \end{bmatrix}$, the second-order term in (66) only involves the diagonal elements of $\nabla^2_{\boldsymbol{z}} h(\bar{\boldsymbol{z}})$. We extract the corresponding blocks from (69) as:

$$\frac{\partial^2}{\partial z_1(i)^2} h(\bar{\boldsymbol{z}}) = -\frac{1}{2\,\|\bar{\boldsymbol{c}}\|_2^2}\left\{\frac{-4}{\|\bar{\boldsymbol{c}}\|_2^2}\bar{\alpha} c_i^2 + 4\bar{\alpha}\right\};$$

$$\frac{\partial^2}{\partial z_1(i)^2} h(\bar{\boldsymbol{z}}) = 0.$$

Thus

$$\begin{aligned} E\left\{(\boldsymbol{z} - \bar{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\bar{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} &= \sum_i \sigma_1^2 \frac{\partial^2}{\partial z_1(i)^2} h(\bar{\boldsymbol{z}}) \\ &= \frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2}(2\bar{\alpha} - 2\bar{\alpha} n) \\ &= 2(1-n)\bar{\alpha}\frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2}. \end{aligned} \tag{70}$$

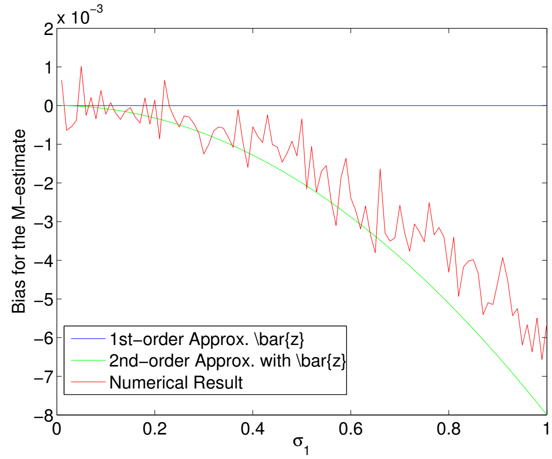
It follows that the second-order estimation for $E[\hat{\alpha}]$ is

$$E^{(2)}[\hat{\alpha}] = E^{(1)}[\hat{\alpha}] + \frac{1}{2} E\left\{(\boldsymbol{z} - \bar{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\bar{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\} = \bar{\alpha} + (1-n)\frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2}\bar{\alpha} = \left\{1 + (1-n)\frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2}\right\}\bar{\alpha}. \tag{71}$$

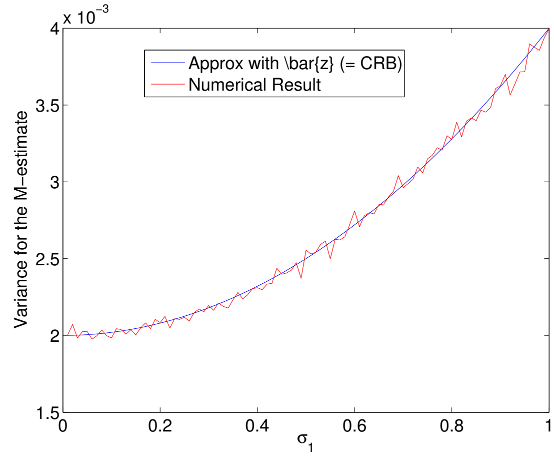
For $n > 1$ and reasonable signal-to-noise ratio, $E^{(2)}[\hat{\alpha}]$ implies shrinkage in magnitude, which WLOG, we refer to as "negative bias" hereafter.

Notice that the choice of $\check{\boldsymbol{z}} = \bar{\boldsymbol{z}}$ is mainly due to computation convenience (so that $\boldsymbol{z} - \check{\boldsymbol{z}}$ is zero mean Gaussian). It is feasible to perform the same routine for different data point $\check{\boldsymbol{z}}$. [8,9] proved that under certain regular conditions, the M-estimate is asymptotically normal with mean $\tilde{\alpha}$ where

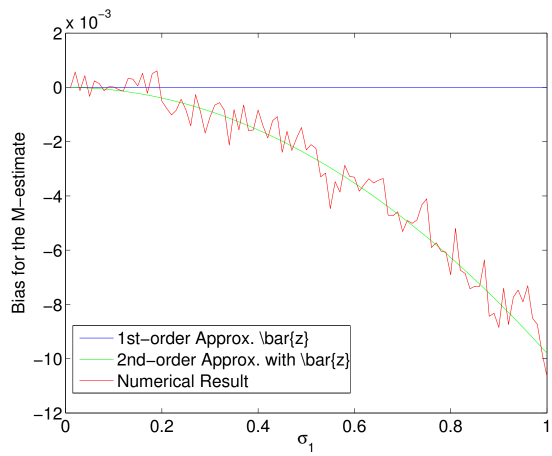
$$E[\frac{\partial}{\partial\alpha}\Phi(\tilde{\alpha}, \boldsymbol{z})] = 0.$$



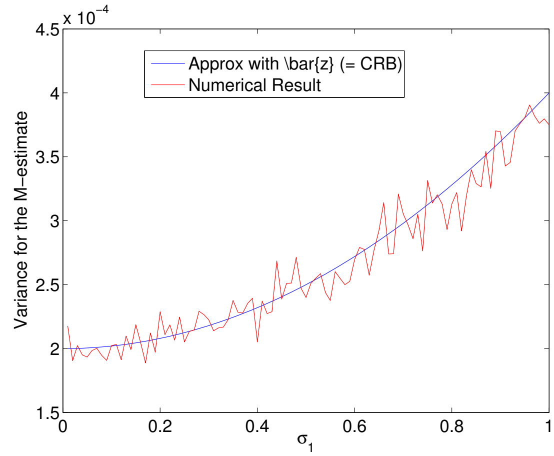
(a1) bias approximation $n = 5$



(a2) variance approximation $n = 5$



(b1) bias approximation $n = 50$



(b2) variance approximation $n = 50$

Fig. 2. Bias and variance approximation for M-estimate obtained from expansion about $(\bar{\alpha}, \bar{\boldsymbol{z}})$.

Under reasonable regularity conditions, we can exchange the order of expectation and differentiation, and take

$$\frac{\partial}{\partial \alpha} E[\Phi(\tilde{\alpha}, \boldsymbol{z})]] = 0.$$

Note that $\tilde{\alpha}$ can be interpreted as a local minima for an "average" cost function $E[\Phi(\alpha, \boldsymbol{z})]$, *i.e.*,

$$\tilde{\alpha} = \arg\min_{\alpha} E[\Phi(\alpha, \boldsymbol{z})]. \tag{72}$$

The expectation of the objective function with respect to the distribution of the observation noise

$$\begin{aligned} E[\Phi(\alpha, \boldsymbol{z})] &= E[\left\| \begin{bmatrix} -\alpha I & I \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{c}} + \epsilon_1 \\ \bar{\alpha}\bar{\boldsymbol{c}} + \epsilon_2 \end{bmatrix} \right\|_2^2] \\ &= E[\begin{bmatrix} \bar{\boldsymbol{c}}^T + \epsilon_1^T & \bar{\alpha}\bar{\boldsymbol{c}}^T + \epsilon_2^T \end{bmatrix} \begin{bmatrix} -\alpha I \\ I \end{bmatrix} \begin{bmatrix} -\alpha I & I \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{c}} + \epsilon_1 \\ \bar{\alpha}\bar{\boldsymbol{c}} + \epsilon_2 \end{bmatrix}] \\ &= (\alpha - \bar{\alpha})^2 \|\bar{\boldsymbol{c}}\|_2^2 + n(\alpha^2\sigma_1^2 + \sigma_2^2) \\ &= (\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1^2)\alpha^2 - 2\bar{\alpha}\|\bar{\boldsymbol{c}}\|_2^2 \alpha + \bar{\alpha}^2 \|\bar{\boldsymbol{c}}\|_2^2 \end{aligned} \tag{73}$$

is convex quadratic in α and the minimizer reads

$$\begin{aligned} \breve{\alpha} &= \arg\min_{\alpha} E[\Psi(\alpha, \boldsymbol{z})] \\ &= \frac{\|\bar{\boldsymbol{c}}\|_2^2}{\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1^2}\bar{\alpha}. \end{aligned} \tag{74}$$

For simplicity, let $\beta \triangleq \frac{\|\bar{c}\|_2^2+n\sigma_1^2}{\|\bar{c}\|_2^2}$, then $\check{\alpha} = \frac{1}{\beta}\bar{\alpha}$. Since $\beta > 1$, the expansion point $\check{\alpha}$ is a shrinkage with respect to the true scale $\bar{\alpha}$.

