SPARSE ESTIMATION: AN MMSE APPROACH

TONGYAO PANG AND ZUOWEI SHEN

Abstract. This paper is to estimate parameters with sparse prior via the minimum mean
square error (MMSE) approach. We model the sparsity by the Bernoulli-uniform prior. The
MMSE estimator gives the posterior mean of the parameter to be estimated. However, its com-
putation involves multiple integration of many variables that is hard to implement numerically.

In order to overcome this difficulty, we develop a coordinate minimization algorithm to
approximate the MMSE estimator for arbitrary given prior. We connect this algorithm to a
variational model and establish a comprehensive convergence analysis. The algorithm converges
to a special stationary point of the variational model, which attains the minimum of the mean
square error at each coordinate when others are fixed. Then, this general algorithm is applied
to the Bernoulli-uniform sparse prior and leads to a stable estimator that provides a good balance
between sparsity and unbiasedness. The advantages of our sparsity model and algorithm over
other approaches (e.g., the maximum a posterior approaches) are analyzed in detail and further
demonstrated by numerical simulations. The applications of the general theory and algorithm
developed here go beyond the sparse estimation.

1. Introduction

This paper develops an algorithm and its associated model and theory on Bayesian estimation.
Consider a linear regression model:
\begin{equation}
Z = \mathbf{A}X + \varepsilon,
\end{equation}
where \( z \in \mathbb{R}^n \) is the observation, \( \mathbf{A} \in \mathbb{R}^{n \times p} \) is the design matrix, \( x \in \mathbb{R}^p \) is the parameter of
interest, and \( \varepsilon \in \mathbb{R}^n \) is the Gaussian noise with mean zero and variance \( \sigma^2 I \). The linear regression
is to estimate the parameter \( x \) from the given observation \( z \).

In Bayesian statistics, the underlying truth is represented as a random variable that follows some
prior distribution. Without loss of generality, we assume that \( x_i (i = 1, 2, \cdots, n) \) are independent
and identically distributed, and follow the distribution \( p(x) : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\} \).

Our aim is to find an efficient estimator \( \mathbf{T}(z) : \mathbb{R}^n \rightarrow \mathbb{R}^p \), for the parameter \( x \) from the
contaminated observation \( z \). Minimizing different Bayesian risk function results in a variety of
estimators. We will concentrate on the minimum mean square error (MMSE) estimator which
minimizes a quadratic cost function:
\begin{equation}
\mathcal{T}^{\text{MMSE}} = \arg \min_{g: \mathbb{R}^n \rightarrow \mathbb{R}^p} \mathbb{E}_{(x,z)}(x - g(z))^2.
\end{equation}
The MMSE estimator gains its popularity since it is stable in observation and balances the trade-
off between variances and bias well. In addition, the MMSE estimator is given in an explicit form
that is the mean of the posterior distribution:
\begin{equation}
\mathcal{T}^{\text{MMSE}} (z) = \mathbb{E}_{x|z} (x|z) = \frac{\int x \prod_{j=1}^{p} p(x_j) \exp(-\frac{\|Ax - z\|^2_{2}}{2\sigma^2}) dx}{\int \prod_{j=1}^{p} p(x_j) \exp(-\frac{\|Ax - z\|^2_{2}}{2\sigma^2}) dx}.
\end{equation}
However, the posterior mean is usually difficult to compute due to the involvement of multiple
integrals of many variables. In this paper, we propose an iterative algorithm to approximate it.
We start from the singleton case
\begin{equation}
z = x + \varepsilon,
\end{equation}
where \( x, z \in \mathbb{R} \) and \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \). The corresponding singleton MMSE estimator is given by
\begin{equation}
\mathbf{s}^{\text{MMSE}}_x = \mathbb{E}_{x|x} (x|z),
\end{equation}
which only involves single integrals and is easy to compute. Furthermore, when the columns of $A$ are orthogonal, the MMSE estimator (1.3) is separable and reduces to the singleton case:

$$
\hat{x}_i^{\text{MMSE}} = \frac{\int x_i \prod_{j=1}^{p} p(x_j) \exp\left(-\frac{||Ax-z||^2}{2\sigma^2}\right)dx}{\int \prod_{j=1}^{p} p(x_j) \exp\left(-\frac{||Ax-z||^2}{2\sigma^2}\right)dx} = \frac{\int x_i \prod_{j=1}^{p} \left\{ p(x_j) \exp\left(-\frac{(x_j - a_j^T z/||a_j||^2_j)^2}{2\sigma^2/||a_j||^2_j}\right)\right\}dx}{\int \prod_{j=1}^{p} \left\{ p(x_j) \exp\left(-\frac{(x_j - a_j^T z/||a_j||^2_j)^2}{2\sigma^2/||a_j||^2_j}\right)\right\}dx}
$$

which avoids multiple integration and can be computed easily. However, in the general case, the MMSE estimator (1.3) is still hard to compute. Instead of calculating the involved multiple integrals directly, we convert the approximation of the MMSE estimator to the problem of finding the minimizer of

$$\min_{y \in \mathbb{R}^p} \mathbb{E}(x, z)(x - y)^2. \quad (1.6)$$

Next, we develop an iterative algorithm that only refines one component each time to reduce the mean square error in (1.6). Given an index $i_k$ at iteration $k$, we minimize the mean square error along the direction of $x_{i_k}^k$ by fixing the remaining components:

$$
x_{i_k}^{k+1} = \arg \min_{x \in \mathbb{R}} \mathbb{E}(x_{i_k}, z | x_{i_k}^k) (x_{i_k} - x)^2 = \mathbb{E}(x_{i_k} | z, x_{j \neq i_k}^k) = S_{\sigma ||a_i||^2} (a_i^T z/||a_i||^2_i), \quad i = 1, 2, \ldots, p, \quad (1.7)
$$

where $a_j$ is the $j$-th column of matrix $A$. This iterative algorithm is called the coordinate minimization algorithm in optimization. At each iteration, it only involves single integration which is easy to implement.

