Bayesian Spatial Field Reconstruction with Unknown Distortions in Sensor Networks

Qikun Xiang\textsuperscript{1}, Ido Nevat\textsuperscript{2} and Gareth W. Peters\textsuperscript{3},
\textsuperscript{1} School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
\textsuperscript{2} TUMCREATE, Singapore
\textsuperscript{3} Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK

Abstract

Spatial regression of random fields based on unreliable sensing information is proposed in this paper. One major concern in such applications is that since it is not known \textit{a-priori} what the accuracy of the collected data from each sensor is, the performance can be negatively affected if the collected information is not fused appropriately. For example, the data collector may measure the phenomenon inappropriately, or alternatively, the sensors could be out of calibration, thus introducing random gain and bias to the measurement process. Such readings would be systematically distorted, leading to incorrect estimation of the spatial field. To combat this detrimental effect, we develop a robust version of the spatial field model based on a mixture of \textit{Gaussian process} experts. We then develop two different approaches for Bayesian spatial field reconstruction: the first algorithm is the Spatial Best Linear Unbiased Estimator (S-BLUE), in which one considers a quadratic loss functions and restricts the estimator to the linear family of transformations; the second algorithm is based on empirical Bayes, which utilises a two-stage estimation procedure to produce accurate predictive inference in the presence of “misbehaving” sensors. We present extensive simulation results of both synthetic and real-world scenarios and draw useful conclusions regarding the performance of each of the algorithms.

\textbf{Keywords:} Sensor Networks, Gaussian Process, Spatial Linear Unbiased Estimator (S-BLUE), Empirical Bayes, Cross Entropy method (CEM), Iterated Conditional Modes (ICM)

I. \textbf{Introduction}

In recent years, Wireless Sensor Networks (WSNs) have attracted considerable attention due to their applications in environment monitoring [1], [2], [3], [4], forecasting [5], surveillance [6],
event detection [7] and tracking [8]. For example, the United States Environmental Protection Agency (EPA) proposed to promote the use of sensor networks for air quality monitoring [9]. In this paper, we focus on environmental monitoring applications in which a WSN consists of a collection of spatially distributed sensor nodes with limited energy and communication bandwidth. The sensors make observations of spatial physical phenomena (e.g. the concentration of air pollutants such as carbon monoxide and ozone, temperature, humidity, etc. [1], [2]) and communicate the observations to a Fusion Centre (FC) [10], [11]. The FC then reconstructs the spatial phenomena from these observations at any spatial location of interest, based on which decisions and actions can be made.

As these low-cost sensor nodes are unreliable, assessing and guaranteeing the veracity, quality and reliability of collected data is crucial [12], [13], [4]. One threat to data reliability comes from people with malicious intent. They may physically compromise sensor nodes and launch false data injection attacks [14], [15] in which the sensor observations are maliciously altered to disrupt the operation of the WSN. Another threat is related to sensor calibration, as miscalibrated sensors can lead to biased conclusions in scientific data analysis [16]. From a calibrated sensor, the physical input can be recovered from the sensor output. Sensor calibration involves estimating the calibration parameters of individual sensors under a calibration model. Among the classical calibration models, the gain-offset response model is widely-used [17], [1], [2], [18]. Traditionally, sensors are calibrated in a controlled environment where the physical input is known and then their performance in a given calibration range is tested and verified over certain operating ranges of the environment, before such WSN are deployed. This is infeasible for large-scale sensor networks due to prohibitive cost as well as inhomogeneity in deployment schedules. Thus, the calibration has to be done through the so-called blind or self-calibration techniques [19], [20], [21]. In addition, the reliability of sensor can deteriorate over time [3], making it very challenging to guarantee the quality of information even for an a priori calibrated network.

Much effort in research has been dedicated to detecting, analysing and mitigating data reliability threats in WSNs [22], many of which focus on threats related to distributed binary detection [7], [14], [15], [23], [24], distributed discrete M-ary inference [25] and multi-agent systems [5]. Regarding data reliability threats in environment monitoring systems, an estimation procedure has been proposed in [26] to detect and exclude malicious sensing agents, while accurately performing spatial field reconstruction. For mobile sensing networks, many of the blind sensor
calibration techniques are based on *rendezvous* events, which corresponds to occasions where two sensors are in close spatial and temporal vicinity [17], [1], [2], [3]. Dorffer et al. [18] proposed a sensor self-calibration technique based on non-negative matrix factorization. Some of the recent works [27], [28], [29], [30], [11] utilize the spatial-temporal correlations to reconstruct spatial physical phenomena at all locations. Wang et al. [31] proposed to use re-sampling and Bayesian techniques to assess the quality of data in sparse mobile crowd-sensing. They also studied the optimal task allocation problem in such applications.

The main goal of this paper is to develop statistical procedures that reconstruct spatial fields using observations from sensors with possibly unknown calibration parameters. We refer to such observations as *distorted observations*. We use Gaussian processes as the probabilistic model for spatial phenomena, and the sensors are assumed to follow the *gain-offset* distortion model with multi-modal priors on the distortion parameters.

The main contributions are as follows:

1) We derive the posterior distribution and the posterior predictive distribution of the model, and show that the exact computation of Bayes estimators is intractable.

2) We develop the Spatial Best Linear Unbiased Estimator (S-BLUE) for the model, which is highly computationally efficient.

3) We develop a two-stage Bayesian inference algorithm that jointly infers the distortions of sensors and reconstructs the spatial field at all locations of interest. The algorithm estimates the distortion parameters in an empirical Bayes manner. We then solve the corresponding optimization problem via two efficient methods, the Cross-Entropy method (CEM) and the Iterated Conditional Mode (ICM) method.

4) We perform a combination of real data and synthetic data experiments to validate our model and estimation procedures. The real data studies involve application to a real temperature dataset from US EPA to show the real-world applicability of the model.

The remainder of the paper is organized as follows. We present our Bayesian sensor network model in Section II, which includes the prior distribution of the distortion parameters. In Section III, we derive the posterior distribution of the parameters as well as the posterior predictive distribution. Section IV introduces the S-BLUE and its properties. Section V introduces the approximation of the Bayes estimators via empirical Bayes, and shows that the maximization of the posterior distribution can be done through the CEM method or ICM. In Section VI and Section VII, we perform experiments using synthetic and real datasets. Finally, Section VIII
concludes the paper.

II. SENSOR NETWORK MODEL AND ASSUMPTIONS

We begin by presenting the statistical model for the spatial physical phenomena, followed by the system model.

A. Spatial Gaussian Random Fields Background

We model the physical phenomenon as spatially dependent continuous process with a spatial correlation structure. Such models have recently become popular due to their mathematical tractability and accuracy [26], [32], [33], [34], [35], [5]. The degree of the spatial correlation in the process increases with the decrease of the separation between two observing locations and can be accurately modelled as a Gaussian random field\(^1\) [29], [30], [36], [37], [38]. A Gaussian process (GP) defines a distribution over a space of functions and it is completely specified by the equivalent of sufficient statistics for such a process, and is formally defined as follows.

**Definition 1.** (Gaussian process [39],[40]): Let \(\mathcal{X} \subset \mathbb{R}^d\) be some bounded domain of a \(d\)-dimensional real valued vector space. Denote by \(f(x) : \mathcal{X} \mapsto \mathbb{R}\) a stochastic process parametrized by \(x \in \mathcal{X}\). Then, the random function \(f(x)\) is a Gaussian process if all its finite dimensional distributions are Gaussian, where for any \(m \in \mathbb{N}\), the random vectors \((f(x_1), \ldots, f(x_m))\) are normally distributed.

We can therefore interpret a GP as formally defined by the following class of random functions:

\[ \mathcal{F} := \{ f(\cdot) : \mathcal{X} \mapsto \mathbb{R} \text{ s.t. } f(\cdot) \sim GP(\mu(\cdot), \mathcal{C}(\cdot, \cdot)) \}, \]

where at each point the mean of the function is \(\mu(\cdot)\), and the spatial dependence between any two points is given by the covariance function (Mercer kernel) \(\mathcal{C}(\cdot, \cdot)\) (see detailed discussion in [39]).

B. Sensor Network System Model

We begin by presenting the system model followed by the prior distribution specifications:

\(^{1}\)We use Gaussian Process and Gaussian random field interchangeably.
A1. Consider a random real-valued spatial phenomenon \( f : \mathcal{X} \mapsto \mathbb{R} \) defined on the \( d \)-dimensional domain \( \mathcal{X} \subseteq \mathbb{R}^d \).

A2. Consider a sensor network with \( N \) sensors that sense and transmit data to a Fusion Centre (FC) over perfect communication channels. The spatial locations of the sensors, denoted \( (x_n)_{n=1:N} \) (\( x_n \in \mathcal{X}, n = 1, \ldots, N \)), are known at the FC.

A3. The sensor \( n \) transmit \( M_n \in \mathbb{N} \) observations to the FC. The observations \( (y_{n,m})_{m=1:M_n} \) are generated according to the following Acquisition+Distortion mechanism:

\[
\begin{align*}
\tilde{y}_{n,m} & = f(x_n) + \epsilon_{n,m} \quad \text{(Acquisition)} \\
y_{n,m} & = T(\tilde{y}_{n,m}; \psi_n) \quad \text{(Distortion)}
\end{align*}
\]

for \( m = 1, \ldots, M_n \), where \( f(x_n) \) is the realisation of the random field at location \( x_n \), \( \epsilon_{n,m} \) represents the additive random noise at the \( n \)-th sensor, and \( T : \mathbb{R} \mapsto \mathbb{R} \) is the distortion transformation function, parametrized by \( \psi_n \).

A4. The distortion transformation \( T \) has the following generic gain-offset form:

\[
T(u ; \psi_n = (a_n, b_n)) := a_n u + b_n,
\]

where \( a_n \in \mathbb{R}_+ \) and \( b_n \in \mathbb{R} \) represent the gain and offset of the \( n \)-th sensor, respectively. This gain-offset model has been widely used to describe sensor characteristics [17], [1], [2], [18].