We can construct an expansion point $\check{\boldsymbol{z}} = [\beta\bar{\boldsymbol{c}}; \bar{\alpha}\bar{c}]$. Then the minimizer of $\Phi(\check{\boldsymbol{z}}) = \frac{1}{\beta}\bar{\alpha} = \check{\alpha}$, which satisfies the requirement (72).

Evaluating (63) at $(\check{c}, \check{\boldsymbol{z}})$ results in:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{z}}\alpha(\check{\boldsymbol{z}}) &= -\|z_1\|_2^{-2}\,\boldsymbol{z}^T \begin{bmatrix} 2\alpha I & -I \\ -I & 0 \end{bmatrix} \\ &= -\frac{1}{\beta^2\,\|\bar{\boldsymbol{c}}\|_2^2} \begin{bmatrix} \beta\bar{\boldsymbol{c}}^T & \check{\alpha}\bar{\boldsymbol{c}} \end{bmatrix} \begin{bmatrix} 2\check{\alpha} I & -I \\ -I0 & \end{bmatrix} \\ &= -\frac{\bar{\boldsymbol{c}}^T}{\beta^2\,\|\bar{\boldsymbol{c}}\|_2^2} \begin{bmatrix} \frac{2\beta-1}{\beta}\bar{\alpha} I & -\beta I \end{bmatrix}. \end{aligned} \tag{75}$$

The approximated covariance of $\hat{\alpha}$ evaluated at the point $(\check{\alpha}, \check{\boldsymbol{z}})$ is given by:

$$\begin{aligned} \mathsf{Cov}\{\hat{\alpha}\}\,|_{\boldsymbol{z}=\check{\boldsymbol{z}},\hat{\alpha}=\check{\alpha}} &= \frac{\partial}{\partial \boldsymbol{z}}\alpha(\check{\boldsymbol{z}})\,\mathsf{Cov}\{\boldsymbol{z}\}\,\frac{\partial}{\partial \boldsymbol{z}}\alpha(\check{\boldsymbol{z}}) \\ &= \beta^{-4}\,\|\bar{\boldsymbol{c}}\|_2^{-4}\,\bar{\boldsymbol{c}}^T \begin{bmatrix} \frac{2\beta-1}{\beta}\bar{\alpha} I & -\beta I \end{bmatrix} \begin{bmatrix} \sigma_1^2 I & 0 \\ 0 & \sigma_2^2 I \end{bmatrix} \begin{bmatrix} \frac{2\beta-1}{\beta}\bar{\alpha} I \\ -\beta I \end{bmatrix} \bar{\boldsymbol{c}} \\ &= \|\bar{\boldsymbol{c}}\|_2^{-2}\,\beta^{-4}((2-\frac{1}{\beta})^2\bar{\alpha}^2\sigma_1^2 + \beta^2\sigma_2^2). \end{aligned} \tag{76}$$

We know from previous analysis that the M-estimate is asymptotically unbiased, so its variance is to be bounded below by Cramér-Rao Bound asymptotically. Therefore, it is curious to find whether there exists a consistent relationship between the pre-asymptotic variance in (76) and the Cramér-Rao Bound , *i.e.*,

$$\|\bar{\boldsymbol{c}}\|_2^{-2}\,\beta^{-4}((2-\frac{1}{\beta})^2\bar{\alpha}^2\sigma_1^2 + \beta^2\sigma_2^2) \gtrless \|\bar{\boldsymbol{c}}\|_2^{-2}\,(\bar{\alpha}^2\sigma_1^2 + \sigma_2^2)? \tag{77}$$

The quantity on the right-hand-side is the Cramér-Rao Bound obtained from the statistical generative model (to be shown later).

Claim 1: The covariance of the M-estimator is bounded *above* by the Cramér-Rao Bound . Moreover, it asymptotically approaches the Cramér-Rao Bound as $\sigma_1 \to 0$.

Proof: To compare the left and right hand sides in (77), it suffices determine the sign of their difference:

$$\text{RHS} - \text{LHS} = \|\bar{\boldsymbol{c}}\|_2^{-2}\,\beta^{-2}(\beta^6 - 4\beta^2 + 4\beta - 1)\bar{\alpha}^2\sigma_1^2 + (\beta^4 - 1)\sigma_2^2.$$

For simplicity, we drop the positive quantity $\|\bar{\boldsymbol{c}}\|_2^{-2}$ in later analysis as it does not affect the sign. Let $A \triangleq \bar{\alpha}^2\sigma_1^2$, $B \triangleq \sigma_2^2$, and we want to determine the sign for:

$$\pi(A, B; \beta) = \beta^{-2}(\beta^6 - 4\beta^2 + 4\beta - 1)A + (\beta^4 - 1)B.$$

The polynomial $(\beta^6 - 4\beta^2 + 4\beta - 1)$ factors into

$$\beta^6 - 4\beta^2 + 4\beta - 1 = (\beta - 1)(\beta^2 + \beta - 1)(\beta^3 + 2\beta - 1).$$

By construction, $\beta > 1$, thus $(\beta^6 - 4\beta^2 + 4\beta - 1) > 0$, so π is linear in A, B with positive coefficients. Meanwhile, A,B are both positive, so $\pi(A, B; \beta) > 0$. This result translates into the claim that in nondegenerative case ($\sigma_1 \neq 0$), the variance of the M-estimate is bounded above by the Cramér-Rao Bound . It is easy to check that when $\sigma_1 = 0$, the variance equals the Cramér-Rao Bound . ■

Now we approximate $E[\hat{\alpha}]$ with (65) and (66) by expanding corresponding terms about $(\check{\alpha}, \check{\boldsymbol{z}})$.
The first order coefficient $\nabla_{\boldsymbol{z}} h$ is obtained in (75), and the corresponding first-order approximation for the mean is:

$$\begin{aligned}
E^{(1)}[\hat{\alpha}] &= h(\check{\boldsymbol{z}}) + E\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}}) \\
&= \frac{\bar{\alpha}}{\beta} + E\left\{ -\frac{\bar{\boldsymbol{c}}^T}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left[\begin{array}{cc} \frac{2\beta-1}{\beta}\bar{\alpha} I & -\beta I \end{array} \right] \left[\begin{array}{c} \bar{\boldsymbol{c}} + \epsilon_1 - \beta \bar{\boldsymbol{c}} \\ \bar{\alpha}\bar{\boldsymbol{c}} + \epsilon_2 - \bar{\alpha}\bar{\boldsymbol{c}} \end{array} \right] \right\} \\
&= \frac{\bar{\alpha}}{\beta} + \frac{\bar{\alpha}}{\beta^3 \|\bar{\boldsymbol{c}}\|_2^2} \bar{\boldsymbol{c}}^T (2\beta - 1)(\beta - 1)\bar{\boldsymbol{c}} \\
&= \frac{\bar{\alpha}}{\beta}[1 + \frac{(2\beta - 1)(\beta - 1)}{\beta^2}] \\
&= \frac{3\beta^2 - 3\beta + 1}{\beta^3}\bar{\alpha}. \qquad (78)
\end{aligned}$$

Since $\beta > 1$, $(\beta - 1)^3 = \beta^3 - 3\beta^2 + 3\beta - 1 = \beta^3 - (3\beta^2 - 3\beta + 1) > 0$. Consequently $\frac{3\beta^2 - 3\beta + 1}{\beta^3} < 1$. Equivalently, $\frac{E[\hat{\alpha}]}{\bar{\alpha}} < 1$, indicating a shrinkage in magnitude, which agrees qualitatively with the result from exact solution.
Expression in (78) can be rewritten as:

$$\begin{aligned}
E[\hat{\alpha}] &= \frac{3\beta^2 - 3\beta + 1}{\beta^3}\bar{\alpha} \\
&= [1 - \frac{(\beta - 1)^3}{\beta^3}]\bar{\alpha}. \qquad (79)
\end{aligned}$$