The remaining question is whether this algorithm converges. It is hard to analyze the convergence of the coordinate minimization algorithm from the perspective of minimizing mean square error in (1.6). Note that the singleton MMSE estimator $S_{\sigma}^{\text{MMSE}}(z)$ corresponds to the unique solution of the following strongly convex variational model

$$\min_{x \in \mathbb{R}} \frac{1}{2} (x - z)^2 + \varphi_{\sigma}^{\text{MMSE}}(x), \quad (1.8)$$

where

$$\varphi_{\sigma}^{\text{MMSE}}(x) = \int x r_{\sigma}^{\text{MMSE}}(u - u)du, \quad (1.9)$$

and $r_{\sigma}^{\text{MMSE}}$ is the inverse function of $S_{\sigma}^{\text{MMSE}}$. Motivated by this fact, we use the following variational model

$$\min_{x \in \mathbb{R}^m} \frac{1}{2} ||Ax - z||^2 + \sum_{i=1}^{m} ||a_i||^2_2 \varphi_{\sigma ||a_i||^2_i}(x_i), \quad (1.10)$$

to analyze the convergence of our algorithm. As we will show, the coordinate minimization algorithm for solving (1.10) is the same as the iteration (1.7). Using the available theory developed in the field of optimization recently, the variational model (1.10) helps to prove the convergence of (1.7). Basically, we will prove that the sequences generated by (1.7) converges to a stationary point of the variational model (1.10). Moreover, it attains the minimum of the mean square error in (1.6) along each coordinate when the others are fixed. Specifically, the algorithm converges to the MMSE estimator when the columns of $A$ are orthogonal.
While the algorithm and theory developed here apply for a wide variety of prior distributions, we illustrate their power by three special prior distributions, namely, Gaussian, Gaussian mixture and sparse prior. The Gaussian prior is commonly-used since it is simple and natural. Gaussian mixture goes beyond Gaussian and can approximate any continuous distribution with arbitrary accuracy. Sparse prior is popular in regression model as it helps to reduce the model complexity. Sparse prior is also used in many other scenarios beyond regression, such as, compressed sensing [8, 11], image analysis and restorations [6, 7]. The sparse prior distribution chosen here takes the form of

$$p(x) = p_0 \delta(x) + \frac{1 - p_0}{2(U - L)} \mathbb{I}_{[-U, -L] \cup [L, U]}(x),$$

where $\delta(x)$ stands for the Dirac-delta function, and $\mathbb{I}_S(x)$ means the indicator function of the set $S$. It consists of a Dirac-delta distribution at the origin and an uniform distribution on $[-U, -L] \cup [L, U]$. The greater the value of $p_0$ is, the larger the portion of zero components is, that is, the sparser the underlying truth is. The uniform distribution on $[-U, -L] \cup [L, U]$ reflects a high degree of uncertainty in the non-zero value. The uniform distribution assumption also implies the underlying truth is bounded, which is the case for the majority of the applications. Finally, when $L > 0$, it means that the non-zero coefficients are significant to some degree. We call (1.11) the Bernoulli-uniform sparse prior as it can be interpreted as the composition of a Bernoulli distribution and a uniform distribution.

Since the MMSE estimator (1.3) involves multiple integrals and is hard to compute, many approaches have been developed to overcome this difficulty in last decades. One way is to constrain the estimator to be linear, that is, to minimize the mean square error among all possible linear maps. The resulting estimator is called the linear minimum mean square error (LMMSE) estimator. The LMMSE estimator for the problem (1.1) is summarized by Bayesian Gauss-Markov theorem [16]. It gives an explicit form of the LMMSE estimator, which depends only on the mean and variance of the prior and noise distribution. The LMMSE estimator will not be optimal unless the conditional expectation $\mathbb{E}_u(u|x)$ happens to be linear.

Another way for a tractable Bayesian estimate is to find the maximum a posteriori (MAP) point estimate instead:

$$x^{MAP} = T^{MAP}(z) := \arg \max_{x \in \mathbb{R}^p} p(x|z) = \arg \max_{x \in \mathbb{R}^p} \frac{p(z|x)p(x)}{p(z)},$$

$$= \arg \max_{x \in \mathbb{R}^p} \log p(z|x) + \log p(x)$$

$$= \arg \min_{x \in \mathbb{R}^p} \frac{1}{2} \|Ax - z\|^2_2 + \varphi(x),$$

where $\varphi(x) = -\sigma^2 \log p(x)$ is called the penalty function. The discarding of the term $p(z)$ is because $z$ has been observed and thus $p(z)$ is constant. Actually, the MAP estimator $T^{MAP}$ can also be treated as

$$T^{MAP} = \arg \min_{g: \mathbb{R}^n \rightarrow \mathbb{R}^p} \mathbb{E}(x,z)(-\delta_0(x - g(z))).$$

That is, the MAP estimator results from a minus delta loss function while the MMSE estimator from a quadratic loss function. The MAP estimator may be unstable for some prior as it gives a single point estimate while the MMSE estimator is more stable as it averages all potential points under the posterior distribution. The advantage of MAP method is the existence of the variational model (1.14). Benefiting from that, the MAP estimator can be solved by iterative algorithms, such as coordinate descent method [15] and proximal alternating minimization method [2]. However, the penalty term under some prior, for example, the Gaussian mixture prior and the sparse prior (1.11), may fail to be smooth, convex or coercive, hence it is usually hard to establish a convergence result. Even if the algorithm has shown to be convergent, it often converges to a stationary point. For non-smooth and non-convex penalty, how the stationary point performs in terms of maximizing the posterior probability is unclear.
As we have extensively studied the Gaussian, Gaussian mixture and sparse prior (1.11), the comparison of LMMSE, MAP and our method under these three prior distributions are summarized here. For Gaussian prior, all three approaches reduce to one model and one solution; and all of them attain the MMSE estimate. For Gaussian mixture prior, since the posterior distribution is still Gaussian mixture, it is easy to obtain the MMSE estimator, which is the posterior mean. Our method which uses the singleton mean iteratively is also easy to perform. In contrast, the MAP approach is much harder to implement since it seeks the maximum of the Gaussian mixture density function which is difficult to find. In comparison with the LMMSE estimator, the numerical results show that our method outperforms it generally in the approximation of the MMSE estimator.