A5. We assume there are \( K + 1 \) “categories” of possible distortion transformations. The auxiliary indicator random variable \( Z_n \) indicates the category to which each sensor’s parameters \( \psi_n \) belong. We denote by \( Z_n = 0 \), the default distortion transformation category, \( \psi_n = \psi^0 = (1, 0) \) i.e. no distortion \( (T(u, \psi^0) = u) \), whereas \( Z_n = k, k \in \{1, \ldots, K\} \) indicates that the sensor \( n \) belongs to the \( k \)-th non-default distortion transformation.

C. Prior Distribution Specifications

P1. The spatial random field, \( f \), is modelled as a Gaussian process (GP) \( F \) with a known mean function \( \mu : \mathcal{X} \mapsto \mathbb{R} \) and a known covariance function \( C : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \), that is,

\[
F \sim \mathcal{GP}(\mu(\cdot), C(\cdot, \cdot)).
\]

P2. The additive random noise, \( \epsilon_{n,m} \), follows a normal distribution with mean zero and a fixed known variance \( \varsigma^2 \),

\[
\epsilon_{n,m} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \varsigma^2),
\]

\[ (3) \]

\[ (4) \]
P3. We place a prior distribution on $Z_n$, denoted $\pi(Z_n)$, which is a categorical distribution given by

$$
\pi(Z_n = k) = q_k^{(n)}, \quad k = 0, \ldots, K,
$$

(5)

where $q_k^{(n)} \geq 0$ for $k = 0, \ldots, K$, and $\sum_{k=0}^{K} q_k^{(n)} = 1$.

P4. Each category of distortion characteristics has a distinct sub-population distribution, denoted $\pi_k$, which translates to the following. For $k = 0, \ldots, K$,

$$
(\psi_n | Z_n = k) \sim \pi_k,
$$

(6)

where $\pi_0$ is a degenerate distribution (or an atom) at $\psi^0$,

$$
\pi_0(\psi_n) = \delta_{\psi^0}.
$$

(7)

For $k = 1, \ldots, K$, assume that $\pi_k$ has density (and we slightly abuse the notation $\pi_k$ to also denote the density function). Thus, the prior density of $\psi_n$ (marginalizing over $Z_n$), denoted by $\pi(\psi_n)$ is a mixture with an atom at $\psi^0$, given by

$$
\pi(\psi_n) = q_0^{(n)} \delta_{\psi^0} + \sum_{k=1}^{K} q_k^{(n)} \pi_k(\psi_n).
$$

(8)

P5. We assume the independence among $(\psi_n)_{1:N}$, and denote them collectively as $\psi$. Let $\pi(\psi)$ denote the prior density function of $\psi$, which factorizes due to independence,

$$
\pi(\psi) = \prod_{n=1}^{N} \pi(\psi_n).
$$

(9)

The graphical structure of the proposed Bayesian model is shown in Figure 1 as a directed-acyclic-graph (DAG) using plate notations. This graphical illustration is helpful for visualizing the dependencies and conditional independence relations between parameters and random variables.

Our objective is to find an estimator $h(y)$ for $f_\ast := f(x_\ast)$, the spatial field at location $x_\ast$, based on observations $y := (y_{n,m})_{n=1:N, m=1:M_n}$.

III. POSTERIOR DENSITIES OF THE MODEL

In this section we derive the following quantities of interest, based on which the Bayesian estimators will be developed in Sections IV and V:

1) The posterior distribution of the model parameters $\psi$, given by $p(\psi|y)$ (Theorem 1).
2) The posterior predictive distribution $p(f_\ast|y)$ (Theorem 2).
Fig. 1: Directed acyclic graph (DAG) of the model using the plate notation. Shaded rectangles represent constants and observed covariates. White circles represent unobserved random variables. The shaded circle represents observed random variable. Arrows represent conditional dependence between two quantities.

The following theorem gives the posterior density function of $\psi$.

**Theorem 1.** Let $\mu := (\mu(x_n))_{1:N} \in \mathbb{R}^N$ be the expected values of the $F$ process at locations $(x_n)_{1:N}$, and let $C \in \mathbb{R}^{N \times N}$ be the covariance matrix, where $(C)_{ij} = C(x_i, x_j)$. For $n = 1, \ldots, N$, define $g_n := \sum_{m=1}^{M_n} y_{n,m}$, $s_n := \sum_{m=1}^{M_n} y_{n,m}^2$. Let $A := \text{diag}([a_n]_{1:N})$, $b := (b_n)_{1:N}$, $g := (g_n)_{1:N}$, $s := (s_n)_{1:N}$. Let $\tilde{g}_n := a_n^{-1}(M_n^{-1} g_n - b_n)$, $\tilde{g} := (\tilde{g}_n)_{1:N} = A^{-1}(M^{-1} g - b)$, $\Upsilon := C + \varsigma^2 M^{-1}$. Then, the log posterior density function of $\psi$ is given by

$$
\log p(\psi|y) = -\frac{1}{2} \left[ \text{tr}(M) \log 2\pi + \text{tr}(M \log(\varsigma^2 A^2)) - \log |\varsigma^2 M^{-1}| \right] \\
+ \log |\Upsilon| + \varsigma^{-2} 1^T A^{-2} s - \varsigma^{-2} g^T M^{-1} A^{-2} g \\
+ 1^T A^{-1} (\tilde{g} - \mu) \Upsilon^{-1} (\tilde{g} - \mu) + \log \pi(\psi) - \log p(y). 
$$

(10)

$p(y)$ is a normalizing constant that is analytically intractable.

**Proof.** See Appendix A-A.

**Remark 1.** As shown in the proof of Theorem 1, the statistics $g_n$ and $s_n$ are sufficient for $\psi$. In fact, as we show later, the estimators depend on $y$ only through $(M_n, g_n, s_n)_{n=1:N}$. Thus, from now on we
take } \mathbf{y} := (M_n, g_n, s_n)_{n=1:N} \text{ as the summary of observations from all sensors.}

The following theorem gives the posterior predictive distribution of the model.

**Theorem 2.** The posterior predictive density is given by

\[ p(f_*|\mathbf{y}) = \int p(f_*|\mathbf{y}, \psi)p(\psi|\mathbf{y})d\psi. \]  

(11)

Let \( \mu_* := \mu(\mathbf{x}_n), C_* := C(\mathbf{x}_n, \mathbf{x}_n) \), and let \( \mathbf{k}_* := (C(\mathbf{x}_n, \mathbf{x}_n))_{n=1:N} \in \mathbb{R}^N \) be a column vector, and we have

\[ p(f_*|\mathbf{y}, \psi) = \frac{1}{\sqrt{2\pi \sigma_*^2}} \exp \left( -\frac{(f_* - \bar{f}_*)^2}{2\sigma_*^2} \right), \]  

(12)

\[ \bar{f}_* = \mu_* + \mathbf{k}^T \mathbf{Y}^{-1} (\bar{g} - \mu), \]  

(13)

\[ \sigma_*^2 = C_* - \mathbf{k}^T \mathbf{Y}^{-1} \mathbf{k}. \]  

(14)

**Proof.** See Appendix A-B.

\[ \square \]

**IV. SPATIAL BEST LINEAR UNBIASED ESTIMATOR (S-BLUE)**

We now derive the Spatial Best Linear Unbiased Estimator (S-BLUE). Let \( l(h(\mathbf{y}), f_*) \) denote the loss function, i.e. the loss incurred when using estimator \( h \) when in fact the target quantity is \( f_* \). Let \( R[\Pi, h] \) denote the Bayes risk of \( h \) associated with the prior distribution \( \Pi \), which is defined as the expected value of loss taken over \( \Pi \), i.e.

\[ R[\Pi, h] = \mathbb{E}[l(h(\mathbf{y}), f*)]. \]

To derive the S-BLUE, we restrict the estimator to be a member of the family of linear estimators, that is \( \mathcal{H} := \{h(\mathbf{y}) = \mathbf{w}^T \mathbf{y} + b\} \), where \( \mathbf{w} \) is a weight vector and \( b \) is an intercept, both of which do not depend on \( \mathbf{y} \) or any unknown variables. Hence, the S-BLUE is defined as the optimal linear estimator under quadratic loss, \( l(h(\mathbf{y}), f_*) = (h(\mathbf{y}) - f_*)^2 \), and is given by

\[ \hat{h}_{\text{S-BLUE}} = \arg \min_{h \in \mathcal{H}} R[\Pi, h] = \arg \min_{h \in \mathcal{H}} \mathbb{E} \left[ (h(\mathbf{y}) - f_*)^2 \right], \]  

(15)

where the expectation is taken over the joint distribution of r.v.’s \( (f_*, \mathbf{y}, \psi) \). The next theorem shows that \( \hat{h}_{\text{S-BLUE}} \) can be expressed in closed-form.

**Theorem 3.** \( \hat{h}_{\text{S-BLUE}} \) is given by

\[ \hat{h}_{\text{S-BLUE}}(\mathbf{y}) = \mu_* + \text{Cov}[\bar{g}, f_*]^{-1} \text{Cov}[\bar{g}]^{-1} (\bar{g} - \mathbb{E}[\bar{g}]), \]  

(16)
where \( \bar{g} = M^{-1}g \). \( E[\bar{g}], \text{Cov}[\bar{g}, f_*], \text{Cov}[\bar{g}] \) can all be expressed in closed-form (let \( \odot \) denotes matrix entry-wise multiplication),

\[
E[\bar{g}] = \text{diag}(E[a]) \mu + E[b],
\]

(17)

\[
\text{Cov}[\bar{g}, f_*] = \text{diag}(E[a]) k_*,
\]

(18)

\[
\text{Cov}[\bar{g}] = E[aa^T] \odot (C + \varsigma^2 M^{-1} + \mu \mu^T) + \text{diag}(\mu) (E[ab^T] + E[ab^T]^T)
\]

(19)

\[+ \text{Cov}[\bar{g}] = \text{E}[bb^T] - \text{E}[\bar{g}]\text{E}[\bar{g}]^T.
\]

The various terms in the above equations can all be computed in closed-form, and the details are given in the proof in Appendix A-C.

**Proof.** See Appendix A-C.

The next corollary shows the unbiasedness property of \( \hat{h}_{S\text{-BLUE}} \).