Denote the signal-to-noise ratio in z_1 as $s \triangleq \frac{\|\bar{\boldsymbol{c}}\|_2^2}{n\sigma_1^2}$ and

$$\frac{E[\hat{\alpha}]}{\bar{\alpha}} = 1 - \frac{1}{(s+1)^3}.$$

To approximate the bias with second-order Taylor expansion, we use (68) and evaluate at $(\bar{\alpha}/\beta, \check{\boldsymbol{z}} =)$.

$$\begin{aligned}
\nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}}) &= -\frac{1}{\|\beta\bar{\boldsymbol{c}}\|_2^2}\{ -\frac{1}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left[\begin{array}{c} \beta\bar{\boldsymbol{c}} \\ 0 \end{array} \right] \left[\begin{array}{cc} \frac{2\beta-1}{\beta}\bar{\alpha}\bar{\boldsymbol{c}}^T & -\beta\bar{\boldsymbol{c}}^T \end{array} \right] - \frac{1}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left[\begin{array}{c} \frac{2\beta-1}{\beta}\bar{\alpha}\bar{\boldsymbol{c}} \\ -\beta\bar{\boldsymbol{c}} \end{array} \right] \left[\begin{array}{cc} \beta\bar{\boldsymbol{c}}^T & 0 \end{array} \right] + \ldots \\
&\quad + \left[\begin{array}{cc} 2\frac{\bar{\alpha}}{\beta} I & -I \\ -I & 0 \end{array} \right] \}.
\end{aligned}$$

To compute $(\boldsymbol{z} - \check{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})$ in (66), it suffices to use only the diagonal blocks of $\nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})$, because the components of $\boldsymbol{z} - \check{\boldsymbol{z}} = \left[\begin{array}{c} \bar{\boldsymbol{c}} + \epsilon_1 - \beta\bar{\boldsymbol{c}} \\ \bar{\alpha}\bar{\boldsymbol{c}} + \epsilon_2 - \bar{\alpha}\bar{\boldsymbol{c}} \end{array} \right] = \left[\begin{array}{c} (1-\beta)\bar{\boldsymbol{c}} + \epsilon_1 \\ \epsilon_2 \end{array} \right]$ are independent. Partition $\boldsymbol{z} - \boldsymbol{z}$ into the deterministic ψ and random part η so that $\psi = \left[\begin{array}{c} (1-\beta)\bar{\boldsymbol{c}} \\ 0 \end{array} \right]$ and $\eta = \left[\begin{array}{c} \epsilon_1 \\ \epsilon_2 \end{array} \right]$. Then the quadratic term in the second-order Taylor expansion in (66) can be written as:

$$E[(\psi + \eta)^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\psi + \eta)] = \psi^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})\psi + E[\eta^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})\eta],$$

where expectation of cross terms between ψ and η are dropped since η is zero-mean.
The diagonal portion of $\nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})$ reads:

$$\nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}}) = \frac{2}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left\{ \frac{1}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left[\begin{array}{cc} \beta(2\beta - 1)\bar{\alpha}\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T & \\ & 0 \end{array} \right] - \left[\begin{array}{cc} \frac{\bar{\alpha}}{\beta} I & \\ & 0 \end{array} \right] \right\}. \qquad (80)$$

It follows that

$$\begin{aligned}
\psi^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}}) \psi &= \frac{2(\beta-1)^2}{\|z_1\|_2^2} \left\{ \frac{(\beta-1)\|\bar{\boldsymbol{c}}\|_2^4}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \bar{\alpha} - \frac{\|\bar{\boldsymbol{c}}\|_2^2}{\beta} \bar{\alpha} \right\} \\
&= \frac{2(\beta-1)^2}{\beta^3} [\frac{(2\beta-1)}{\beta} - 1] \bar{\alpha} \\
&= \frac{2(\beta-1)^3}{\beta^4} \bar{\alpha}.
\end{aligned} \tag{81}$$

$$\begin{aligned}
\eta^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}}) \eta &= \frac{2}{\beta^2 \|\bar{\boldsymbol{c}}\|_2^2} \left\{ \frac{\beta(2\beta-1)\sigma_1^2 \|\bar{\boldsymbol{c}}\|_2^2}{\beta^3 \|\bar{\boldsymbol{c}}\|_2^2} \bar{\alpha} - \frac{n\sigma_1^2}{\beta} \bar{\alpha} \right\} \\
&= \frac{2\sigma_1^2}{\beta^6 \|\bar{\boldsymbol{c}}\|_2^2} [(2\beta-1) - n\beta] \bar{\alpha}.
\end{aligned} \tag{82}$$

Summing (81) and (82) yields:

$$E[(\boldsymbol{z}-\check{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z}-\check{\boldsymbol{z}})] = \frac{2(\beta-1)^3}{\beta^4} \bar{\alpha} + \frac{2\sigma_1^2}{\beta^6 \|\bar{\boldsymbol{c}}\|_2^2} [(2\beta-1) - n\beta] \bar{\alpha}. \tag{83}$$

Combining (83) with the first order estimation of $E[\hat{\alpha}]$, we obtain the second order approximation for $E[\hat{\alpha}]$ as:

$$\begin{aligned}
E^{(2)}[\hat{\alpha}] &= h(\check{\boldsymbol{z}}) + E\left\{ \nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z}-\check{\boldsymbol{z}}) + \frac{1}{2}(\boldsymbol{z}-\check{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z}-\check{\boldsymbol{z}}) \right\} \\
&= E^{(1)}[\hat{\alpha}] + \frac{1}{2} E[(\boldsymbol{z}-\check{\boldsymbol{z}})^T \nabla^2_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z}-\check{\boldsymbol{z}})] \\
&= \left\{ \frac{\beta^3 - (\beta-1)^3}{\beta^3} + \frac{(\beta-1)^3}{\beta^4} \right\} \bar{\alpha} + \frac{(2-n)\beta - 1}{\beta^5} \frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2} \bar{\alpha} \\
&= \frac{\beta^4 - (\beta-1)^4}{\beta^4} \bar{\alpha} + \frac{(2-n)\beta - 1}{\beta^5} \frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2} \bar{\alpha}
\end{aligned} \tag{84}$$

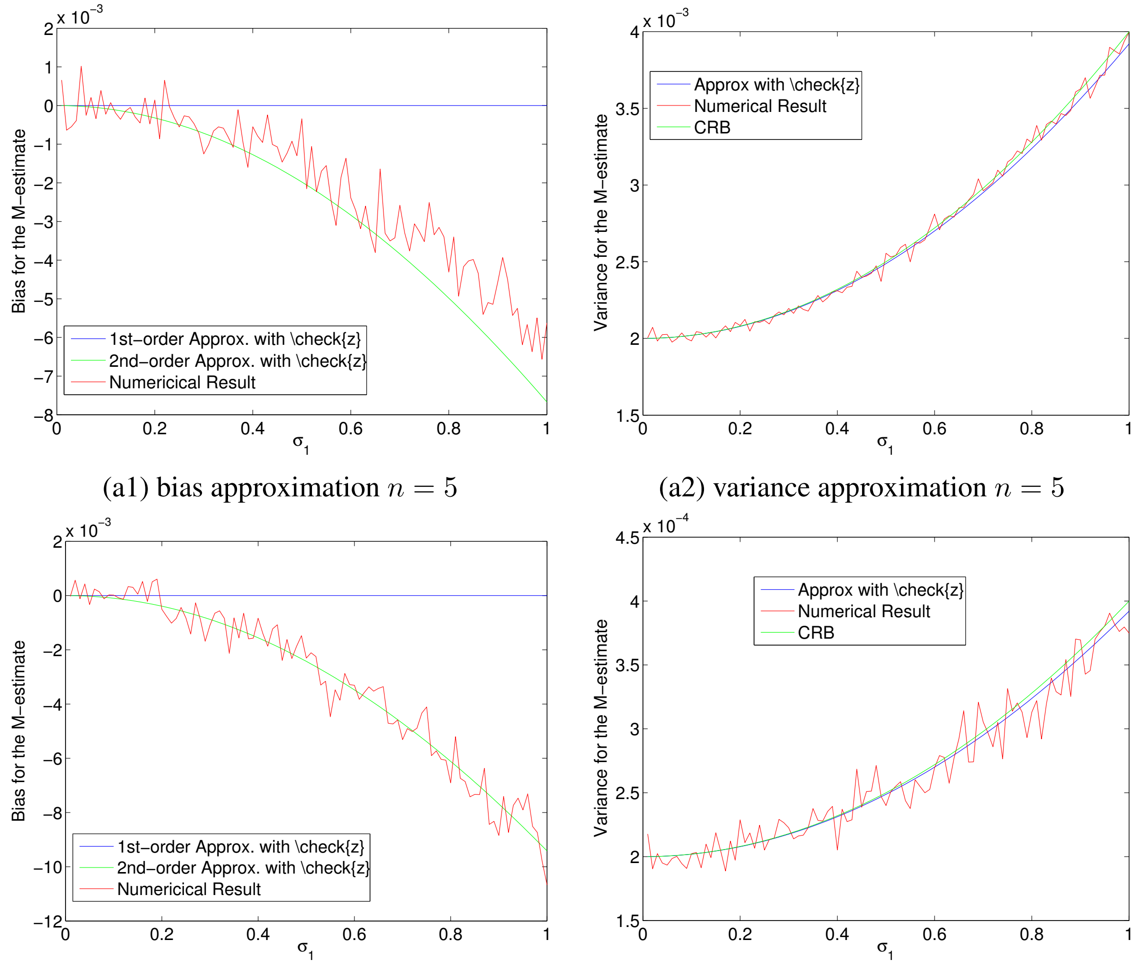
Recall that $\beta = \frac{\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2}$, so for reasonable SNR, $\frac{(2-n)\beta-1}{\beta} \approx 1-n$. Using the $s = \frac{\|\bar{\boldsymbol{c}}\|_2^2}{n\sigma_1^2}$, we can rewrite $E^{(2)}[\hat{\alpha}]$ approximately as:

$$E^{(2)}[\hat{\alpha}] = [1 - \frac{1}{(s+1)^4} + \frac{(1-n)s^3}{n(1+s)^4}] \bar{\alpha}. \tag{85}$$

Notice that when SNR is high (large s), then

$$\begin{aligned}
E^{(2)}[\hat{\alpha}] &= [1 - \frac{1}{(s+1)^4} + \frac{(1-n)s^3}{n(1+s)^4}] \bar{\alpha} \\
&\approx [1 + \frac{1-n}{n(1+s)}] \bar{\alpha} \\
&= [1 - \frac{1-n}{n} \frac{n\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1^2}] \bar{\alpha} \\
&\approx [1 + (1-n) \frac{\sigma_1^2}{\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1^2}] \bar{\alpha},
\end{aligned} \tag{86}$$

which closely resembles the result (71) obtained from expanding about noiseless data $\bar{\boldsymbol{z}}$. In fact, for high enough SNR, $\frac{\|\bar{\boldsymbol{c}}\|_2^2 + n\sigma_1}{\sigma_1^2} \approx \frac{\|\bar{\boldsymbol{c}}\|_2^2}{\sigma_1^2}$ so that (86) and (71) are approximately equal. This relation is expected, as for small SNR, $\check{\boldsymbol{z}} \approx \bar{\boldsymbol{z}}$ and $\check{\alpha} \approx \bar{\alpha}$, the small error analysis is essentially performed on the same neighborhood!



(a1) bias approximation $n = 5$ (a2) variance approximation $n = 5$

(b1) bias approximation $n = 50$ (b2) variance approximation $n = 50$

Fig. 3. Bias and variance approximation of M-estimate obtained from expansion about $(\check{\alpha}, \check{\boldsymbol{z}})$.

.3 ML Estimator for the Statistical Model

The maximum likelihood estimator from (23) aims to jointly estimate $\boldsymbol{c}$ and α via:

$$[\hat{\alpha}, \hat{\boldsymbol{c}}] = \arg\min_{\alpha, \boldsymbol{c}} \frac{1}{\sigma_1^2} \|z_1 - \boldsymbol{c}\|_2^2 + \frac{1}{\sigma_2^2} \|z_2 - \alpha \boldsymbol{c}\|_2^2 . \tag{87}$$

Note that conditioned on α, (87) is quadratic in $\boldsymbol{c}$ with the solution $\hat{\boldsymbol{c}}(\alpha, \boldsymbol{z})$ given by:

$$\begin{aligned}
\hat{\boldsymbol{c}} &= \left\{ \begin{bmatrix} I \\ \alpha I \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} I & 0 \\ 0 & \frac{1}{\sigma_2^2} I \end{bmatrix} \begin{bmatrix} I \\ \alpha I \end{bmatrix} \right\}^{-1} \begin{bmatrix} I & \alpha I \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} I & 0 \\ 0 & \frac{1}{\sigma_2^2} I \end{bmatrix} \boldsymbol{z} \\
&= \left(\frac{1}{\sigma_1^2} + \frac{\alpha^2}{\sigma_2^2}\right)^{-1} \left(\frac{1}{\sigma_1^2} z_1 + \frac{\alpha^2}{\sigma_2^2} z_2\right) \\
&= \frac{1}{\alpha^2 \sigma_1^2 + \sigma_2^2} (\sigma_2^2 z_1 + \alpha \sigma_1^2 z_2).
\end{aligned} \tag{88}$$

Remark:

- In the limiting case when $\sigma_1 \to 0$ (with non-vanishing σ_2), z_1 is a noise-free observation of $\boldsymbol{c}$, it is natural to estimate $\boldsymbol{c}$ solely on z_1 as (88) reduces to

$$\lim_{\sigma_1 \to 0} \hat{\boldsymbol{c}} = z_1,$$

which coincides with (47) in the conventional method. On the other hand, as the noise level in z_2 becomes small relative to that in z_1 ($\sigma_2 \to 0$ with non-vanishing σ_1), the estimate reduces to:

$$\lim_{\sigma_2 \to 0} \hat{\boldsymbol{c}} = z_2/\alpha,$$

which corresponds to the case of estimating $\boldsymbol{c}$ solely from z_2.
More precisely,

$$\begin{aligned} \lim \hat{\boldsymbol{c}} = z_1 \quad \text{as } \sigma_1/\sigma_2 \to 0; \\ \lim \hat{\boldsymbol{c}} = z_2/\alpha \quad \text{as } \sigma_1/\sigma_2 \to \infty. \end{aligned} \tag{89}$$

- It is easy to check the estimator in (88) is unbiased with variance

$$\mathsf{Var}\{\hat{\boldsymbol{c}}\} = \frac{\sigma_1^2\sigma_2^2}{\alpha^2\sigma_1^2 + \sigma_2^2} I = \frac{\sigma_1^2}{1 + \alpha^2 \frac{\sigma_1^2}{\sigma_2^2}} I.$$

  It immediately follows that this quantity is upper-bounded by the covariance $\sigma_1^2 I$ of the estimator for $\boldsymbol{c}$ (47) resulting from conventional methods.

Now we can plug in the expression of $\hat{\boldsymbol{c}}$ in (88) and (87) reduces to a minimization problem over α only:

$$\begin{aligned} \hat{\alpha} &= \arg\min_{\alpha} \Psi(\alpha, \boldsymbol{z}) \\ &= \arg\min_{\alpha} \frac{1}{\sigma_1^2} \left\| z_1 - \frac{1}{\alpha^2\sigma_1^2 + \sigma_2^2}(\sigma_2^2 z_1 + \alpha\sigma_1^2 z_2) \right\|_2^2 + \frac{1}{\sigma_2^2} \left\| z_2 - \frac{\alpha}{\alpha^2\sigma_1^2 + \sigma_2^2}(\sigma_2^2 z_1 + \alpha\sigma_1^2 z_2) \right\|_2^2 \\ &= \arg\min_{\alpha} \frac{1}{\alpha^2\sigma_1^2 + \sigma_2^2} \|\alpha z_1 - z_2\|_2^2 \end{aligned} \tag{90}$$

This function Ψ is nonlinear in α. Note that $\Psi \geq 0$. In the case of noise-free observation $\boldsymbol{z} = \bar{\boldsymbol{z}}$, $\bar{\alpha}$ achieves the zero value and is the global minimizer (we will justify this more precisely later). Therefore, we can utilize the techniques for M-estimate as before, and analyze the behavior of $\hat{\alpha}$ in the neighborhood $\hat{\alpha}(\bar{\boldsymbol{z}}) = \bar{\alpha}$.