Finally, we consider the Bernoulli-uniform sparse prior for the problem (1.1). The MAP estimator under the Bernoulli-uniform sparse prior for the singleton problem (1.4) is obtained by simple hard-thresholding. As pointed out by many [see 12], the hard-thresholding suffers from the drawback of instability. To remedy this drawback, the continuous soft-thresholding [9] is proposed. However, it introduces bias. The hard-thresholding is unbiased but unstable, while the soft one is stable but biased. Consequently, many efforts have been made to combine the advantages of the hard and soft thresholding and avoid their disadvantages. These efforts are initiated by [12] and the aim is to seek estimators that simultaneously satisfy the conditions of “sparsity”, “continuity” and “unbiasedness”. The SCAD thresholding [12] and MCP thresholding [27] are two examples of these efforts. The key idea is to construct thresholding by interpolating between the hard and soft thresholding so that it satisfies those three conditions. The constructed thresholding is raised to a variational model that can be interpreted as a MAP model with a different prior from the Bernoulli-uniform sparse prior. The “sparsity” and “unbiasedness” is actually the result of negotiation between variance and bias. Thus, the conditions of “sparsity”, “continuity” and “unbiasedness” require a stable way to balance variance and bias. The singleton MMSE estimator achieves the goals of stability and balance between variance and bias naturally. On one hand, it is smooth as we will show; on the other hand, since the mean square error can be written as the sum of variance and squared bias of the estimate, the singleton MMSE estimator attains balance between variance and bias. For the general problem (1.1) with the sparse prior, all of SCAD, MCP and our method adopt iterative algorithm relating to the corresponding variational models. However, our iterative algorithm always converges to an estimate that attains the minimum of the mean square error along single coordinate while the iterative algorithms for the SCAD and MCP variational model may diverge as their models are not coercive. Even if they converge, it is still unclear how the obtained estimates are connected to the goal of maximizing the posterior probability density. Finally, the numerical results demonstrate the advantages of our method over SCAD and MCP methods in terms of various loss functions. As for the LMMSE estimator, it is not listed for numerical comparison as it fails to enhance sparsity.

The rest of the paper is organized as follows. In section 2, we first introduce MMSE estimator from a singleton case and analyze its good properties. We next build up a connection between the singleton MMSE estimator and some variational model. At the end of this section, we give three examples of the prior distributions, i.e., Gaussian, Gaussian mixture and the Bernoulli-uniform sparse prior for illustration. In section 3, we propose to approximate the MMSE estimator by the cyclic coordinate minimization algorithm in the general case and give the main theorem that establishes the convergence of our algorithm. The numerical experiments follow to show the efficiency of our algorithm. The proof of our main theorem is given in section 4. Finally, the paper concludes in section 5.

2. Singleton MMSE estimator

For the general regression problem (1.1), the iterative scheme (1.7) approximates the MMSE estimator by performing a singleton MMSE estimation at each iteration. In order to analyze the convergence of (1.7), we ultimately need to understand the various properties of the singleton MMSE estimator and its corresponding variational model.
In this paper, we use the notation $\delta_{c_0}(x)$ to stand for the Dirac-delta function centered at point $c_0 \in \mathbb{R}$, and $\mathbb{I}_S(x)$ for the indicator function of set $S$. The Dirac-delta function $\delta_{c_0}(x)$ is roughly defined as
\[
\delta_{c_0}(x) = \begin{cases} 
+\infty, & x = c_0, \\
0, & x \neq c_0,
\end{cases}
\]
and is constrained to satisfy $\int \delta_{c_0}(x) = 1$ [see 23, for a rigorous definition]. We denote the random variable with a lower case letter in plain typeface (e.g., $x$ is a random variable) and the values it can take on with lower case script letters (e.g., $x_1$ is a possible value of $x$). The closure of a set $S$ is denoted by $\overline{S}$.

2.1. Singleton MMSE estimator and its properties. To set up our platform, we assume that the prior distributions satisfy the following conditions.

Assumption 1. Assume the prior distribution $p(x) : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ satisfies that

(i) it takes the form of
\[
(2.1) \quad p(x) = \sum_{i=1}^{d_1} w_{i,1} \delta_{c_i}(x) + \sum_{j=1}^{d_2} w_{j,2} h_j(x), \quad d_2 > 0, \quad \sum_{i=1}^{d_1} w_{i,1} + \sum_{j=1}^{d_2} w_{j,2} = 1,
\]
where $\delta_{c_i}(x) \ (i = 1, 2, \ldots, d_1)$ are the Dirac delta functions and $h_j(x) : \mathbb{R} \rightarrow \mathbb{R} \ (j = 1, 2, \ldots, d_2)$ are bounded probability density functions;

(ii) its mean, i.e. $\int xp(x)dx$, exists and is finite.

For the problem (1.4), when the prior distribution $p(x)$ satisfies Assumption 1, it is well known that the singleton MMSE estimator is (1.5) [see 16]. Since our theory and algorithm start from this basic fact, we summarize the derivation in the following.

The singleton MMSE estimator for (1.4) refers to the estimator which attains the minimum of the mean square error among all possible maps from $\mathbb{R}$ to $\mathbb{R}$:
\[
(2.2) \quad S_{\sigma}^{\text{MMSE}} = \arg\min_{y \in \mathbb{R}} \mathbb{E}_{(x,z)}((x-g(z))^2)
= \arg\min_{y \in \mathbb{R}} \int ((x - g(z))^2 p(x,z) dx dz
= \arg\min_{y \in \mathbb{R}} \int p(z) \left( \int (x^2 - 2zx + z^2) p(x|z) dx \right) dy
\]
where $\sigma$ is the standard variance of the noise. It is equivalent to finding $S_{\sigma}^{\text{MMSE}}(z) = \arg\min_{y \in \mathbb{R}} \int (x^2 - 2xy + y^2)p(x|z) dx$ for every observation $z$. As the noise follows Gaussian distribution, the conditional distribution $p(x|z)$ takes the form of
\[
p(x|z) = \frac{p(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right)}{\int p(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dx}.
\]
When Assumption 1 is satisfied, the integral $\int x^n p(x|z) dx$ exists and is finite for any $n \in \mathbb{N}$. Then, by dominated convergence theorem, we obtain
\[
\frac{d}{dy} \int (x^2 - 2xy + y^2)p(x|z) dx = 2 \int (y - x)p(x|z) dy
= 2y \int p(x|z) dx - 2 \int xp(x|z) dx = 2y - 2 \int xp(x|z) dx.
\]
Setting the differential to zero, we get a closed form of the singleton MMSE estimator:
\[
S_{\sigma}^{\text{MMSE}}(z) = \mathbb{E}_{x}(x|z) = \int xp(x|z) dx = \frac{\int xp(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dx}{\int p(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dx}.
\]
The single integrals in (2.3) are easy to compute for every observed $z$, and so is the singleton MMSE estimator $S_{\sigma}^{\text{MMSE}}(z)$. In contrast, the MMSE estimator (1.3) for the general problem (1.1), which can be derived in a similar way as the singleton MMSE estimator, is much harder to compute numerically due to the involvement of the multiple integrals. Consequently, we propose an iterative algorithm (1.7) to approximate the MMSE estimator (1.3) by a sequence of singleton MMSE estimators. For the convergence analysis of the algorithm (1.7), we study the properties of the singleton MMSE estimator in the following.
In the form of (2.3), the singleton MMSE estimator is proved to be analytic in Proposition (2.1). This means that the singleton MMSE estimator provides a stable estimation for the parameter of interest.