**Corollary 1.** \( \hat{h}_{S\text{-BLUE}} \) is unbiased, that is,

\[
E[\hat{h}_{S\text{-BLUE}}(y)] = E[f_*].
\]

(20)

**Proof.** It is shown via the linearity of expectation.

The following corollary gives the closed-form expression of \( R[\Pi, \hat{h}_{S\text{-BLUE}}] \) (under quadratic loss).

**Corollary 2.** Under quadratic loss, the Bayes risk associated with \( \hat{h}_{S\text{-BLUE}} \) is given by

\[
R[\Pi, \hat{h}_{S\text{-BLUE}}] = C_* - \text{Cov}[\bar{g}, f_*]Cov[\bar{g}]^{-1}\text{Cov}[\bar{g}, f_*].
\]

(21)

**Proof.** Substituting (16) into \( R[\Pi, h] = E[l(h(y), f_*)] \) gives the proof.

The complete S-BLUE algorithm is shown in Algorithm 1.

**V. Empirical Bayes Estimators**

We now derive an algorithm in which we do not restrict the estimator to be linear. The idea of empirical Bayes is to plug in a point estimate \( \hat{\psi} \) into (12) to approximate the posterior predictive distribution, i.e. \( p(\psi|y) \approx \delta_{\hat{\psi}} \). This gives us the corresponding empirical Bayes estimators, which minimize the expected posterior loss, conditional on \( \hat{\psi} \):

\[
\hat{h}_{EB}(y, \hat{\psi}) = \arg \min_{h(y)} E[l(h(y), f_*)|y, \hat{\psi}];
\]

(22)
Algorithm 1: Spatial-Best Linear Unbiased Estimator (S-BLUE)

**Input:** \( x^*, (x_n)_{1:N}, y, (q^{(n)})_{n=1:N, k=0:K} \)

**Output:** Estimator \( \hat{h}_{\text{S-BLUE}}(y) \), Bayes risk \( R[\Pi, \hat{h}_{\text{S-BLUE}}] \)

1. Compute \( \mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[aa^T], \mathbb{E}[bb^T], \mathbb{E}[ab^T] \) (see Appendix A-C).
2. Compute \( \mu^*, \mu, k^*, C, C^* \) by their respective definitions.
3. Compute \( \mathbb{E}[\bar{g}], \text{Cov}[\bar{g}, f^*], \text{Cov}[\bar{g}] \) by Equations (17) - (19).
4. Compute \( \hat{h}_{\text{S-BLUE}}(y) \) by (16).
5. Compute \( R[\Pi, \hat{h}_{\text{S-BLUE}}] \) by (21).
6. return \( \hat{h}_{\text{S-BLUE}}(y), R[\Pi, \hat{h}_{\text{S-BLUE}}] \).

To complete the specification of the estimator we are required to define appropriate *loss functions*. We present a few loss functions which are widely used and their corresponding approximated Bayes estimators.

1) **Quadratic loss function:**

\[
l_{\text{quad}}(h(y), f^*) = (h(y) - f^*)^2,
\]

the Bayes estimator is the conditional expectation (minimum mean squared error estimator, or MMSE estimator),

\[
\hat{h}_{\text{MMSE}}(y) = \mathbb{E}[f^*|y] = \int_R f^* p(f^*|y) df^*,
\]

where \( p(f^*|y) \) is given in Theorem 2. The empirical Bayes version of \( \hat{h}_{\text{MMSE}}(y) \) is given by

\[
\hat{h}_{\text{EB-MMSE}}(y, \hat{\psi}) = \mathbb{E}[f^*|y, \hat{\psi}] = \int_R f^* p(f^*|y, \hat{\psi}) df^* 
\approx \hat{h}_{\text{MMSE}}(y),
\]

where \( p(f^*|y, \hat{\psi}) \) is given in (12).

2) **Absolute difference loss function**

\[
l_{\text{abs}}(h(y), f^*) = |h(y) - f^*|,
\]

the Bayes estimator is the conditional median (least absolute deviation estimator, or LAD estimator),

\[
\hat{h}_{\text{LAD}}(y) = \text{median}(f^*|y).
\]

The empirical Bayes version of \( \hat{h}_{\text{LAD}}(y) \) is given by

\[
\hat{h}_{\text{EB-LAD}}(y, \hat{\psi}) = \text{median}(f^*|y, \hat{\psi}) \approx \hat{h}_{\text{LAD}}(y).
\]
3) 0 – 1 loss function

\[ l_{0-1}(h(y), f^*) = \mathbb{1}_{\{f^*, h(y) \leq f^* + \delta_f\}} \]

the Bayes estimator is the conditional mode (maximum a posteriori estimator, or MAP estimator),

\[ \hat{h}_{\text{MAP}}(y) = \arg \max_{f^*} p(f^* | y). \]

The empirical Bayes version of \( \hat{h}_{\text{MAP}}(y) \) is given by

\[ \hat{h}_{\text{EB-MAP}}(y, \hat{\psi}) = \arg \max_{f^*} p(f^* | y, \hat{\psi}) \approx \hat{h}_{\text{MAP}}(y). \]

For all the aforementioned Bayes estimators, we first need to find a point estimator for \( \psi \). To achieve this we find the MAP estimator of \( \psi \), which aims at maximizing the posterior density \( p(\psi | y) \), given in Theorem 1. The MAP estimator is then given by

\[ \hat{\psi} = \arg \max_{\psi} p(\psi | y) \]

\[ = \arg \max_{\psi} \left[ -\frac{1}{2} \left( \text{tr}(M) \log 2\pi + \text{tr}(M \log(\varsigma^2 A^2)) \right) \right. \]

\[ - \log |\varsigma^2 M^{-1}| + \log |\Upsilon| + \varsigma^{-2} \mathbf{1}^T A^{-2} s \]

\[ - \varsigma^{-2} g^T M^{-1} A^{-2} g + (\hat{g} - \mu)^T \Upsilon^{-1} (\hat{g} - \mu) \]

\[ + \log \pi(\psi) \left. \right]. \tag{23} \]

Note that the optimization objective does not involve the \( p(y) \), since it is a constant. Thus, the empirical Bayes estimators could be computed in the following two-stage algorithm:

I. Compute \( \hat{\psi} \) by solving the optimization problem \( \arg \max_{\psi} p(\psi | y) \).

II. Plug in \( \hat{\psi} \) to compute \( \hat{h}_{\text{EB}}(y, \hat{\psi}) \).

In order to solve the optimization problem in Step I, we develop two algorithms. The first approach is a stochastic optimization method named Cross-Entropy method (CEM), and the second approach is Iterated Conditional Modes (ICM) which is based on iterative greedy search.

A. Cross-Entropy Method (CEM)

The Cross-Entropy method (CEM) is an stochastic algorithm that is suitable for solving combinatoric or continuous optimization problems. Suppose we have a maximization problem with a unique optimizer,

\[ \varphi = \arg \max_{\varphi \in \Phi} J(\varphi), \]
where \( J(\cdot) \) is the objective function, \( \Phi \) is the domain, and \( \varphi \) is the parameter vector. We solve the optimization problem by considering the level sets of the objective function \( \{ \varphi : J(\varphi) \geq \gamma \} \), for \( \gamma \in \mathbb{R} \). When \( \gamma = \hat{J} = \max_{\varphi \in \Phi} J(\varphi) \), we have \( \{ \varphi : J(\varphi) \geq \gamma \} = \{ \hat{\varphi} \} \). Next, let us define a family of probability measures \( \{ \mathbb{P}_\theta : \theta \in \Theta \} \) on \( \Phi \) with densities \( \{ w_\theta : \theta \in \Theta \} \) that are parameterized by \( \theta \in \Theta \). Let \( \mathbb{E}_\theta \) denote the expectation taken with respect to \( \mathbb{P}_\theta \). Let us fix \( \theta \) and \( \gamma \), and define a rare event probability problem,

\[
\mathbb{P}_\theta[ J(\varphi) \geq \gamma ] = \mathbb{E}_\theta[1_{\{J(\varphi) \geq \gamma \}}] = \int_{\Phi} 1_{\{J(\varphi) \geq \gamma \}} w_\theta(\varphi) d\varphi.
\]

Instead of approximating this probability naively by sampling from \( w_\theta \), the importance sampling method is used. Let \( \tilde{\theta} \) denote the importance sampler, where \( \tilde{\theta} \in \Theta \). Importance sampling approximates the rare event probability by,

\[
\mathbb{P}_\theta[ J(\varphi) \geq \gamma ] = \mathbb{E}_{\tilde{\theta}} \left[ \frac{1_{\{J(\varphi) \geq \gamma \}} w_\theta(\varphi)}{w_{\tilde{\theta}}(\varphi)} \right] 
\approx \frac{1}{S} \sum_{s=1}^{S} 1_{\{J(\tilde{\varphi}[s]) \geq \gamma \}} \frac{w_\theta(\tilde{\varphi}[s])}{w_{\tilde{\theta}}(\tilde{\varphi}[s])},
\]

where \( \tilde{\varphi}[1], \ldots, \tilde{\varphi}[S] \) are \( S \) independent samples generated from \( w_{\tilde{\theta}} \). The optimal importance sampler \( w_{\tilde{\theta}} \) is selected through the cross-entropy criterion,

\[
\hat{\theta} = \arg \min_{\tilde{\theta} \in \Theta} \int_{\Phi} 1_{\{J(\varphi) \geq \gamma \}} w_\theta(\varphi) \log \frac{w_\theta(\varphi)}{w_{\tilde{\theta}}(\varphi)} d\varphi 
\approx \arg \max_{\tilde{\theta} \in \Theta} \frac{1}{S} \sum_{s=1}^{S} 1_{\{J(\tilde{\varphi}[s]) \geq \gamma \}} \log w_{\tilde{\theta}}(\tilde{\varphi}[s]),
\]

where \( \varphi[1], \ldots, \varphi[S] \) are \( S \) independent samples generated from \( w_{\tilde{\theta}} \). Notice that the last line of (25) corresponds to the maximum likelihood estimation (MLE) of \( \hat{\theta} \) when the samples are \( \{ \varphi[s] : J(\varphi[s]) \geq \gamma \} \). The CEM starts from an initial sampling distribution \( w_{\theta_0} \) and iteratively updates the threshold \( \hat{\gamma} \) and the sampling distribution \( w_{\tilde{\theta}} \). For a detailed introduction of CEM, see [41], [42]. The complete procedure is detailed in Algorithm 2.