Let $\hat{\alpha}$ be the minimizer of the function $\Psi(\alpha, \boldsymbol{z})$, then it is true that

$$\frac{\partial}{\partial \alpha}\Psi(\alpha, \boldsymbol{z}) = \frac{\partial}{\partial \alpha} \frac{1}{\alpha^2\sigma_1^2 + \sigma_2^2} \left\| \begin{bmatrix} \alpha I & -I \end{bmatrix} \boldsymbol{z} \right\|_2 = 0 \quad \forall \boldsymbol{z}.$$

$$\begin{aligned} \frac{\partial}{\partial \alpha}\Psi(\alpha, \boldsymbol{z}) &= \frac{1}{(\alpha^2\sigma_1^2 + \sigma_2^2)^2}(\alpha z_1 - z_2)^T[2z_1(\alpha^2\sigma_1^2 + \sigma_2^2) - 2\alpha\sigma_1^2(\alpha z_1 - z_2)] \\ &= \frac{2}{(\alpha^2\sigma_1^2 + \sigma_2^2)^2} \boldsymbol{z}^T \begin{bmatrix} \alpha I \\ -I \end{bmatrix} \begin{bmatrix} \sigma_2^2 I & \alpha\sigma_1^2 \end{bmatrix} \boldsymbol{z}. \end{aligned} \tag{91}$$

Let $Q \triangleq \begin{bmatrix} \alpha I \\ -I \end{bmatrix} \begin{bmatrix} \sigma_2^2 I & \alpha\sigma_1^2 \end{bmatrix} = \begin{bmatrix} \alpha\sigma_2^2 I & \alpha^2\sigma_1^2 I \\ -\sigma_2^2 I & -\alpha\sigma_1^2 I \end{bmatrix}$, then the derivative of $\frac{\partial}{\partial \alpha}\Psi$ with respect to $\boldsymbol{z}$ is given by:

$$\begin{aligned} \frac{\partial^2}{\partial \alpha \partial \boldsymbol{z}}\Psi &= \frac{2}{(\alpha^2\sigma_1^2 + \sigma_2^2)^2} \boldsymbol{z}^T (Q + Q^T) \\ &= \frac{2}{(\alpha^2\sigma_1^2 + \sigma_2^2)^2} \boldsymbol{z}^T \begin{bmatrix} 2\alpha\sigma_2^2 I & (\alpha^2\sigma_1^2 - \sigma_2^2) I \\ (\alpha^2\sigma_1^2 - \sigma_2^2) I & -2\alpha\sigma_1^2 I \end{bmatrix}. \end{aligned} \tag{92}$$

Evaluating (92) at $\boldsymbol{z} = \bar{\boldsymbol{z}}$ and $\alpha = \bar{\alpha}$ yields:

$$\frac{\partial^2}{\partial \alpha \partial \boldsymbol{z}}\Psi(\bar{\alpha}, \bar{\boldsymbol{z}}) = \frac{2}{\bar{\alpha}^2\sigma_1^2 + \sigma_2^2} \bar{\boldsymbol{c}}^T \begin{bmatrix} \bar{\alpha} I & -I \end{bmatrix} \tag{93}$$

Now we compute the derivative of $\frac{\partial}{\partial \alpha}\Psi$ with respect to α and evaluate at the minimizer $\hat{\alpha} = \bar{\alpha}$ with $\boldsymbol{z} = \bar{\boldsymbol{z}}$:

$$\begin{aligned}
\frac{\partial^2}{\partial \alpha^2}\Psi &= 2\frac{\partial}{\partial \alpha}\left\{\frac{(\alpha z_1 - z_2)^T(\sigma_2^2 z_1 + \alpha\sigma_1^2 z_2)}{(\alpha^2\sigma_1^2 + \sigma_2^2)^2}\right\} \qquad (94)\\
&= 2\left\{-2\frac{2\alpha\sigma_1^2}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}(\alpha z_1 - z_2)^T(\sigma_2^2 z_1 + \alpha\sigma_1^2 z_2) + \frac{1}{(\alpha^2\sigma_1^2+\sigma_2^2)^2}[z_1^T(\sigma_2^2 z_1 + \alpha\sigma_1^2 z_2) + (\alpha z_1 - z_2)^T\sigma_1^2 z_2]\right\}
\end{aligned}$$

This is a convenient form to be evaluated at $\boldsymbol{z} = \bar{\boldsymbol{z}}$, and we obtain:

$$\frac{\partial^2}{\partial \alpha^2}\Psi(\bar{\boldsymbol{z}}) = \frac{2}{\bar{\alpha}^2\sigma_1^2 + \sigma_2^2}\left\|\bar{\boldsymbol{c}}\right\|_2^2. \qquad (95)$$

To prepare for future use, we simplify the general form of (95) into:

$$\begin{aligned}
\frac{\partial^2}{\partial \alpha^2}\Psi &= \frac{2}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}\left\{(-3\alpha^2\sigma_1^2+\sigma_2^2)\sigma_2^2\left\|z_1\right\|_2^2 + 2(3\sigma_2^2 - \alpha^2\sigma^2)\alpha\sigma_1^2 z_1^T z_2 + (3\alpha^2\sigma_1^2 - \sigma_2^2)\sigma_1^2\left\|z_2\right\|_2^2\right\}\\
&= \frac{2}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}\boldsymbol{z}^T\left[\begin{array}{cc} (-3\alpha^2\sigma_1^2+\sigma_2^2)\sigma_2^2 I & (3\sigma_2^2-\alpha^2\sigma^2)\alpha\sigma_1^2 I \\ (3\sigma_2^2-\alpha^2\sigma^2)\alpha\sigma_1^2 I & (3\alpha^2\sigma_1^2-\sigma_2^2)\sigma_1^2 I \end{array}\right]\boldsymbol{z}. \qquad (96)
\end{aligned}$$

Estimating $\frac{\partial}{\partial \boldsymbol{z}}\alpha$ yields:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{z}}\hat{\alpha}|_{\bar{\boldsymbol{z}},\bar{\alpha}} &= -\frac{\partial^2}{\partial \alpha^2}\Psi^{-1}\frac{\partial^2}{\partial\alpha\partial \boldsymbol{z}}\Psi\\
&= -\frac{1}{\left\|\bar{\boldsymbol{c}}\right\|_2^2}\bar{\boldsymbol{c}}^T\left[\begin{array}{cc}\bar{\alpha}I & -I\end{array}\right]. \qquad (97)
\end{aligned}$$

The covariance evaluated at $(\bar{\alpha}, \bar{\boldsymbol{z}})$ is

$$\begin{aligned}
\mathsf{Cov}\{\hat{\alpha}\}\,|_{(\bar{\boldsymbol{z}},\bar{\alpha})} &= \frac{\partial}{\partial \boldsymbol{z}}\alpha(\bar{\boldsymbol{z}})\,\mathsf{Cov}\{\boldsymbol{z}\}\,\frac{\partial}{\partial \boldsymbol{z}}\alpha(\bar{\boldsymbol{z}})\\
&= \left\|\bar{\boldsymbol{c}}\right\|_2^{-4}\bar{\boldsymbol{c}}^T\left[\begin{array}{cc}\bar{\alpha}I & -I\end{array}\right]\left[\begin{array}{cc}\sigma_1^2 I & 0\\ 0 & \sigma_2^2 I\end{array}\right]\left[\begin{array}{c}\bar{\alpha}I\\ -I\end{array}\right]\bar{\boldsymbol{c}}\\
&= \left\|\bar{\boldsymbol{c}}\right\|_2^{-2}(\bar{\alpha}^2\sigma_1^2 + \sigma_2^2) \qquad (98)
\end{aligned}$$

.4 Lower Bound for Covariance From Cramér-Rao Bound

The negative log-likelihood is given as the objective function in (87). It is straight-forward to compute the sub-matrices for the Fisher-Information Matrix.