**Proposition 2.1.** Assume \( p(x) \) satisfies Assumption 1. The singleton MMSE estimator \( S_{\sigma}^{MMSE}(z) \) (2.3) is analytic.

**Proof.** Since the denominator of (2.3) is non-zero, it is enough to show both the numerator
\[
\int xp(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right)dx
\]
and the denominator
\[
\int p(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right)dx
\]
are analytic in terms of \( z \). We only prove the numerator is analytic in the following, as the proof for the denominator is similar.

Let
\[
f(z) = \int xp(x)\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right)dx.
\]
It can be rewritten as
\[
f(z) = \exp\left(-\frac{z^2}{2\sigma^2}\right)h(z), \quad h(z) = \int xp(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\frac{xz}{\sigma^2}\right)dx.
\]
The Taylor series for \( \exp\left(\frac{zx}{\sigma^2}\right) \) with respect to \( x \) is
\[
\exp\left(\frac{zx}{\sigma^2}\right) = \sum_{k=1}^{+\infty} \frac{z^k}{\sigma^{2k}k!} x^k,
\]
which converges for all \( x \in \mathbb{R} \). Denote the first \( n \) terms of the Taylor series by \( g_n(x) = \sum_{k=1}^{n} \frac{z^k}{\sigma^{2k}k!} x^k \). We have
\[
|\exp\left(\frac{zx}{\sigma^2}\right) - g_n(x)| = \left| \sum_{k=n+1}^{+\infty} \frac{z^k}{\sigma^{2k}k!} x^k \right| \leq \sum_{k=n+1}^{+\infty} \frac{|z|^k}{\sigma^{2k}k!} |x|^k
\]
\[
\leq \sum_{k=1}^{+\infty} \frac{|z|^k}{\sigma^{2k}k!} |x|^k = \exp\left(\frac{|xz|}{\sigma^2}\right).
\]
Since the integral \( \int xp(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\frac{xz}{\sigma^2}\right)dx \) exists and is finite, by dominated convergence theorem, we can obtain
\[
\int xp(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\frac{xz}{\sigma^2}\right)dx = \int xp(x)\exp\left(-\frac{x^2}{2\sigma^2}\right) \lim_{n->+\infty} g_n(x)dx
\]
\[
= \lim_{n->+\infty} \int xp(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)g_n(x)dx
\]
\[
= \sum_{k=1}^{+\infty} \frac{1}{\sigma^{2k}k!} \int x^{k+1}p(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)dx z^k.
\]
for all \( z \in \mathbb{R} \). By the definition of analytic function, we have \( h(z) \) is analytic. Therefore, \( f(z) \) is also analytic.

Next we prove that \( S_{\sigma}^{MMSE}(z) \) is monotonous. Roughly speaking, the posterior distribution density of \( x \) is likely to concentrate around \( z \), and a larger \( z \) results in a larger posterior mean, i.e., the singleton MMSE estimate. The analytic and monotonous property imply that \( S_{\sigma}^{MMSE}(z) \) does not oscillate heavily in its domain. This will be used in the proof of the convergence of the iteration (1.7).

**Proposition 2.2.** Assume \( p(x) \) satisfies Assumption 1. The singleton MMSE estimator \( S_{\sigma}^{MMSE}(z) \) (2.3) is strictly increasing with respect to \( z \).
derivative and the integral to obtain the MMSE variational model.

2.2. MMSE variational model. Next, we establish a connection between the singleton MMSE estimator (2.3) and the variational model (1.8). For most applications, e.g., the MAP approach, the variational model is used, as its minimizer is the desired estimator. In contrast, we have already obtained the singleton MMSE estimator, and its corresponding variational model is built up for a different purpose: it is lifted to the general variational model (1.10) for the purpose of the convergence analysis of the iteration (1.7).

As the noise follows Gaussian distribution, the variational model should have an $\ell_2$ fidelity term. Hence, we choose the variational model in the following form:

$$\min_{x \in \mathbb{R}} \frac{1}{2} (x - z)^2 + \psi^\text{MMSE}_\sigma(x),$$

and the singleton MMSE estimator $S^\text{MMSE}_\sigma$ is the minimizer of this variational model, that is, $S^\text{MMSE}_\sigma$ is a solution of

$$x - z + \partial \psi^\text{MMSE}_\sigma(x) = 0.$$

Since $S^\text{MMSE}_\sigma$ is strictly monotonous by Proposition 2.2, we define $r^\text{MMSE}_\sigma$ as the inverse of $S^\text{MMSE}_\sigma$ on the range of $S^\text{MMSE}_\sigma$. With the definition of $r^\text{MMSE}_\sigma$, it is clear $\hat{x} = S^\text{MMSE}_\sigma(z)$ satisfies

$$\hat{x} - z + (r^\text{MMSE}_\sigma(\hat{x}) - \hat{x}) = 0.$$  

Then, $\psi^\text{MMSE}_\sigma(x) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ should be