We can now link CEM to the MAP estimation problem in (23). We define the objective function as the (un-normalized) log-posterior conditional density,

\[
J(\psi) = \log p(\psi|y) + \log \pi(\psi).
\]

For the purpose of demonstrating the CEM, let us assume here that under prior distribution \( \pi_{k^*}(a_n, b_n) \) have a bivariate normal distribution,

\[
\left( \begin{array}{c} \log a_n \\ b_n \end{array} \right) \sim \mathcal{N}(\nu_k, \Xi_k).
\]
Algorithm 2: Cross-Entropy Method (CEM)-based Optimizer

**Input:** number of importance samples \( S \), \( \rho \in (0, 1) \) (typically \( 0.001 \leq \rho \leq 0.01 \)), initial sampler parameter \( \hat{\theta}_0 \), objective function \( J \)

**Output:** \( \hat{\varphi} = \arg \max_{\varphi \in \Phi} J(\varphi) \)

1. \( \hat{\gamma}_0 \leftarrow -\infty \), \( t \leftarrow 1 \).

2. repeat
   
   3. Generate \( S \) independent samples \( \varphi^{[1]}, \ldots, \varphi^{[S]} \) from \( w_{\hat{\theta}_{t-1}} \).
   
   4. Compute \( J(\varphi^{[1]}), \ldots, J(\varphi^{[S]}) \).
   
   5. \( \hat{\gamma}_t \leftarrow \) the \((1 - \rho)\)-sample quantile of \( J(\varphi^{[1]}), \ldots, J(\varphi^{[S]}) \).
   
   6. \( \hat{\theta}_t \leftarrow \arg \max_{\theta \in \Theta} \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}_{\{J(\varphi^{[s]}) \geq \hat{\gamma}_t\}} \log w_{\theta}(\varphi^{[s]}) \).
   
   7. \( t \leftarrow t + 1 \).

   until termination condition is triggered;

8. Set \( \hat{\varphi} \) to be the sample with the largest \( J(\hat{\varphi}) \) so far.

9. return \( \hat{\varphi} \).

Note that this can be easily adapted for other prior distributions. We choose the family of sampling distributions such that,

\[
  w_{\theta}(\psi) = \prod_{n=1}^{N} w_{\theta^{(n)}}(\psi^{(n)}),
\]

\[
  w_{\theta^{(n)}}(\psi^{(n)}) = \delta_{\psi^{(n)}}^{(n)} + \sum_{k=1}^{K} r^{(n)} \mathcal{N}((\log a_n, b_n)^T; \nu^{(n)}_k, \Xi^{(n)}_k).
\]

Here, we have \( \theta = (\theta^{(n)})_{1:N} = ((r^{(n)}_{k=k:K}), (\nu^{(n)}_k)_{k=1:K}, (\Xi^{(n)}_k)_{k=1:K})_{n=1:N} \). Before running the CEM algorithm, we set \( \hat{\theta}_0 \) such that \( w_{\hat{\theta}_0} \) coincides with the prior distribution \( \pi \). Under this setting, the optimization in Line 6 of Algorithm 2 corresponds to the maximum likelihood estimation (MLE) of \( \theta \), given independent samples \( \{\psi^{[s]} : J(\psi^{[s]}) \geq \hat{\gamma}_t\} \). This decomposes into sub-problems

\[
  \hat{\theta}^{(n)} = \arg \max_{\theta^{(n)}} \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}_{\{J(\psi^{[s]}) \geq \hat{\gamma}_t\}} \log w_{\theta^{(n)}}(\psi^{[s]}).
\]

Since \( w_{\theta^{(n)}} \) is a mixture distribution, the MLE does not admit a closed-form solution and we use the expectation-maximization (EM) algorithm. The EM algorithm is an iterative procedure that computes a local optimum of the likelihood function. For notational simplicity, we drop the superscripts and subscripts with \( n \) for now, and denote the samples used to obtain the MLE.
as $\psi[1], \ldots, \psi[S]$. To apply the EM algorithm, let us first introduce the auxiliary variables. Let $z[s] \in \{0, \ldots, K\}$ for $s = 1, \ldots, S$ be the discrete auxiliary variables, such that

$$p(\psi[s] | z[s] = 0; \theta) = \delta_{\psi[0]}$$
$$p(\psi[s] | z[s] = k; \theta) = \mathcal{N}((\log a, b)^T; \tilde{\nu}_k, \tilde{\Xi}_k),$$

for $k = 1, \ldots, K,$

$$p(z[s] = k; \theta) = r_k,$$ for $k = 0, \ldots, K.$

This gives the marginal distributions $w_\theta(\psi[s])$ above. The EM algorithm starts with an initial estimate $\hat{\theta}_0$, and iteratively updates the estimated parameters through two steps. In the expectation step (E-step), a lower bound of the log-likelihood function is constructed by first computing the conditional distributions of the auxiliary variables given the estimate of the parameters in the $t$-th iteration

$$p(z[s] = k | \psi[s]; \hat{\theta}_t) = \frac{p(\psi[s] | z[s] = k; \hat{\theta}_t)p(z[s] = k; \hat{\theta}_t)}{\sum_{k'=0}^K p(\psi[s] | z[s] = k'; \hat{\theta}_t)p(z[s] = k'; \hat{\theta}_t)},$$

for $k = 0, \ldots, K,$

and then compute the expected value of the log-likelihood function with respect to this conditional distribution, given by

$$Q(\theta; \hat{\theta}_t) = \sum_{s=1}^S \mathbb{E}_{(z[s] | \psi[s], \hat{\theta}_t)}[\log w_\theta(\psi[s], z[s])].$$

In the maximization step (M-step), the estimated parameters in the $(t + 1)$-th iteration are computed by maximizing the lower bound $Q(\theta; \hat{\theta}_t)$, that is,

$$\hat{\theta}_{t+1} = \arg \max_{\theta \in \Theta} Q(\theta; \hat{\theta}_t).$$

The algorithm is summarized in Algorithm 3. For details about the EM algorithm, see [43].

B. Iterated Conditional Modes (ICM)

We propose a second optimization method to find the MAP estimator $\hat{\psi}$ which is based on iterative greedy search. Since $\hat{\psi} \in \mathbb{R}^{2N}$, the dimensionality of the optimization problem can be high if $N$ is large. In addition, for $n = 1, \ldots, N$, the distribution of $\psi_n$ contains an atom. Hence, to improve the computational efficiency in these settings, we seek to reduce the complexity of the MAP estimation by reducing the global search problem to a sequence of iterative local solutions. In this approach, we therefore decompose the problem involving high dimensional non-linear
**Algorithm 3: Expectation-Maximization Algorithm**

**Input:** $S$ samples $\psi^{[1]}, \ldots, \psi^{[S]}$, initial estimate $\hat{\theta}_0$

**Output:** Estimated parameter $\hat{\theta}$

1. $\hat{p} \leftarrow -\infty$, $t \leftarrow 0$.
2. repeat
   3. for $s = 1 \ldots S$ do
   4.   Compute $p_k^{[s]} := p(z^{[s]} = k|\psi^{[s]}; \hat{\theta}_t)$, for $k = 0, \ldots, K$.
5.   (E-step) Construct $Q(\theta; \hat{\theta}_t) = \sum_{s=1}^{S} \sum_{k=0}^{K} p_k^{[s]} p(\psi^{[s]}|z^{[s]} = k; \theta)$.
6.   (M-step) $\hat{\theta}_{t+1} \leftarrow \arg \max_{\theta \in \Theta} Q(\theta; \hat{\theta}_t)$. This decomposes into $K + 1$ weighted MLE problems.
7.   $t \leftarrow t + 1$.
   until termination condition is triggered;
8. $\hat{\theta} \leftarrow \hat{\theta}_t$.
9. return $\hat{\theta}$.

---

root search into a sequence of iterated local root search problems of iterated conditional modes (ICM) [44]. Let $\psi_{(-n)} := (\psi_m)_{m \neq n}$. In each iteration of ICM, we fix $\psi_{(-n)}$ and compute the mode of the conditional posterior distribution $\hat{\psi}_n = \arg \max_{\psi_n} p(\psi_n|y, \psi_{(-n)})$ through the conjugate gradient algorithm. ICM converges to a local maximum of the objective function in the sense that $\hat{\psi}_n$ is the mode of the conditional posterior distribution for $n = 1, \ldots, N$.

To optimize the conditional posterior distributions, we decomposed them (up to a normalizing constant) as follows, for $n = 1, \ldots, N$,

$$p(\psi_n|y, \psi_{(-n)}) \propto p(y|\psi) \pi(\psi_n) \propto p(y_n|y_{(-n)}, \psi) \pi(\psi_n),$$

where $y_n = (y_{n,m})_{m=1:M_n}$, $y_{(-n)} = (y_{i,m})_{i \neq n, m=1:M_i}$. Let $m_n := \mu(x_n)$, $\mu_{(-n)} := (\mu_i)_{i \neq n}$. Let $C_n := C(x_n, x_n)$. Let $\Upsilon_{(-n,n)} \in \mathbb{R}^{(N-1)}$ denote the sub-matrix of $\Upsilon$ involving the cross-terms between sensor $n$ and the rest of sensors. Let $\Upsilon_{(-n)} \in \mathbb{R}^{(N-1) \times (N-1)}$ denote the sub-matrix of $\Upsilon$ related to sensors other than $n$. Let $\tilde{g}_{(-n)} := (\tilde{g}_i)_{i \neq n}$. Completely analogous to Theorem 2, we have that

$$(F(x_n)|y_{(-n)}, \psi) \sim \mathcal{N}(\nu_n, \zeta_n),$$
where
\[
\nu_n = \mu_n + Y_{(-n,n)}^T Y_{(-n)}^{-1} (\tilde{g}_{(-n)} - \mu_{(-n)}) ,
\]
\[
\zeta_n = \mathcal{C}_n + \varsigma^2 - Y_{(-n,n)}^T Y_{(-n)}^{-1} Y_{(-n,n)}.
\]

One verifies that \(\nu_n\) and \(\zeta_n\) do not depend on \(\psi_n\). Therefore, following a derivation similar to that in Theorem 1, we have,
\[
\log p\left(y_n|y_{(-n)}, \psi\right) = \log \left[ \int p(y_n|f(x_n), \psi)p(f(x_n)|y_{(-n)}, \psi)df(x_n) \right]
= -\frac{1}{2} \left[ M_n \log 2\pi + (M_n - 1) \log(\varsigma^2 a_n^2) \right.
+ \log(a_n^2 M_n \zeta_n + \varsigma^2 a_n^2) + \varsigma^{-2} a_n^{-2} (s_n - M_n^{-1} g_n^2)
+ (\zeta_n + \varsigma^2 M_n^{-1})^{-1} (\tilde{g}_n - \nu_n)^2 \bigg].
\]

Thus, the log-conditional likelihood as well as its partial derivatives can be efficiently evaluated. The ICM algorithm then separately treats the continuous and discrete part of the parameter space, that is, comparing \(\sup_{\psi_n \neq \psi_0} \log p\left(\psi_n|y, \psi_{(-n)}\right)\) and \(\log p\left(\psi_0|y, \psi_{(-n)}\right)\).