$$\frac{\partial}{\partial\alpha}\Lambda = -\frac{1}{\sigma_2^2}(\alpha\boldsymbol{c} - z_2)^T\boldsymbol{c};$$

$$\frac{\partial^2}{\partial\alpha^2}\Lambda = -\frac{1}{\sigma_2^2}\boldsymbol{c}^T\boldsymbol{c}.$$

$$\frac{\partial^2}{\partial\alpha\partial\boldsymbol{c}}\Lambda = \frac{-1}{\sigma_2^2}(2\alpha\boldsymbol{c}^T - z_2^T),$$

resulting in

$$E[\frac{\partial^2}{\partial\alpha^2}\boldsymbol{c}] = \frac{-1}{\sigma_2^2}\alpha\boldsymbol{c}^T.$$

The Fisher-information matrix (FIM) is thus given by:

$$FIM = \frac{1}{\sigma_2^2}\left[\begin{array}{cc}\boldsymbol{c}^T\boldsymbol{c} & \alpha\boldsymbol{c}^T\\ \alpha\boldsymbol{c} & (\alpha^2 + \frac{\sigma_2^2}{\sigma_1^2})I\end{array}\right].$$

Invoking block-matrix inversion, we obtain:

$$\begin{aligned}\mathsf{Cov}\{\hat{\alpha}\} &\geq \sigma^2[\boldsymbol{c}^T\boldsymbol{c} - \alpha\boldsymbol{c}^T(\alpha^2 + \frac{\sigma_2^2}{\sigma_1^2})^{-1}\alpha\boldsymbol{c}] \\ &= \|\boldsymbol{c}\|_2^{-2}(\alpha^2\sigma_1^2 + \sigma_2^2). \end{aligned} \tag{99}$$

Since the ML estimator is known to be asymptotically unbiased, the coincidence between (98) and (99) justifies the well-known fact that the ML estimator is asymptotically efficient (thus is asymptotically a uniformly minimal variance and unbiased estimator (UMVUE)).

.5 Approximate Bias of the ML Estimator

Not withstanding the value of asymptotic analysis for the ML estimator, it is often of great interest to analyze the bias and variance before the the estimator enters the asymptotic zone. Hereafter, we focus on deriving analytical approximation for the bias of the ML estimator. As in the covariance analysis previously, we assume the estimate is over continuous parameter's α and is computed by "completely" maximizing the objective function (likelihood in this case) without "stopping rules" that terminates the iterations before the maximum is reached. We derive the approximation using implicit function theorem, the Taylor expansion (with different orders of approximation accuracy), and the chain rule.

The objective function Ψ in (90) implicitly defines the M-estimate $\hat{\alpha}$ as a function of $\boldsymbol{z}$. Yet the absence of an explicit analytical expression of the form $\hat{\alpha} = h(\boldsymbol{z})$ (as the one in (49)) makes it difficult to study the mean of $\hat{\alpha}$ directly. As in the previous section, we apply Taylor expansion, chain rules and implicit function theorem to estimate the bias with the first and second order approximation given by:

$$E[\hat{\alpha}] \approx h(\check{\boldsymbol{z}}) + E\left\{\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\}. \tag{100}$$

$$E[\hat{\alpha}] \approx h(\check{\boldsymbol{z}}) + E\left\{\nabla_{\boldsymbol{z}} h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}}) + \frac{1}{2}(\boldsymbol{z} - \check{\boldsymbol{z}})^T\nabla_{\boldsymbol{z}}^2 h(\check{\boldsymbol{z}})(\boldsymbol{z} - \check{\boldsymbol{z}})\right\}. \tag{101}$$

We now determine the point of expansion $\check{z}$ and the approximation for first (linear) and second order (Hessian) coefficients $\nabla_{\boldsymbol{z}} h$, $\nabla_{\boldsymbol{z}}^2 h$. To obtain the best choice for $\check{\alpha}$

$$\check{\alpha} = \arg\min_{\alpha} E[\Psi(\alpha, \boldsymbol{z})], \tag{102}$$

where $\check{\alpha}$ and $\check{\boldsymbol{z}}$ in the Taylor expansions are related by $\check{\alpha} = h(\check{\boldsymbol{z}})$, we compute $E[\Psi(\alpha, \boldsymbol{z})]$ as follows:

$$E[\Psi(\alpha, \boldsymbol{z})] = \frac{1}{\alpha^2\sigma_1^2 + \sigma_2^2}\sum_{i=1}^{n}(\alpha z_1(i) - z_2(i))^2.$$

For each index i,

$$\begin{aligned} E[(\alpha z_1(i) - z_2(i))^2] &= E[\alpha^2 z_1(i)^2 - 2\alpha z_1(i)z_2(i) + z_2(i)^2] \\ &= \alpha^2(\bar{c}_i^2 + \sigma_1^2) - 2\alpha\bar{\alpha}\bar{c}_i^2 + \bar{\alpha}^2\bar{c}_i^2 + \sigma_2^2 \\ &= (\alpha^2 - 2\alpha\bar{\alpha} + \bar{\alpha}^2)\bar{c}_i^2 + (\alpha^2\sigma_1^2 + \sigma_2^2), \end{aligned} \tag{103}$$

where $\bar{c}_i$ and $\bar{\alpha}$ are the underlying "true" parameter values.

Substituting (103) yields:

$$E[\Psi(\alpha, \boldsymbol{z})] = \frac{1}{\alpha^2\sigma_1^2 + \sigma_2^2}(\alpha - \bar{\alpha})^2\|\bar{\boldsymbol{c}}\|_2^2 + n. \tag{104}$$

Even though $E[\Psi(\alpha, \boldsymbol{z})]$ is nonlinear in α, its global minimizer is immediately observed as $\alpha = \bar{\alpha}$, because $E[\Psi(\bar{\alpha}, \boldsymbol{z})] = n$ achieves the lower bound for $E[\Psi(\alpha, \boldsymbol{z})]$ as a function of α. Thus we have found the proper point to expand around $\check{\alpha} = \bar{\alpha}$.

Note that when noise free data is observed, *i.e.*, $\boldsymbol{z} = \bar{\boldsymbol{z}}$, the minimizer $\hat{\alpha}$ in (90) is obtained as:

$$\begin{aligned}\hat{\alpha}(\bar{\boldsymbol{z}}) &= \arg\min_{\alpha} \frac{1}{\alpha^2\sigma_1^2+\sigma_2^2}\left\|\alpha\bar{z}_1-\bar{z}_2\right\|_2^2 \\ &= \arg\min_{\alpha} \frac{1}{\alpha^2\sigma_1^2+\sigma_2^2}\left\|\alpha\bar{\boldsymbol{c}}-\bar{\alpha}\bar{\boldsymbol{c}}\right\|_2^2 \\ &= \arg\min_{\alpha} \frac{(\alpha-\bar{\alpha})^2\left\|\boldsymbol{c}\right\|_2^2}{\alpha^2\sigma_1^2+\sigma_2^2}. \end{aligned} \tag{105}$$

Note this function is nonnegative, its global minimizer is obtained at $\alpha = \bar{\alpha}$, *i.e.*, $h(\bar{\boldsymbol{z}}) = \bar{\alpha} = \check{\alpha}$. This indicates that $\check{\boldsymbol{z}} = \bar{\boldsymbol{z}}$ is the proper choice to expand h around, without requiring to know the precise value of $\bar{\alpha}$.

In this case, the bias analysis with first-order Taylor expansion as in (100) is simple by noting that $(\boldsymbol{z}-\bar{\boldsymbol{z}}) \sim \mathcal{N}(0, \begin{bmatrix} \sigma_1^2 I & \\ & \sigma_2^2 I \end{bmatrix})$, so that

$$\begin{aligned} E[\hat{\alpha}] &= h(\bar{\boldsymbol{z}}) + E\left\{\nabla_{\boldsymbol{z}} h(\bar{\boldsymbol{z}})(\boldsymbol{z}-\bar{\boldsymbol{z}})\right\} \\ &= \bar{\alpha} \end{aligned} \tag{106}$$

This states that the estimator is unbiased if we approximate its first moment up to first order dependence on the data.

The first order expansion is usually sufficient in practice and has been extensively used. However, there are situations where (100) may be inadequate. We next derive a mean approximation based on the second-order Taylor expansion (101) which is expected to be more accurate, but also computationally more intensive.