$$\partial \psi^\text{MMSE}_\sigma(x) = r^\text{MMSE}_\sigma(x) - x.$$

That is,

$$\psi^\text{MMSE}_\sigma(x) = \int_0^x (r^\text{MMSE}_\sigma(u) - u) du.$$  

While $r^\text{MMSE}_\sigma$ is only defined on the range of $S^\text{MMSE}_\sigma$, we extend the domain of $\psi^\text{MMSE}_\sigma(x)$ to $\mathbb{R}$. Since $S^\text{MMSE}_\sigma$ is analytic by Lemma 2.1, $r^\text{MMSE}_\sigma$ is also analytic by the Lagrangian inversion theorem. The range of $r^\text{MMSE}_\sigma$ is $\mathbb{R}$, so its domain must be an open interval. We denote the domain of $r^\text{MMSE}_\sigma$ (i.e. the range of $S^\text{MMSE}_\sigma$) by $(l_\sigma, u_\sigma)$, where $-\infty \leq l_\sigma < u_\sigma \leq +\infty$, and also call it...
the effective domain of $\varphi^{\text{MMSE}}_\sigma(x)$. Outside the effective domain $(l_\sigma, u_\sigma)$, we define the value of $\varphi^{\text{MMSE}}_\sigma(x)$ as

$$
\varphi^{\text{MMSE}}_\sigma(x) = \begin{cases} 
\lim_{x \to u_\sigma^-} \varphi^{\text{MMSE}}_\sigma(x), & \text{if } x = u_\sigma < +\infty, \\
\lim_{x \to l_\sigma^+} \varphi^{\text{MMSE}}_\sigma(x), & \text{if } x = l_\sigma > -\infty, \\
+\infty, & \text{if } x \in \mathbb{R} \setminus (l_\sigma, u_\sigma).
\end{cases}
$$

Note that $\varphi^{\text{MMSE}}_\sigma$ is continuous on the closure of its effective domain if it is finite on that. In summary, we have the following lemma.

**Lemma 2.3.** Assume $p(x)$ satisfies Assumption 1. The singleton MMSE estimator $S^{\text{MMSE}}_\sigma(x) \in (l_\sigma, u_\sigma)$ is the unique solution of the singleton variational model

$$
\min_{x \in \mathbb{R}} h(x) = \frac{1}{2} (x - z)^2 + \varphi^{\text{MMSE}}_\sigma(x).
$$

**Proof.** Firstly, we consider $x \in (l_\sigma, u_\sigma)$ and show that $h(x) > h(S^{\text{MMSE}}_\sigma(z))$ if $x \neq S^{\text{MMSE}}_\sigma(z)$. We can rewrite $h(x)$ as

$$
h(x) = \int_0^x r^{\text{MMSE}}_\sigma(u) du - xz + \frac{1}{2} z^2 = \int_0^x (r^{\text{MMSE}}_\sigma(u) - z) du + \frac{1}{2} z^2.
$$

Thus, we have

$$
h(x) - h(S^{\text{MMSE}}_\sigma(z)) = \int_{S^{\text{MMSE}}_\sigma(z)}^x (r^{\text{MMSE}}_\sigma(u) - z) du.
$$

Since $S^{\text{MMSE}}_\sigma(x)$ and $r^{\text{MMSE}}_\sigma(x)$ are strictly increasing, it is easy to get that $h(x) > h(S^{\text{MMSE}}_\sigma(z))$ if $x \neq S^{\text{MMSE}}_\sigma(z)$.

Next we show it always holds that

$$
h(u_\sigma) > h(S^{\text{MMSE}}_\sigma(z)), \forall z \in \mathbb{R},
$$

when $u_\sigma$ is finite. Since $r^{\text{MMSE}}_\sigma(x)$ is strictly increasing and its domain is $\mathbb{R}$, we have

$$
\lim_{x \to u_\sigma^-} r^{\text{MMSE}}_\sigma(x) \to +\infty, \text{ as } x \to u_\sigma.
$$

Thus, $h(x)$ is monotonously increasing as $x$ is sufficiently close to $u_\sigma$. Then there exists $\tilde{x} \in (l_\sigma, u_\sigma)$ which is very close to $u_\sigma$ and not equal to $S^{\text{MMSE}}_\sigma(z)$ such that

$$
h(u_\sigma) > h(\tilde{x}) > h(S^{\text{MMSE}}_\sigma(z)).
$$

Similarly we can prove

$$
h(l_\sigma) > h(S^{\text{MMSE}}_\sigma(z)), \forall z \in \mathbb{R},
$$

when $l_\sigma$ is finite. Moreover, when $x \in \mathbb{R} \setminus (l_\sigma, u_\sigma)$, $h(x) = +\infty > h(S^{\text{MMSE}}_\sigma(z))$. So $x = S^{\text{MMSE}}_\sigma(z)$ is the unique solution to the minimization problem (2.7).

Furthermore, the objective function of the variational model (2.7), i.e. $h(x)$, is strongly convex.

**Proposition 2.4.** Assume $p(x)$ satisfies Assumption 1. The objective function $h(x)$ of the singleton variational model (2.7) is $\Delta$-strongly convex on its effective domain $(l_\sigma, u_\sigma)$, where

$$
\Delta = 1 / \sup_{x \in \mathbb{R}} \left\{ \frac{dS^{\text{MMSE}}_\sigma(x)}{dx} \right\}.
$$

**Proof.** Denote $h_\Delta(x) = h(x) - \frac{\Delta}{2} x^2$. We can get that

$$
h''_\Delta(x) = \frac{dr^{\text{MMSE}}_\sigma(x)}{dx} - \Delta > 0.
$$

If $\Delta$ is positive, it directly gives that $h(x)$ is $\Delta$-strongly convex. To prove $\Delta$ is positive, we only need to show the differential $\frac{dS^{\text{MMSE}}_\sigma(x)}{dx} > 0$ is bounded. Since we have shown that $\frac{dS^{\text{MMSE}}_\sigma(x)}{dz}$ is smooth in Proposition (2.1), it is sufficient to prove that

$$
-3z - \eta \leq S^{\text{MMSE}}_\sigma(z) \leq 3z + \eta, \ z \geq 0,
$$

(2.8)
and

\( 3z - \eta \leq S_\sigma^{\text{MMSE}}(z) \leq -3z + \eta \), \( z < 0 \),

for some \( \eta > 0 \) when \(|z|\) is large enough. The proof of (2.9) is similar to that of (2.8), and we only prove (2.8) here. As \( S_\sigma^{\text{MMSE}}(z) \) is monotonously increasing, the left inequality of (2.8) always holds when \( z \) is sufficiently large, and next we prove the right inequality of (2.8).