The details of the ICM algorithm are shown in Algorithm 4. To account for the multi-modality of the posterior distribution, we adopt a standard multiple start initialization strategy, that is to run ICM from a number of random initial estimates. This corresponds to running Algorithm 4 multiple times with different initial values.

VI. EXPERIMENTS WITH SYNTHETIC DATA

We conduct two experiments with synthetically generated data to study the performance of the methods we proposed including CEM, ICM and S-BLUE. In Section VI-A, we study the sensitivity of the proposed methods to the strength of distortions. In Section VI-B, we perform a realistic simulation and analyse the overall performance of the proposed methods. In the studies, the proposed methods are compared to two baselines, the “oracle” case in which the distortions \(\psi\) are known exactly, and the “naive” case in which distortions are disregarded in the prediction, i.e. \(\hat{\psi}_n = \psi_0\), for \(n = 1, \ldots, N\).

A. Synthetic Experiment 1: Estimation Accuracy and Sensitivity to Distortions

In this experiment, we study an ideal scenario where the signal-to-noise ratio (SNR) is high and observations are plentiful. We fix the number of observations per sensor to be 50, and the
Algorithm 4: Iterative Conditional Modes Algorithm

Input: \((x_n)_{1:N}, y, (q^{(n)}_k)_{n=1:N, k=0:K}, (\pi_k)_{k=1:K}\)

Output: Estimation of posterior mode \(\hat{\psi}\)

1. Randomly initialize \(\hat{\psi}\).

2. repeat
   3. for \(n = 1 \ldots N\) do
      4. Compute \(\nu_n, \zeta_n\) from (26) and (27).
      5. \(\hat{\psi}_n \leftarrow \arg \max_{\psi_n \in \mathbb{R}_+ \times \mathbb{R}, \psi_n \neq \psi^0} \log p(\psi_n | y, \hat{\psi}_{(-n)})\), by running the conjugate gradient algorithm disregarding the atom at \(\psi^0\).
      6. if \(\log p(\psi^0 | y, \hat{\psi}_{(-n)}) < \log p(\hat{\psi}_n | y, \hat{\psi}_{(-n)})\) then
         7. \(\hat{\psi}_n \leftarrow \hat{\psi}_n\).
      else
         8. \(\hat{\psi}_n \leftarrow \psi^0\).
   until termination condition is triggered;

9. return \(\hat{\psi}\).

SNR to be 15dB. Notice that for the ease of comparison, all SNRs are measured at the sensor level, that is, the SNRs of aggregated observations. Under the i.i.d. noise assumption, we define
\[
\text{SNR} = 10 \log_{10} \left( \frac{\text{Var}(F)}{\varsigma^2} \right),
\]
where \(\text{Var}(F)\) denotes the signal variance, \(\varsigma^2\) denotes observation noise variance, and \(M\) denotes the number of observations per sensor.

We simulate a spatial field defined on the two-dimensional square \(X = [0, 1]^2\) with mean 10, i.e. \(\forall x \in X, \mu(x) = 10\), and a Matern covariance function with \(\nu = 3/2\), \(\text{Var}(F) = 100\), and length scales=0.3, i.e. \(\forall x, x' \in X, C(x, x') = 100 \left( 1 + \sqrt{3} ||x - x'||_{0.3} \right) \exp \left( -\sqrt{3} ||x - x'||_{0.3} \right) \cdot ||x - x'||\) is a distance metric on \(X\), which corresponds to the Euclidean distance in this case. The contour plot of the simulated spatial field is shown in Figure 2a.

Subsequently, 100 sensors are randomly placed in the square. 50 out of the 100 sensors are fixed to have identical distortion parameters, and the rest are set to have the default transformation parameters \(\psi^0\), that is, non-distorting. For the sensors with distortions, we first fix the offset parameter \(b_n\) at 5, and vary the gain parameter \(a_n\) from 1 to 1.6. Then we fix the gain parameter at 1.2, and vary the offset parameter from 0 to 12. With each setting of distortion parameters, 100 sets of noisy observations are randomly simulated. For each set of observations,
the three proposed methods: CEM, ICM and S-BLUE, along with the two baselines oracle and naive, are used to reconstruct the spatial field. Here, weakly informative prior for the distortion parameters are used, which has a single category ($K = 1$) given by $q^{(n)}_1 = 0.5$ for all $n$, where under $\pi_1$, $a_n \sim \log \mathcal{N}(0.25, 0.1^2)$, $b_n \sim \mathcal{N}(6, 3^2)$. The reconstruction accuracy is evaluated by the mean-squared-errors (MSE) at a $100 \times 100$ grid on $[0, 1]^2$.

Figure 3 shows the reconstruction accuracy averaged over 100 realizations. For better interpretability, the ratio between the MSE and the prior variance, referred to as the relative MSE, is shown. From Figure 3, one observes that the MSE of the oracle stayed constant, as expected, while the MSE of the naive baseline increased rapidly when the distortion parameters increased. The MSE of S-BLUE first decreased and then increased slightly. This is due to the way the prior distribution of distortions were set up. The prior mean of $a_n$ and $b_n$ were 1.29 and 6.0, respectively. Since S-BLUE makes predictions based purely on the prior information, its performance is best when the actual distortion is closer to the prior mean. The two empirical Bayes-based methods shows decreasing MSE when the gain parameter increases and slightly increasing MSE when the offset parameter increases. The reason is that since the gain parameter affects both the location and the spread of the observations, while the offset parameter only affects the location, the gain in the distortion was more noticeable and thus easier to detect. The MSEs of ICM and CEM were almost identical. Figure 4 shows the average false positive rate (FPR) and false negative rate (FNR) of ICM and CEM. The FPR is defined as the proportion of non-distorting sensors that were estimated to be distorting, and the FNR is defined as the proportion of distorting sensors that were estimated to be non-distorting. Again the FPR and FNR of the two methods were almost identical. Also notice that the high FNR when the gain was 1 and the offset was 5. This was caused by the short length scale (0.3), which made detection of the offset hard due to low spatial correlation.

B. Synthetic Experiment 2: Realistic Scenario

In the second synthetic experiment, we study a realistic scenario where each sensor has differing distortion parameters, and vary the SNR as well as the number of observations.

We again simulate a spatial field defined on the two-dimensional square $[0, 1]^2$. This time, however, length scales is set to be 0.5, so the spatial correlation decays at a slower rate. The contour plot of the simulated spatial field is shown in Figure 2b. The 100 sensors are placed at the same locations as in the synthetic experiment 1. 50 out of the 100 sensors are randomly selected to have the different distortion parameters generated from the following prior with
three categories \((K = 3)\), given by \(q_1^{(n)} = q_2^{(n)} = q_3^{(n)} = \frac{1}{6}\) for all \(n\), where under \(\pi_1\), \(a_n \sim \log \mathcal{N}(-0.4, 0.05^2)\), \(b_n \sim \mathcal{N}(0, 0.2^2)\), under \(\pi_2\), \(a_n \sim \log \mathcal{N}(0.2, 0.05^2)\), \(b_n \sim \mathcal{N}(0, 0.2^2)\), under \(\pi_3\), \(a_n \sim \log \mathcal{N}(0, 0.05^2)\), \(b_n \sim \mathcal{N}(10, 2^2)\), and the rest of the sensors are set to be non-distorting. With each setting of distortion parameters, we randomly simulate 100 sets of noisy observations. For
Fig. 4: Synthetic experiment 1 – false positive rate (FPR) and false negative rate (FNR) resulted from varying strengths of distortions.

each set of observations, we test the proposed methods along with the baselines as in the synthetic experiment 1.

Figure 5 shows the relative MSE averaged over 100 realizations, with different number of observations per sensor and different SNR. Figure 6 shows the FPR and FNR of ICM and CEM averaged over 100 realizations. Observe that the MSE of the baselines and S-BLUE did not change with different number of observations per sensor because they depend only on the mean of observations from each sensor. ICM and CEM, on the other hand, benefited from having access to more observations. The naive baseline showed a peculiar trend that first decreased and then increased when SNR increased. The reason is that the naive estimator is a linear combination of the prior mean and the observations, where the weights depend on the SNR. With high SNR, the naive estimator placed a high weight on the distorted observations, thus making the MSE high. The MSE of S-BLUE decreased steadily when SNR increased, and eventually flattened out. The MSE of ICM and CEM depended highly on the number of observations. CEM had the best reconstruction quality among the proposed methods when observations were plentiful or when the SNR was high. In comparison, ICM performed considerably worse than CEM when the number of observations was 5 and 20. This was due to the higher spatial correlation, which made the dependency between distortion parameters higher in the posterior and reduced the effectiveness of iterative greedy search. Notice that the FPR of CEM was close to 1 when the
Fig. 5: Synthetic experiment 2 – mean-squared-errors resulted from varying number of observations.