The first two (0th and 1st order) terms in (101) are (100), so it suffices to study the Hessian $\nabla_{\boldsymbol{z}}^2$.

For scalar α, we follow the simplified expression in [6] to obtain the Hessian of $h(\boldsymbol{z})$ as:

$$\nabla_{\boldsymbol{z}}^2 h = [-\frac{\partial^2}{\partial\alpha^2}\Psi]^{-1}\left\{\frac{\partial^3}{\partial\alpha^3}\Psi\nabla_{\boldsymbol{z}}h^T\nabla_{\boldsymbol{z}}h + \frac{\partial^3}{\partial\alpha^2\partial\boldsymbol{z}}\Psi^T\nabla_{\boldsymbol{z}}h + \nabla_{\boldsymbol{z}}h^T\frac{\partial^3}{\partial\alpha^2\partial\boldsymbol{z}}\Psi + \frac{\partial}{\partial\alpha}\nabla_{\boldsymbol{z}}^2\Psi\right\}. \tag{107}$$

Some of the key gradients are already available: $\nabla_{\boldsymbol{z}}h$ is given in (97) as well as $\frac{\partial^2}{\partial\alpha^2}\Psi$ in (95) (before evaluation) and $\frac{\partial^2}{\partial\alpha\partial\boldsymbol{z}}\Psi$ in (92). We still need to compute $\frac{\partial^3}{\partial\alpha^3}\Psi(\bar{\alpha},\bar{\boldsymbol{z}})$, $\frac{\partial^3}{\partial\alpha^2}\partial\boldsymbol{z}\Psi(\bar{\alpha},\bar{\boldsymbol{z}})$ and $\frac{\partial}{\partial\alpha}\nabla_{\boldsymbol{z}}^2\Psi$.

Evaluating (95) at $(\bar{\alpha},\bar{\boldsymbol{z}})$ yields:

$$\frac{\partial^2}{\partial\alpha^2}\Psi(\bar{\alpha},\bar{\boldsymbol{z}}) = \frac{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}{\bar{\alpha}^2\sigma_1^2+\sigma_2^2}.$$

Taking derivative of (96) with respect to $\boldsymbol{z}$ yields:

$$\frac{\partial^3}{\partial\alpha^2\partial\boldsymbol{z}}\Psi = \frac{4}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}\boldsymbol{z}^T\begin{bmatrix} (-3\alpha^2\sigma_1^2+\sigma_2^2)\sigma_2^2 I & (3\sigma_2^2-\alpha^2\sigma^2)\alpha\sigma_1^2 I \\ (3\sigma_2^2-\alpha^2\sigma^2)\alpha\sigma_1^2 I & (3\alpha^2\sigma_1^2-\sigma_2^2)\sigma_1^2 I \end{bmatrix} \tag{108}$$

Evaluating (108) at $(\bar{\alpha},\bar{\boldsymbol{z}})$ yields:

$$\frac{\partial^3}{\partial\alpha^2\partial\boldsymbol{z}}\Psi(\bar{\alpha},\bar{\boldsymbol{z}}) = \frac{4}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^3}\bar{\boldsymbol{c}}^T\begin{bmatrix} (\sigma_2^4-\bar{\alpha}^4\sigma_1^4)I & 2\bar{\alpha}\sigma_1^2(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)I \end{bmatrix} \tag{109}$$

Taking derivative of (95) with respect to α yields:

$$\begin{aligned}\frac{\partial^3}{\partial\alpha^3}\Psi \quad = \quad & \frac{-12\alpha\sigma_1^2}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}z_1^T(\sigma_2^2 z_1+\alpha\sigma_1^2 z_2) + \frac{2}{(\alpha^2\sigma_1^2+\sigma_2^2)^3}[-4\alpha\sigma_1^2 z_1^T(\sigma_2^2 z_1+\alpha\sigma_1^2 z_2)+\ldots \\ & +2\alpha\sigma_1^2 z_1^T(\sigma_2^2 z_1+\alpha\sigma_1^2 z_2)+2(\alpha^2\sigma_1^2+\sigma_2^2)\sigma_1^2 z_1^T z_2]. \end{aligned} \tag{110}$$

Evaluating (110) at (110) at $(\bar{\alpha},\bar{\boldsymbol{z}})$ yields:

$$\frac{\partial^3}{\partial\alpha^3}\Psi(\bar{\alpha},\bar{\boldsymbol{z}}) = \frac{-12\bar{\alpha}\sigma_1^2\left\|\bar{\boldsymbol{c}}\right\|_2^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)}. \tag{111}$$

The term $\frac{\partial}{\partial\alpha}\nabla^2_{\boldsymbol{z}}\Psi$ is obtained by taking derivative of $\frac{\partial^2}{\partial\alpha\partial\boldsymbol{z}}\Psi$ in (92) with respect to $\boldsymbol{z}$ as:

$$\frac{\partial}{\partial\alpha}\nabla^2_{\boldsymbol{z}}\Psi = 2(\frac{1}{\alpha^2\sigma_1^2+\sigma_2^2})^2 \begin{bmatrix} 2\alpha\sigma_2^2 I & (\alpha^2\sigma_1^2-\sigma_2^2)I \\ (\alpha\sigma_1^2-\sigma_2^2)I & -2\alpha\sigma_1^2 I \end{bmatrix} \tag{112}$$

Evaluating at $\bar{\alpha}$ yields:

$$\frac{\partial}{\partial\alpha}\nabla^2_{\boldsymbol{z}}\Psi(\bar{\alpha}) = 2(\frac{1}{\bar{\alpha}^2\sigma_1^2+\sigma_2^2})^2 \begin{bmatrix} 2\bar{\alpha}\sigma_2^2 I & (\bar{\alpha}^2\sigma_1^2-\sigma_2^2)I \\ (\bar{\alpha}\sigma_1^2-\sigma_2^2)I & -2\bar{\alpha}\sigma_1^2 I \end{bmatrix}. \tag{113}$$

Substituting the expressions of all components into the right-hand-side of (107) yields:

$$\begin{aligned} \nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}}) \quad = \quad & -\frac{\bar{\alpha}^2\sigma_1^2+\sigma_2^2}{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\Big\{\frac{-12\bar{\alpha}\sigma_1^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\begin{bmatrix}\bar{\alpha}I \\ -I\end{bmatrix}\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T\begin{bmatrix}\bar{\alpha}I & -I\end{bmatrix} + \ldots \\ & -\frac{4}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^3\left\|\bar{\boldsymbol{c}}\right\|_2^2}\begin{bmatrix}(\sigma_2^4-\bar{\alpha}^4\sigma_1^4)I \\ 2\bar{\alpha}\sigma_1^2(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)I\end{bmatrix}\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T\begin{bmatrix}\bar{\alpha}I & -I\end{bmatrix} + \ldots \\ & -\frac{4}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^3\left\|\bar{\boldsymbol{c}}\right\|_2^2}\begin{bmatrix}\bar{\alpha}I \\ -I\end{bmatrix}\bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T\begin{bmatrix}(\sigma_2^4-\bar{\alpha}^4\sigma_1^4)I & 2\bar{\alpha}\sigma_1^2(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)I\end{bmatrix} + \ldots \\ & +2(\frac{1}{\bar{\alpha}^2\sigma_1^2+\sigma_2^2})^2\begin{bmatrix}2\bar{\alpha}\sigma_2^2 I & (\bar{\alpha}^2\sigma_1^2-\sigma_2^2)I \\ (\bar{\alpha}\sigma_1^2-\sigma_2^2)I & -2\bar{\alpha}\sigma_1^2 I\end{bmatrix}\Big\}. \end{aligned} \tag{114}$$

The second order term in (101) depends on the Hessian $\nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}})$ via $(\boldsymbol{z}-\bar{\boldsymbol{z}})^T\nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}})(\boldsymbol{z}-\bar{z})$ since $\check{\boldsymbol{z}} = \bar{z}$, where $\boldsymbol{z}-\bar{\boldsymbol{z}}$ are exactly the noise component $\epsilon \sim \mathcal{N}(0, \begin{bmatrix}\sigma_1^2 I & \\ & \sigma_2^2 I\end{bmatrix})$. Because the elements of ϵ are mutually independent, $E\left\{(\boldsymbol{z}-\bar{\boldsymbol{z}})^T\nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}})(\boldsymbol{z}-\bar{z})\right\}$ only depends on the diagonal elements of the Hessian $\nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}})$.