Denote

\[ \phi_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \]

When \( z \) is positive, we have

\[
S_\sigma^{\text{MMSE}}(z) = \frac{\int x p(x) \phi_\sigma(x-z) \, dx}{\int p(x) \phi_\sigma(x-z) \, dx} \leq \frac{\int_0^{\infty} x p(x) \phi_\sigma(x-z) \, dx}{\int_{-z}^{\infty} p(x) \phi_\sigma(x-z) \, dx}
\]

\[
\leq \frac{\int_0^{3z} x p(x) \phi_\sigma(x-z) \, dx}{\int_{-z}^{3z} p(x) \phi_\sigma(x-z) \, dx} + \frac{\int_{3z}^{\infty} x p(x) \phi_\sigma(x-z) \, dx}{\int_{-z}^{3z} p(x) \phi_\sigma(x-z) \, dx}
\]

\[
\leq 3z + \frac{\int_{3z}^{\infty} x p(x) \phi_\sigma(x-z) \, dx}{\int_{-z}^{3z} p(x) \phi_\sigma(x-z) \, dx}.
\]

Since \( p(x) \) satisfies Assumption 1, we can get that

\[ \int_{3z}^{\infty} x p(x) \phi_\sigma(x-z) \, dx \leq \phi_\sigma(2z) \int_{3z}^{\infty} x p(x) \, dx, \]

and

\[ \int_{-z}^{3z} p(x) \phi_\sigma(x-z) \, dx \geq \phi_\sigma(2z) \int_{-z}^{3z} p(x) \, dx. \]

It yields that

\[ S_\sigma^{\text{MMSE}}(z) \leq 3z + \frac{\int_{3z}^{\infty} x p(x) \, dx}{\int_{-z}^{3z} p(x) \, dx}. \]

It is obvious that \( \int_{-z}^{3z} p(x) \to 1 \) and \( \int_{3z}^{\infty} x p(x) \, dx \to 0 \) as \( z \to +\infty \). Thus, for any \( \eta > 0 \), there exits \( K > 0 \) such that

\[ \frac{\int_{3z}^{\infty} x p(x) \, dx}{\int_{-z}^{3z} p(x) \, dx} < \eta, \text{ when } z > K. \]

Directly, we obtain that \( S_\sigma^{\text{MMSE}}(z) \leq 3z + \eta \) when \( z > K \).

\[ \square \]

2.3. Examples. We focus on three examples with special prior, namely, Gaussian, Gaussian mixture and sparse prior (1.11), which all satisfy Assumption 1. They are widely used in many applications. We also compare the MMSE estimator with the LMMSE and MAP estimator.

2.3.1. Gaussian prior. The first example is one of the most commonly used prior, the Gaussian prior distribution

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_s)^2}{2\sigma_s^2}\right). \]

The posterior distribution \( p(x|z) \) is also Gaussian, as claimed in [16]. In this form the singleton MMSE estimator is found as

\[ \hat{x}^{\text{MMSE}} = \alpha z + (1 - \alpha)\mu_s, \]

where

\[ \alpha = \frac{\sigma_s^2}{\sigma_s^2 + \sigma^2}. \]
As (2.13) is linear, the linear MMSE estimator coincides with the MMSE estimator. Moreover, since the mode of Gaussian distribution is the same as its mean, the MAP estimator is also identical to the MMSE estimator.

2.3.2. Gaussian mixture prior. The second example goes beyond the Gaussian setting and extends to the Gaussian mixture prior distribution

\[
p(x) = \sum_{i=1}^{k} p_i \frac{1}{\sqrt{2\pi} \sigma_i^2} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right), \quad \sum_{i=1}^{k} p_i = 1.
\]

The Gaussian mixture prior is widely applicable due to several properties. Firstly, Gaussian mixture distributions can approximate any continuous density with arbitrary accuracy [see 24]. Secondly, Gaussian mixture prior together with Gaussian noise setting gives a Gaussian mixture posterior distribution [see 16]. Therefore, the MMSE estimator, i.e., the posterior mean, is easy to obtain and takes the form of [see 14]

\[
x_{\text{MMSE}} = \sum_{i=1}^{k} \beta_i(z)(\alpha_i z + (1 - \alpha_i)\mu_i),
\]

where

\[
\alpha_i = \frac{\sigma_i^2}{\sigma_i^2 + \sigma^2}, \quad \beta_i(z) = \frac{p_i \exp\left(-\frac{(z - \mu_i)^2}{2(\sigma_i^2 + \sigma^2)}\right)}{\sum_{i=1}^{k} p_i \exp\left(-\frac{(z - \mu_i)^2}{2(\sigma_i^2 + \sigma^2)}\right)}.
\]

By Bayesian Gauss-Markov theorem (Theorem 12.1 of [16]), the LMMSE estimator is

\[
x_{\text{LMMSE}} = \hat{\alpha} z + (1 - \hat{\alpha})\hat{\mu},
\]

where

\[
\hat{\mu} = \sum_{i=1}^{k} p_i \mu_i, \quad \hat{\sigma}^2 = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i \mu_i^2 - \hat{\mu}^2, \quad \hat{\alpha} = \frac{\sigma^2}{\hat{\sigma}^2 + \sigma^2}.
\]

Actually, \(\hat{\mu}\) and \(\hat{\sigma}^2\) are the mean and variance of (2.14). Comparing with (2.13), we find that the LMMSE estimator (2.16) behaves as if to minimize the mean square error without the linear constraint and under the assumption of Gaussian prior with mean \(\hat{\mu}\) and variance \(\hat{\sigma}^2\).

Compared to the MMSE and LMMSE estimator, the MAP estimator has no explicit form and is not easy to compute. It seeks the maximum of the Gaussian mixture posterior probability density function while the Gaussian mixture distribution has multiple modes.