SNR was low and the observations were scarce. This indicates that ICM and CEM estimated all of the sensors as distorting. The reason was that with high noise variance, the likelihood had a flat shape, and thus the posterior mode did not contain non-distorting sensors. It can also be observed that the FPR and FNR were low when the number of observations was 100 and the SNR was high.
VII. Experiments with Real Data

We study the 2017 US temperature dataset from the US EPA\textsuperscript{2}. The dataset contains 309,226 rows, with the following fields:

- **State.Code**: the numerical code of the state.
- **County.Code**: the numerical code of the county.
- **Site.Num**: the numerical code of the temperature monitoring site.
- **Longitude**: the longitude of the temperature monitoring site.
- **Latitude**: the latitude of the temperature monitoring site.
- **Date.Local**: the local date on which the temperature measurement is taken.
- **X1st.Max.Value**: the maximum hourly temperature measurement of a day.

A. Preprocessing

The first step of preprocessing is to remove the irrelevant fields from the dataset. For temperature measurements, we take the maximum hourly measurement of every day. The temperature are converted from Fahrenheit into Celsius. We noticed that there is an obvious outlier which corresponds to 125°C (the next highest measurement is 55°C). Therefore, this measurement is removed from the dataset.

\textsuperscript{2}https://aqs.epa.gov/aqswaib/airdata/download_files.html, retrieved on 5 June 2018.
In the dataset, there are 830 monitoring sites in total. Figure 7 is a scatter plot showing the spatial locations of these monitoring sites. Not all monitoring sites have taken measurements on all 365 days. We refer to these as missing measurements. Out of the 830 monitoring sites, 55 contain more than 30% of missing measurements. Overall, 16,980 measurements are missing, which is 5.6% of the 302,950 measurement (365×830).

**Remark 2.** We have noticed that some monitoring sites have reported temperature measurements by different types of instrument, which explains why we have slightly fewer measurements after we sort measurements into the site & date format. However, measurements on the same day at the same monitoring site with different instruments tend to be close hence we choose an arbitrary set whenever a monitoring site reports multiple sets of measurements.

Figure 8 shows the histogram and the normal quantile-quantile (Q-Q) plot of the temperature measurements.

Figure 8 shows the histogram and the normal Q-Q plot of the temperature measurements.
measurements. From Figure 8a, one can see that the measurements contain quantization artifacts. Nonetheless, the measurements are treated as continuous. From Figure 8b, one sees that the dataset does not contain abnormally large or small values, and is slightly left-skewed compared to a normal distribution. Overall, the normality assumption holds approximately, as we have not yet considered the seasonal shift of temperatures. Here are some summary statistics of the measurements:

| TABLE I: Summary statistics of the temperature measurements. |
|---------------------------------|-----------------|-----------------|-----------------|
| mean                           | variance        | skewness        | kurtosis        |
| 19.7556                        | 120.6565        | -0.5556         | 2.9809          |
| min                            | 1st-quartile    | median          | 3rd-quartile    | max             |
| -50.0000                       | 12.7778         | 21.1111         | 28.3333         | 55.0000         |

B. Smoothing of Spatial Field

The first step of the spatial analysis is to model the daily observations as noisy samples from a Gaussian process without distortions and estimate its hyperparameters. For subsequent analyses, we take only data from the 20 days between day 181 to day 200 from the dataset, and restrict ourselves to the contiguous United States (that is, excluding Hawaii and Alaska) since Hawaii and Alaska are far away from the rest of the Unites States. As for the covariance function, we again choose to use the Matérn covariance function with $\nu = 3/2$ to allow for high flexibility while keeping the spatial field mean-square differentiable. For each calendar day, we estimate the signal mean, signal variance, noise variance and length-scale via maximum marginal likelihood estimation (see Chapter 5 of [39] for details) using all available observations on that day. Then, the median of the estimated values on 20 days are taken as the estimated hyperparameters of the GP. The estimated hyperparameters are shown in Table II.

| TABLE II: Estimated GP hyperparameters |
|----------------------------------------|-----------------|-----------------|
| signal mean (°C)                       | signal std. dev. (°C) |
| 30.5034                                 | 4.6587          |
| noise std. dev. (°C)                    | length-scale (km) |
| 1.6340                                  | 174.3699        |
Using the estimated hyperparameters, we reconstruct the spatial field for each calendar day at both a 100 × 100 grid of locations and the 824 sensor locations by computing the posterior mean and the posterior covariance matrix, conditional on all the available observations on that day. Since the posterior mean is usually much smoother compared to the actual spatial field, a sample spatial field is generated from the posterior distribution to make it realistic. The generated spatial field at a grid of locations is treated as the ground truth of the spatial field, and the generated field intensities at the 824 sensor locations are treated as as the noise-free sensor reading from which noisy and distorted observations are generated.

C. Experimental Settings

In order to evaluate the reconstruction accuracy of the proposed methods, 25% of the sensor locations are left out as the test set. For the 618 remaining locations, we randomly generate the distortion parameters from three different settings. The three settings are labelled as “low”, “medium” and “high”, corresponding to the proportion of distorting sensors. In the “low”, “medium” and “high” settings, each sensor has a respective probability of 0.3, 0.5 and 0.7 to introduce distortion. Under all three settings, the distorting sensors have one of the three following categories, with equal probabilities. Under \( \pi_1 \), \( a \sim \log\mathcal{N}(-0.4, 0.05^2) \), \( b \sim \mathcal{N}(0, 0.2^2) \), under \( \pi_2 \), \( a \sim \log\mathcal{N}(0.2, 0.05^2) \), \( b \sim \mathcal{N}(0, 0.2^2) \), under \( \pi_3 \), \( a \sim \log\mathcal{N}(0, 0.05^2) \), \( b \sim \mathcal{N}(10, 2^2) \). For each setting, we randomly simulate noisy observations at each sensor location and then apply the corresponding distortions. In addition, we examine the effect of the number of observations and effective SNR. Specifically, we examine the four following cases:

1) 10 observations per sensor, SNR=5dB;
2) 10 observations per sensor, SNR=15dB;
3) 50 observations per sensor, SNR=5dB;
4) 50 observations per sensor, SNR=15dB.

D. Results and Discussion

Figure 9 shows the relative MSE averaged over 20 days, and Figure 10 shows the FPR and FNR of ICM and CEM, under each setting. Observe that same as in the synthetic experiment 2, the MSE of the baselines and S-BLUE did not change with different number of observations per sensor. ICM and CEM, on the other hand, benefited from having access to more observations.
The naive baseline showed worse MSE when SNR was high. The reason was the same as in the synthetic experiment 2. In addition, the MSE of naive baseline is clearly affected by a higher proportion of distorting sensors.

S-BLUE showed stable accuracies across all settings. Since S-BLUE does not estimate the distortion parameters, the error mainly resulted from smoothing. This can be seen clearly from the reconstructed spatial fields in Figure 11, which will be discussed later.

Compared to the simple method S-BLUE, the more sophisticated methods ICM and CEM have the additional benefit of being able to estimate the distortion parameters. ICM and CEM benefited from more plentiful observations as well as higher SNR. With small number of observations and low SNR, the reconstruction accuracy of CEM was about the same as S-BLUE, and the estimation of distortion parameters showed high FPR (i.e. mistakenly label non-distorting sensors as distorting). With more plentiful observations and higher SNR, both the reconstruction accuracy and parameter estimation accuracy improved significantly. Notice that although ICM is also based on empirical Bayes, its accuracies were much worse compared to CEM. The reason was that the iterative greedy search method tend not to work well with high-dimensional mixed discrete-continuous optimization problems.

**Remark 3.** We noticed that if the SNR is set to be higher than 20dB, the accuracies of both ICM and CEM deteriorates greatly. With a high SNR (hence low noise variance), the posterior density of the model tend to be highly irregular. Thus the optimization problem might be highly ill-posed. In general, there are no universal solution to this. Some form of relaxation might help to improve the performance, but they are not examined in this work.

Finally, let us examine the reconstructed spatial fields. Figure 11 shows the heat maps of spatial fields reconstructed by S-BLUE, ICM, CEM, and the two baselines with the settings: high distortion setting, 10 observations per sensor, SNR=15dB. The ground truth spatial field is also presented for reference. First, notice that the ground truth contained much more details compared to the reconstructions, because all of the estimators here have the smoothing effect. Overall, the reconstruction of S-BLUE is much smoother and contains fewer details. This is due to nature of S-BLUE, as it does not estimate the distortion parameters. ICM and CEM, on the other hand, preserved many of the details, and their reconstructions were overall close to the one produced by the oracle. The naive method, however, produced a noticeably inaccurate reconstruction.
Fig. 9: Mean-squared-errors resulted from varying number of observations and signal-to-noise ratio (SNR).

Fig. 10: False positive rates (FPR) and false negative rates (FNR) of ICM and CEM.
VIII. Conclusion

This paper addressed the problem of spatial field reconstruction based on distorted sensor readings. A new spatial field model based on a mixture of Gaussian process experts was developed. We developed two approaches to solve the inference problem. The first approach uses a linear Bayes estimator named the Spatial Best Linear Unbiased Estimator (S-BLUE), which is a low-complexity algorithm relying only on prior information. The second approach is a two-stage algorithm based on empirical Bayes, in which the unknown distortion parameters of the sensors are estimated based on distorted observations. In addition, we developed two optimization procedures for the two-stage algorithm, the first is based on the Cross-Entropy method (CEM) and the second is based on Iterated Conditional Mode (ICM) which is an iterative greedy search procedure. We preformed two synthetic experiments as well as an experiment based on real temperature data from US EPA to assess the spatial field reconstruction accuracy of the proposed approaches. The results showed significant improvement compared to the estimation approach that neglects sensor distortions.

References

[1] O. Saukh, D. Hasenfratz, C. Walser, and L. Thiele, “On rendezvous in mobile sensing networks,” in Real-World Wireless Sensor Networks. Springer, 2014, pp. 29–42.
[2] O. Saukh, D. Hasenfratz, and L. Thiele, “Reducing multi-hop calibration errors in large-scale mobile sensor networks,” in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*. ACM, 2015, pp. 274–285.

[3] A. Arfire, A. Marjovi, and A. Martinoli, “Model-based rendezvous calibration of mobile sensor networks for monitoring air quality,” in *2015 IEEE SENSORS*. IEEE, 2015, pp. 1–4.