When a component is located in the z_1 portion of $\boldsymbol{z}$, the noise component $\epsilon(i) \sim \mathcal{N}(0,\sigma_1^2)$, and taking the corresponding element in the Hessian, we obtain:

$$\frac{\partial^2}{\partial z_1(i)^2}h(\bar{\boldsymbol{z}}) = -\frac{1}{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\left\{\frac{-12\bar{\alpha}^3\sigma_1^2c_i^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)\left\|\bar{\boldsymbol{c}}\right\|_2^2} - \frac{8\bar{\alpha}(\sigma_2^4-\bar{\alpha}^4\sigma_1^4)c_i^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^2\left\|\bar{\boldsymbol{c}}\right\|_2^2} + 4\bar{\alpha}\sigma_2^2\right\}. \tag{115}$$

Similarly,

$$\frac{\partial^2}{\partial z_2(i)^2}h(\bar{\boldsymbol{z}}) = -\frac{1}{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\left\{\frac{-12\bar{\alpha}\sigma_1^2c_i^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)\left\|\bar{\boldsymbol{c}}\right\|_2^2} + \frac{16\bar{\alpha}\sigma_1^2c_i^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)\left\|\bar{\boldsymbol{c}}\right\|_2^2} - 4\bar{\alpha}\sigma_1^2\right\}. \tag{116}$$

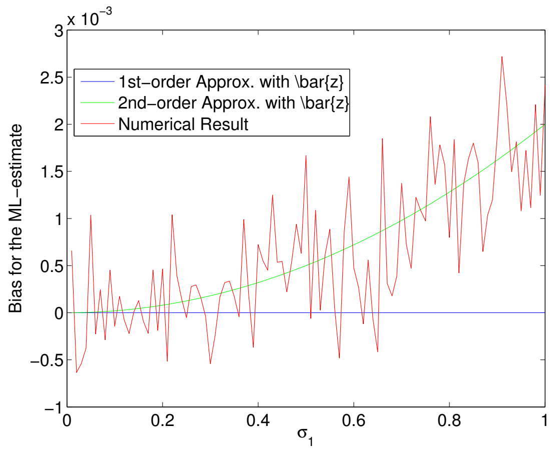
Combining the above to obtain:

$$\begin{aligned} E[\epsilon^T\nabla^2_{\boldsymbol{z}}h(\bar{\boldsymbol{z}})\epsilon] \quad &= \quad \sigma_1^2\sum_{i=1}^n\frac{\partial^2}{\partial z_1(i)^2}h(\bar{\boldsymbol{z}}) + \sigma_2^2\sum_{i=1}^n\frac{\partial^2}{\partial z_2(i)^2}h(\bar{\boldsymbol{z}}) \\ &= \quad -\frac{1}{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\left\{\frac{-12\bar{\alpha}^3\sigma_1^4}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)} - \frac{8\bar{\alpha}(\sigma_2^4-\bar{\alpha}^4\sigma_1^4)\sigma_1^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)^2}\right\} - \frac{1}{2\left\|\bar{\boldsymbol{c}}\right\|_2^2}\left\{\frac{-12\bar{\alpha}\sigma_1^2\sigma_2^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)} + \frac{16\bar{\alpha}\sigma_1^2\sigma_2^2}{(\bar{\alpha}^2\sigma_1^2+\sigma_2^2)}\right\} \\ &= \quad \frac{\bar{\alpha}\sigma_1^2}{\left\|\bar{\boldsymbol{c}}\right\|_2^2}. \end{aligned} \tag{117}$$

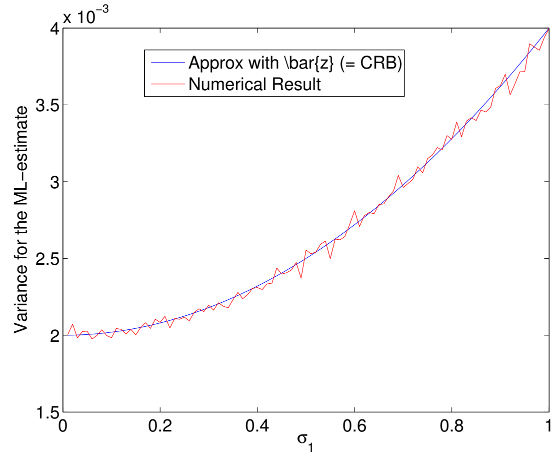
The second order approximation of the estimator yields:

$$E[\hat{\alpha}]/\bar{\alpha} = 1 + \frac{\sigma_1^2}{\left\|\bar{\boldsymbol{c}}\right\|_2^2},$$

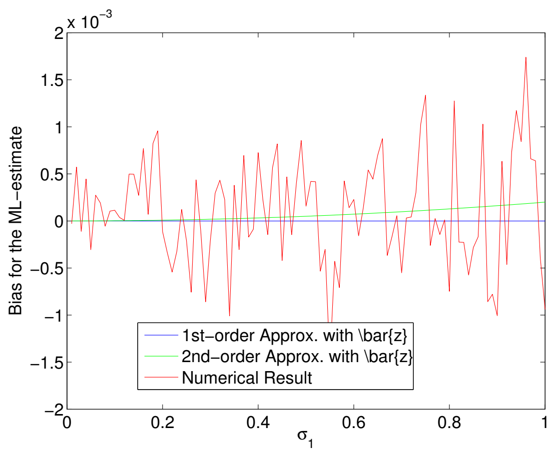
which indicates a bias toward positive magnitude. Comparing with the bias analysis for the conventional M-estimate, the bias of the ML estimate is independent of the data length n, which indicates that even though both estimators are asymptotically unbiased, they approach the asymptotic region with different rate (roughly $1:n$).



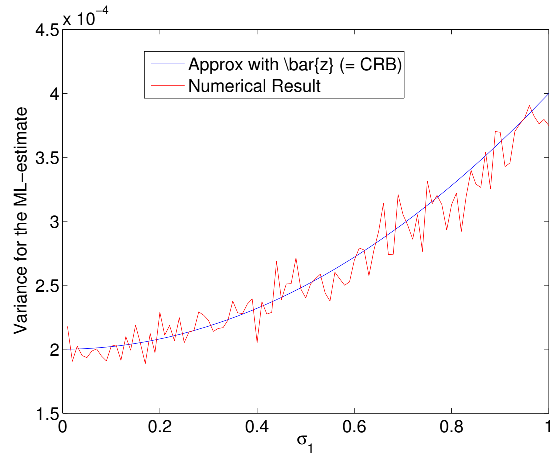
(a1) bias approximation $n = 5$



(a2) variance approximation $n = 5$



(b1) bias approximation $n = 50$



(b2) variance approximation $n = 50$

Fig. 4. Bias and variance approximation of ML-estimate obtained from expansion about $(\check{\alpha}, \check{\boldsymbol{z}})$.

VII. Acknowledgment

The authors would like to thank Prof. Alfred O. Hero for motivating this work.

References

[1] H. L. Van Trees. *Detection, estimation, and modulation theory*. Wiley, New York, 1968.
[2] A. O. Hero and J. A. Fessler. A recursive algorithm for computing Cramer-Rao-type bounds on estimator covariance. *IEEE Trans. Info. Theory*, 40(4):1205–10, July 1994.
[3] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Univ. Press, 2 edition, 1989.
[4] M. Moeneclaey. On the true and the modified Cramer-Rao bounds for the estimation of a scaler parameter in the presence of nuisance parameters. *IEEE Trans. Commun.*, 46(11):1536–44, Nov. 1998.
[5] A. D'Andrea, U. Mengalia, and R. Reggiannini. The modified Cramer-Rao bound and its application to synchronization problems. *IEEE Trans. Commun.*, 42(234):1391–99, Feb-Apr. 1994.
[6] J. A. Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Trans. Im. Proc.*, 5(3):493–506, March 1996.
[7] C. R. Rao. *Linear statistical inference and its applications*. Wiley, New York, 1973.
[8] R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, New York, 1980.
[9] P. J. Huber. *Robust statistics*. Wiley, New York, 1981.