2.3.3. Bernoulli-uniform sparse prior. The third example is the Bernoulli-uniform sparse prior given in (1.11). It assumes that most of the parameters can be set to zero without substantially affecting the fitting of model. Substituting the Bernoulli-uniform sparse prior and the Gaussian noise distribution into (2.2), we get an explicit form of the singleton MMSE estimator as follows

\[
S_{\sigma}^{\text{MMSE}}(z) = \frac{1 - p_0}{2(U - L)} \cdot \int_{[-U,-L]\cup[L,U]} \left\{ \frac{p_0 \phi_\sigma(x - z)}{\phi_\sigma(z)} + \frac{1 - p_0}{2(U - L)} \cdot \int_{[-U,-L]\cup[L,U]} \phi_\sigma(x - z) dx \right\} dx,
\]

where \(\phi_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\). Let \(\Psi_\sigma(t)\) denote the cumulative distribution function of the Gaussian distribution \(\mathcal{N}(0, \sigma^2)\) and

\[
C = \frac{2p_0(U - L)}{1 - p_0},
\]

\[
G_1 = \Psi_\sigma(U - z) + \Psi_\sigma(-L - z) - \Psi_\sigma(L - z) - \Psi_\sigma(-U - z),
\]

\[
G_2 = \sigma^2(\phi_\sigma(L - z) + \phi_\sigma(-U - z) - \phi_\sigma(U - z) - \phi_\sigma(-L - z)).
\]

Then the singleton MMSE estimator can be rewritten as

\[
S_{\sigma}^{\text{MMSE}}(z) = \frac{zG_1 + G_2}{C\phi_\sigma(z) + G_1}.
\]

Figure 1 displays the singleton MMSE estimator obtained by (2.21).
As suggested by [12], a good sparse estimator should satisfy the following three conditions:

(i) **continuity**: the estimator is continuous in data $z$ to avoid instability in model prediction;
(ii) **unbiasedness**: the estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias;
(iii) **sparsity**: the estimator sets small estimated coefficients to zero to reduce model complexity.

The singleton MMSE estimator is a good sparse estimator. Firstly, the singleton MMSE estimator is analytic by Proposition 2.1. Secondly, it is unbiased since

$$E_z S_{\sigma}^{\text{MMSE}}(z) = E_z E_x (x|z) = x.$$ 

Thirdly, it is also sparse. Specifically, when $z$ is very small, $G_1$ in (2.19) and $G_2$ in (2.20) is much smaller than the constant value $C$ in(2.18); hence, the estimate is very close to zero.

Next, we compare the MMSE estimator with the LMMSE estimator. The LMMSE estimator under the Bernoulli-uniform sparse prior is a shrinkage operator

$$x^{\text{LMMSE}} = \frac{\sigma^2}{\sigma^2 + \sigma_z^2} z,$$

where $\sigma_z = \frac{1}{3}(1 - p_0)(L^2 + U^2 + LU)$ is the variance of the Bernoulli-uniform sparse prior. The LMMSE estimator (2.22) is continuous, but it is biased and loses the sparsity.

Finally, we compare the MMSE estimator with the MAP estimator. From (1.14), the MAP model for the singleton problem (1.4) with Gaussian noise takes the form of

$$\min_{x \in \mathbb{R}} \frac{1}{2}(x - z)^2 + \varphi(x),$$

where $\varphi(x) = -\sigma^2 \log p(x)$. Note that the above MAP model is developed for continuous random variables and seeks the maximum of the probability density function. However, the Bernoulli-uniform distribution density function is not continuous at origin and has infinite value at zero. Nevertheless, we relax the rigorousness in analysis and write $\varphi(x)$ formally by assuming $p(0)$ is arbitrarily large:

$$\varphi^a(x) = \begin{cases} a \|x\|_0 + b, & x \in \{0\} \cup [-U, -L] \cup [L, U], \\ +\infty, & \text{otherwise,} \end{cases}$$

where

$$a = \sigma^2 \left( \log p(0) - \log \frac{1 - p_0}{2(U - L)} \right), \quad b = -\sigma^2 \log p(0).$$

The solution of (2.23) with $\varphi(x) = \varphi^a(x)$ can be solved formally. Since the Bernoulli-uniform sparse prior assumes the underlying truth is bounded, the MAP estimate for $z > U$ (resp. $z < -U$) is either 0 or $U$ (resp. $-U$). In the following, we only consider the case $|z| \leq U$. Let
\[ \lambda = \sigma \sqrt{2 \log \frac{2p(0)(U-L)}{1-p_0}} \]  
When \( \lambda > U \), the resulting MAP estimator for the Bernoulli-uniform sparse prior is always zero. When \( L \leq \lambda \leq U \), the MAP estimator is the hard-thresholding 
\[ x_{\text{hard}} = S^\lambda_\gamma(z) := zI(|z| > \lambda). \]

Since the hard-thresholding is a good sparse estimator, it justifies the validity of the Bernoulli-uniform sparse prior. In contrast, the Bernoulli-Gaussian prior \([17, 19]\), which is the combination of a Dirac-delta distribution and a Gaussian distribution, results in an additional \( \ell_2 \)-norm penalization on non-zero values and only gives the hard-thresholding estimator in the limiting case where the variance of the Gaussian distribution goes to infinity [see 25]. We note that one also has to relax the rigorousness when using the Bernoulli-Gaussian prior as a sparse prior for the MAP model. Strictly speaking, there is no prior distribution in the MAP setting corresponding to the hard thresholding.