[4] P. Rathore, D. Kumar, S. Rajasegarar, and M. Palaniswami, “Maximum entropy-based auto drift correction using high- and low-precision sensors,” *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 3, p. 24, 2017.

[5] H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, “Short-term solar power forecasting based on weighted gaussian process regression,” *IEEE Transactions on Industrial Electronics*, vol. 65, no. 1, pp. 300–308, 2018.

[6] K. Sohraby, D. Minoli, and T. Znati, *Wireless sensor networks: technology, protocols, and applications*. John Wiley & Sons, 2007.

[7] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, “Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 205–215, 2013.

[8] B. Ristic, D. E. Clark, and N. Gordon, “Calibration of multi-target tracking algorithms using non-cooperative targets,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 3, pp. 390–398, 2013.

[9] T. Watkins, “Draft roadmap for next generation air monitoring,” *Environmental Protection Agency*, 2013.

[10] F. Fazel, M. Fazel, and M. Stojanovic, “Random access sensor networks: Field reconstruction from incomplete data,” in *2012 Information Theory and Applications Workshop*. IEEE, 2012, pp. 300–305.

[11] J. Matamoros, F. Fabbri, C. Anton-Haro, and D. Dardari, “On the estimation of randomly sampled 2d spatial fields under bandwidth constraints,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4184–4192, 2011.

[12] F. Restuccia, N. Ghosh, S. Bhattacharjee, S. K. Das, and T. Melodia, “Quality of information in mobile crowdsensing: Survey and research challenges,” *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 4, p. 34, 2017.

[13] X. Fang and I. Bate, “Using multi-parameters for calibration of low-cost sensors in urban environment,” in *Proceedings of the 2017 International Conference on Embedded Wireless Systems and Networks*. Junction Publishing, 2017, pp. 1–11.

[14] E. Nurellari, D. McLernon, and M. Ghogho, “A secure optimum distributed detection scheme in under-attack wireless sensor networks,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 325–337, 2018.

[15] B. Kailkhura, S. Brahma, and P. K. Varshney, “Data falsification attacks on consensus-based detection systems,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 145–158, 2017.

[16] T. R. Karl, A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, and H.-M. Zhang, “Possible artifacts of data biases in the recent global surface warming hiatus,” *Science*, vol. 348, no. 6242, pp. 1469–1472, 2015.

[17] E. Miluzzo, N. D. Lane, A. T. Campbell, and R. Olfati-Saber, “Calibree: A self-calibration system for mobile sensor networks,” in *International Conference on Distributed Computing in Sensor Systems*. Springer, 2008, pp. 314–331.

[18] C. Dorffer, M. Puigt, G. Delmaire, and G. Roussel, “Informed nonnegative matrix factorization methods for
mobile sensor network calibration,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 4, pp. 667–682, 2018. 2, 3, 5

[19] Ç. Bilen, G. Puy, R. Gribonval, and L. Daudet, “Convex optimization approaches for blind sensor calibration using sparsity,” *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4847–4856, 2014. 2

[20] B.-T. Lee, S.-C. Son, and K. Kang, “A blind calibration scheme exploiting mutual calibration relationships for a dense mobile sensor network,” *IEEE Sensors Journal*, vol. 14, no. 5, pp. 1518–1526, 2014. 2

[21] M. Üney, B. Mulgrew, and D. E. Clark, “Latent parameter estimation in fusion networks using separable likelihoods,” *IEEE Transactions on signal and information processing over networks*, vol. 4, no. 4, pp. 752–768, 2018. 2

[22] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, “False data injection attacks against state estimation in wireless sensor networks,” in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 5967–5972. 2

[23] W. Hashlamoun, S. Brahma, and P. K. Varshney, “Mitigation of byzantine attacks on distributed detection systems using audit bits,” *IEEE Transactions on Signal Processing and Network* ., vol. 4, no. 1, pp. 18–32, 2018. 2

[24] ———, “Audit bit based distributed bayesian detection in the presence of byzantines,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 4, pp. 643–655, 2018. 2

[25] V. S. S. Nadendla, Y. S. Han, and P. K. Varshney, “Distributed inference with m-ary quantized data in the presence of byzantine attacks,” *IEEE Transactions on Signal Processing*, vol. 62, no. 10, pp. 2681–2695, 2014. 2

[26] Q. Xiang, J. Zhang, I. Nevat, and P. Zhang, “A trust-based mixture of gaussian processes model for reliable regression in participatory sensing.” in *IJCAI*, 2017, pp. 3866–3872. 2, 4

[27] P. Zhang, I. Nevat, G. W. Peters, F. Septier, and M. A. Osborne, “Spatial field reconstruction and sensor selection in heterogeneous sensor networks with stochastic energy harvesting,” *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2245–2257, 2018. 3

[28] G. W. Peters, I. Nevat, and T. Matsui, “How to Utilize Sensor Network Data to Efficiently Perform Model Calibration and Spatial Field Reconstruction,” in *Modern Methodology and Applications in Spatial-Temporal Modeling*. Springer, 2015, pp. 25–62. 3

[29] I. Nevat, G. W. Peters, F. Septier, and T. Matsui, “Estimation of Spatially Correlated Random Fields in Heterogeneous Wireless Sensor Networks,” *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2597–2609, 2015. 3, 4

[30] I. Nevat, G. W. Peters, and I. B. Collings, “Random Field Reconstruction With Quantization in Wireless Sensor Networks,” *IEEE Transactions on Signal Processing*, vol. 61, pp. 6020–6033, 2013. 3, 4

[31] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, “Sparse mobile crowdsensing: challenges and opportunities,” *IEEE Communications Magazine*, vol. 54, no. 7, pp. 161–167, 2016. 3

[32] D. Gu and H. Hu, “Spatial Gaussian Process Regression With Mobile Sensor Networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1279–1290, 2012. 4

[33] Y. Zhang, T. N. Hoang, K. H. Low, and M. Kankanhalli, “Near-optimal active learning of multi-output Gaussian processes,” *arXiv preprint arXiv:1511.06891*, 2015. 4

[34] P. A. Plonski, P. Tokekar, and V. Isler, “Energy-efficient Path Planning for Solar-powered Mobile Robots,” *Journal of Field Robotics*, vol. 30, no. 4, pp. 583–601, 2013. 4

[35] C. K. Ling, K. H. Low, and P. Jaillet, “Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond,” *arXiv preprint arXiv:1511.06890*, 2015. 4
[36] I. Nevat, G. W. Peters, and I. B. Collings, “Location-aware cooperative spectrum sensing via Gaussian Processes,” in *Communications Theory Workshop (AusCTW), 2012 Australian*. IEEE, 2012, pp. 19–24.

[37] P. Agrawal and N. Patwari, “Correlated link shadow fading in multi-hop wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4024–4036, 2009.

[38] C. Fonseca and H. Ferreira, “Stability and contagion measures for spatial extreme value analyses,” *arXiv preprint arXiv:1206.1228*, 2012.

[39] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[40] R. J. Adler and J. E. Taylor, *Random fields and geometry*. Springer Verlag, 2007, vol. 115.

[41] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, “A Tutorial on the Cross-Entropy Method,” *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, Feb 2005. [Online]. Available: http://link.springer.com/10.1007/s10479-005-5724-z

[42] R. Rubinstein, “The cross-entropy method for combinatorial and continuous optimization,” *Methodology and computing in applied probability*, vol. 1, no. 2, pp. 127–190, 1999.

[43] J. A. Bilmes et al., “A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models,” 1998.

[44] J. Besag, “On the statistical analysis of dirty pictures,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 48, no. 3, pp. 259–279, 1986.

### Appendix A

**Proof of Theorems**

**A. Proof of Theorem 1**

**Proof.** For $n = 1, \ldots, N$, the log model likelihood given $\mathbf{f} := (f(x_n))_{1:N}$ is given by

$$
\log p(y_{n,1:M_n} | \mathbf{f}, \psi) = \sum_{n=1}^{M_n} -\frac{1}{2} \log 2\pi - \frac{1}{2} \log a_n^{-2} - \frac{(y_{n,m} - a_n f_n - b_n)^2}{2a_n^{-2}s_n^{-2}} \\
= -\frac{M_n}{2} \log 2\pi - \frac{M_n}{2} \log a_n^{-2} - \frac{s_n}{2a_n^{-2}s_n^{-2}} \\
+ \frac{(a_n f_n + b_n)g_n}{a_n^{-2}s_n^{-2}} - \frac{M_n(a_n f_n + b_n)^2}{2a_n^{-2}s_n^{-2}},
$$

(29)
which depends on \( y_{n,1:M_n} \) only through the statistics \( g_n \) and \( s_n \). Therefore, \((g, s)\) are sufficient for \((f, \psi)\). The joint density of \((y, f)\) conditional on \( \psi \) is given by

\[
\log p(y, f | \psi) = \log p(y | f) + \log p(f | \psi)
\]

\[
= \sum_{n=1}^{N} -\frac{M_n}{2} \log 2\pi - \frac{M_n}{2} \log a_n^2 \varsigma_n^2 - \frac{s_n}{2a_n^2 \varsigma_n^2} + \frac{(a_n f_n + b_n) g_n}{a_n^2 \varsigma_n^2}
\]

\[
- \frac{M_n (a_n f_n + b_n)^2}{2a_n^2 \varsigma_n^2} - \frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} (f - \mu)^T \mathbf{C}^{-1} (f - \mu) \tag{30}
\]

\[
= -\frac{1}{2} \left[ (N + \text{tr}(\mathbf{M})) \log 2\pi + \text{tr}(\mathbf{M} \log(\varsigma^2 \mathbf{A}^2)) + \log |\mathbf{C}|ight.
\]

\[
+ \varsigma^{-2} 1^T \mathbf{A}^{-2} s + \mu^T \mathbf{C}^{-1} \mu + \varsigma^{-2} \mathbf{b}^T \mathbf{M} \mathbf{A}^2 \mathbf{b}
\]

\[
-2\varsigma^{-2} \mathbf{g}^T \mathbf{A}^{-2} \mathbf{b} - \gamma^T \mathbf{Z}^{-1} \gamma + (f - \mathbf{Z}^{-1} \gamma)^T \mathbf{Z} (f - \mathbf{Z}^{-1} \gamma) \right],
\]

where \( \gamma = \varsigma^{-2} \mathbf{M} \mathbf{g} + \mathbf{C}^{-1} \mu, \mathbf{Z} = \varsigma^{-2} \mathbf{M} + \mathbf{C}^{-1} \). Integrating over \( f \), we deduce that,

\[
\log p(y | \psi)
\]

\[
= \log \left[ \int p(y, f | \psi) df \right]
\]

\[
= -\frac{1}{2} \left[ \text{tr}(\mathbf{M}) \log 2\pi + \text{tr}(\mathbf{M} \log(\varsigma^2 \mathbf{A}^2)) + \log |\mathbf{C}| + \log |\mathbf{Z}|ight.
\]

\[
+ \varsigma^{-2} 1^T \mathbf{A}^{-2} s + \mu^T \mathbf{C}^{-1} \mu + \varsigma^{-2} \mathbf{b}^T \mathbf{M} \mathbf{A}^2 \mathbf{b}
\]

\[
-2\varsigma^{-2} \mathbf{g}^T \mathbf{A}^{-2} \mathbf{b} - \gamma^T \mathbf{Z}^{-1} \gamma \right]. \tag{31}
\]

We have by Woodbury’s formula that,

\[
\mathbf{Z}^{-1} = \varsigma^2 \mathbf{M}^{-1} - \varsigma^4 \mathbf{M}^{-1} (\mathbf{C} + \varsigma^2 \mathbf{M}^{-1})^{-1} \mathbf{M}^{-1} \tag{32}
\]

\[
= \mathbf{C} - \mathbf{C} (\mathbf{C} + \varsigma^2 \mathbf{M}^{-1})^{-1} \mathbf{C}, \tag{33}
\]

\[
\log |\mathbf{Z}| = \log |\varsigma^{-2} \mathbf{M}| - \log |\mathbf{C}| + \log |\mathbf{C} + \varsigma^2 \mathbf{M}^{-1}|. \tag{34}
\]

Substituting (32), (33), (34) into (31), and recall that \( \mathbf{g} := \mathbf{A}^{-1} (\mathbf{M}^{-1} \mathbf{g} - \mathbf{b}), \mathbf{Y} := \mathbf{C} + \varsigma^2 \mathbf{M}^{-1} \), we have

\[
\log p(y | \psi)
\]

\[
= -\frac{1}{2} \left[ \text{tr}(\mathbf{M}) \log 2\pi + \text{tr}(\mathbf{M} \log(\varsigma^2 \mathbf{A}^2)) - \log |\varsigma^2 \mathbf{M}^{-1}| + \log |\mathbf{Y}|ight.
\]

\[
+ \varsigma^{-2} 1^T \mathbf{A}^{-2} s - \varsigma^{-2} \mathbf{g}^T \mathbf{M}^{-1} \mathbf{A}^{-2} \mathbf{g} + (\mathbf{g} - \mu)^T \mathbf{Y}^{-1} (\mathbf{g} - \mu) \right]. \tag{35}
\]

Notice that after integrating out \( f \), \((g, s)\) are still sufficient for \( \psi \).

By Bayes’ rule,

\[
p(\psi | y) = \frac{p(y | \psi) \pi(\psi)}{p(y)}, \tag{36}
\]
where \( p(y) = \int p(y|\psi)\pi(\psi)d\psi \) is the normalizing constant that is analytically intractable. The proof is now complete.

### B. Proof of Theorem 2

From (30) we see that

\[
\log p(y, f|\psi) = q_1(\psi) + q_2(s, \psi) + q_3(g, f, \psi),
\]

where \( q_1, q_2, q_3 \) are some functions of the corresponding parameters. Similarly, we can deduce that

\[
\log p(y, f; f_*|\psi) = q_1(\psi) + q_2(s, \psi) + q_3(g, [f, f_*], \psi).
\]

Therefore,

\[
p(f_*|y, \psi) = \frac{p(y, f_*|\psi)}{p(y|\psi)} = \frac{\int p(y, f, f_*|\psi)df}{\int p(y, f|\psi)df}
\]

\[
= \frac{\int \exp(q_3(g, f, \psi))df}{\int \exp(q_3(g, [f, f_*], \psi))df}.
\]

Thus, we have deduced that \( p(f_*|y, \psi) \) depends on \( y \) only through \( g \) (or equivalently, \( \tilde{g} \)), i.e.

\[
p(f_*|y, \psi) = p(f_*|g, \psi) = p(f_*|\tilde{g}, \psi).
\]

We also have that conditional on \( \psi \), \((\tilde{g}, f_*)\) are jointly Gaussian, with \( E[f_*|\psi] = \mu_* \), \( E[\tilde{g}|\psi] = \mu \), \( \text{Cov}[f_*|\psi] = C_* \), \( \text{Cov}[\tilde{g}, f_*|\psi] = k_* \), \( \text{Cov}[\tilde{g}|\psi] = \Upsilon \), which can be easily verified. Thus, we have that \( (f_*|y, \psi) \sim N(\bar{f}_*, \sigma_*^2) \),

\[
\bar{f}_* = \mu_* + k_*^T \Upsilon^{-1} (\tilde{g} - \mu),
\]

\[
\sigma_*^2 = C_* - k_*^T \Upsilon^{-1} k_*.
\]

The distribution of \( (f_*, \psi|y) \) is given by

\[
p(f_*, \psi|y) = p(f_*|y, \psi)p(\psi|y),
\]

and the posterior predictive distribution \( (f_*|y) \) is obtained by marginalizing over \( \psi \),

\[
p(f_*|y) = \int p(f_*|y, \psi)p(\psi|y)d\psi.
\]

The proof is now complete.
C. Proof of Theorem 3

Proof. Firstly, we claim that any linear estimator must have the form $h(y) = w^T \bar{g} + b$, where $w \in \mathbb{R}^N, b \in \mathbb{R}$. Due to the linearity of $h$ in observations $(y_{n,m})_{n=1:N,m=1:M}$, $s$ is not involved. For $n = 1, \ldots, N$, the weights of $(y_{n,m})_{m=1:M}$ must be the same due to symmetry. This proves the claim.

Under quadratic loss, for $h \in \mathcal{H}$, $R[\Pi, h]$ is given by

$$R[\Pi, h] = \mathbb{E} \left[ \left( w^T \bar{g} + b - f_\ast \right)^2 \right]$$

Differentiating $R[\Pi, h]$ with respect to $w$ and $b$, we get

$$\frac{\partial R[\Pi, h]}{\partial w} = 2w^T \mathbb{E} [\bar{g} \bar{g}^T] \bar{g} + 2b \mathbb{E} [\bar{g}] - 2w^T \mathbb{E} [f_\ast y] - 2b \mathbb{E} [f_\ast].$$

$$\frac{\partial R[\Pi, h]}{\partial b} = 2b + 2w^T \mathbb{E} [\bar{g}] - 2\mathbb{E} [f_\ast].$$

Setting the partial derivatives to 0 and solving the equations gives the optimal weight vector and intercept,

$$\hat{b} = \mathbb{E} [f_\ast] - \hat{w}^T \mathbb{E} [\bar{g}],$$

$$\hat{w} = (\mathbb{E} [\bar{g} \bar{g}^T] - \mathbb{E} [\bar{g}] \mathbb{E} [\bar{g}]^T)^{-1} (\mathbb{E} [f_\ast y] - \mathbb{E} [f_\ast] \mathbb{E} [\bar{g}])$$

$$= \text{Cov}[\bar{g}]^{-1} \text{Cov}[\bar{g}, f_\ast].$$

One can verify that $(\hat{w}, \hat{b})$ is indeed minimizing the Bayes risk. Hence,

$$\hat{h}_\text{S-BLUE}(y) = \hat{w}^T \bar{g} + \hat{b}$$

$$= \mathbb{E} [f_\ast] + \text{Cov}[\bar{g}, f_\ast]^T \text{Cov}[\bar{g}]^{-1} (\bar{g} - \mathbb{E} [\bar{g}]).$$

The terms $\mathbb{E} [f_\ast], \mathbb{E} [\bar{g}], \text{Cov}[\bar{g}, f_\ast], \text{Cov}[\bar{g}]$ can all be expressed in closed-form. The closed-form expressions and the details of the computation are given below. Let $\odot$ denote matrix entry-wise
\[ E[f_s] = \mu_s, \]  
\[ E[g] = E[E[g|\psi]] = \text{diag}(E[a]) \mu + E[b], \]  
\[ \text{Cov}[g,f_s] = E[\text{Cov}[g,f_s|\psi]] + E[\text{Cov}[g|\psi] \mu_s] = E[Ak_s] = \text{diag}(E[a]) k_s, \]  
\[ \text{Cov}[g] = E[\text{Cov}[g|\psi]] + E[\text{Cov}[g|\psi]] = E[ACA + \varsigma^2 M^{-1} A^2] + \text{Cov}[A \mu + b] = E[a a^T] \odot (C + \varsigma^2 M^{-1} + \mu \mu^T) \]  
\[ + \text{diag}(\mu) (E[a b^T] + E[a b^T]^T) + E[b b^T] - E[g]E[g]^T, \]  

In the terms \( E[a], E[b], E[aa^T], E[bb^T], E[ab^T], \) the expectations are evaluated entry-wise. For example, entries of \( E[a] \) are given by

\[ E[a_n] = \sum_{k=0}^{K} q_k^{(n)} E[a_n | Z_n = k] = q_0^{(n)} + \sum_{k=1}^{K} q_k^{(n)} E[a_n | Z_n = k]. \]  

Entries of \( E[aa^T] \) are given by

\[ E[a_i a_j] = \begin{cases} \sum_{k=0}^{K} q_k^{(i)} E[a_i^2 | Z_i = k] & \text{if } i = j, \\ \sum_{k=0}^{K} \sum_{k'} q_k^{(i)} q_{k'}^{(j)} E[a_i | Z_i = k] E[a_j | Z_j = k'] & \text{if } i \neq j. \end{cases} \]  

Entries of \( E[b], E[bb^T] \) and \( E[ab^T] \) can be evaluated similarly. With the above equations, we are able to evaluate \( \hat{h}_{S-BLUE} \) efficiently. \( \Box \)