The hard-thresholding is sparse and nearly unbiased for large coefficients, but it is discontinuous. As shown in (1.15), its discontinuity results from the instability of the minus delta loss function of the MAP model. To remedy the drawback of the hard-thresholding, the continuous soft-thresholding \([10]\) is proposed:
\[ x_{\text{soft}} = S^\gamma_\lambda(z) := \text{sign}(z)(|z| - \lambda)_+, \]
which corresponds to the \( \ell_1 \)-norm penalty (the \( \ell_1 \)-norm penalty is the convex relaxation of the \( \ell_0 \)-norm penalty and assumes a Laplace prior distribution). However, the continuity of the soft-thresholding comes at the price of shifting the large coefficients by \( \lambda \) and introducing larger bias. Under the spirit of “continuity”, “unbiasedness” and “sparsity”, \([12]\) proposed a Smoothly Clipped Absolute Deviation Penalty (SCAD) thresholding:
\[ x_{\text{SCAD}} = S^{\gamma,\lambda}_\gamma(z) := \begin{cases} 
\text{sign}(z)(|z| - \lambda)_+, & |z| \leq 2\lambda, \\
\frac{(\gamma-1)}{\gamma}z - \text{sign}(z)\gamma\lambda, & 2\lambda < |z| < \gamma\lambda, \\
z, & |z| \geq \gamma\lambda.
\end{cases} \]

with \( \gamma \geq 1 \). The corresponding SCAD penalty of the MAP model (2.23) is given by
\[ \varphi^{\gamma,\lambda}_{\text{SCAD}}(x) = \begin{cases} 
\lambda|x|, & |x| \leq \lambda, \\
\frac{\gamma}{2}\lambda^2 - \frac{(\gamma\lambda - |x|)^2}{2(\gamma - 1)}, & |x| > \lambda.
\end{cases} \]

The SCAD-thresholding is continuous compared to the hard-thresholding and avoids the large bias introduced by the soft-thresholding. Actually, it can be interpreted as a linear interpolation between the soft-thresholding and the hard-thresholding in the interval \((2\lambda, 2\gamma\lambda)\). The SCAD-thresholding also assumes a different prior from the Bernoulli-uniform sparse prior. The relating prior of SCAD smooths the discontinuity of the Bernoulli-uniform sparse prior around the origin, but still keeps sparsity and is uniform for large parameters. The MCP-thresholding \([27]\) is proposed under a similar spirit as SCAD while the corresponding penalty of MCP is less concave than that of SCAD. The MCP penalty term takes the form of
\[ \varphi^{\text{MCP}}_{\gamma,\lambda}(x) = \begin{cases} 
\lambda|x| - \frac{|x|^2}{2\gamma}, & |x| \leq \gamma\lambda, \\
\frac{1}{2}\gamma^2, & |x| > \gamma\lambda,
\end{cases} \]

and the MCP thresholding is
\[ x_{\text{MCP}} = S^{\text{MCP}}_{\gamma,\lambda}(z) := \begin{cases} 
0, & |z| \leq \lambda, \\
\text{sign}(z)\frac{\gamma(|z| - \lambda)}{\gamma - 1}, & \lambda < |z| < \gamma\lambda, \\
z, & |z| \geq \gamma\lambda,
\end{cases} \]

with \( \gamma \geq 1 \).

By proposing the three conditions of “sparsity”, “unbiasedness” and “continuity”, Fan and Li \([12]\) actually means to find a stable way to balance the variance and bias. In Bayesian estimation, the prior knowledge influences the parameter estimation by shifting its posterior probability mass density towards the region preferred by the prior, e.g., the sparse prior shifts the density towards zero. As seen, this influence has two effects: reducing the variance of the estimate (or the noise effect), but gaining bias simultaneously. The “sparsity” and “unbiasedness” is the result
of negotiation between the variance and bias in different ranges of observations. The “sparsity” means omitting the small observations while the “unbiasedness” means maintaining the large ones. Since the small observations contain more noise than signal information, omitting them reduces much variance and gains little bias. On the contrary, the large observations contain strong signal information such that the noise effect can be neglected; and thus maintaining them is wiser.

The mean square error can be decomposed as the sum of the variance and square bias of the estimator

\[
\text{MSE} = E_x(x - g(z))^2 = E_x(x - EX g(z))^2 + E_z(g(z) - EX g(z))^2 = \text{Bias}(g(z))^2 + \text{Var}(g(z)).
\]

So the singleton MMSE estimator under the Bernoulli-uniform sparse prior naturally provides a balance between the variance and bias, i.e., a balance between “sparsity” and “unbiasedness”. Moreover, it is smooth by Proposition 2.1. In short, the singleton MMSE estimator balances the variance and bias in a stable way and satisfies the three conditions of “continuity”, “unbiasedness” and “sparsity”.

Finally, to demonstrate the efficiency of the singleton MMSE estimator in the sparse case, we implement a parameter estimation test for the singleton model (1.4) and evaluate the performance by \( \ell_2 \)-norm loss \( \|\hat{x} - x\|^2 \). The results are reported in Table 1. The parameter \( x \) is constructed to follow the Bernoulli-uniform sparse prior with \( p_0 = 0.8, L = 3, U = 6 \) in (1.11) and the noise variance \( \sigma^2 \) is 1. The experiments are repeated 1000 times. The LMMSE estimator fails to enhance sparsity and we only test the sparse estimators displayed in Figure 2. The parameters \( \lambda \) and \( \gamma \) in the related thresholding vary from \([0.5, 3]\) and \([1, 3]\) respectively. Only the least average \( \ell_2 \)-norm loss and the corresponding parameters are reported. It comes at no surprise that the singleton MMSE estimator outperforms the others.

<table>
<thead>
<tr>
<th>estimator</th>
<th>hard</th>
<th>soft</th>
<th>SCAD</th>
<th>MCP</th>
<th>MMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )</td>
<td>2.5</td>
<td>0.9</td>
<td>1.4</td>
<td>1.6</td>
<td>–</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>( |\hat{x} - x|^2 )</td>
<td>20.20</td>
<td>22.28</td>
<td>20.05</td>
<td>20.13</td>
<td><strong>19.90</strong></td>
</tr>
</tbody>
</table>

### 3. Approximation of the MMSE estimator

In this section, we will propose the cyclic coordinate minimization algorithm based on the iteration (1.7) and establish a theorem on its convergence. The numerical experiments will follow to show the efficiency of our algorithm.
3.1. Algorithm and main theorem. We consider the regression model
\[ z = Ax + \varepsilon, \]
where \( z \in \mathbb{R}^n \) is the observation, \( A \in \mathbb{R}^{n \times p} \) is the design matrix, \( x \in \mathbb{R}^p \) is the parameter of interest, and \( \varepsilon \in \mathbb{R}^n \) is the Gaussian noise with mean zero and variance \( \sigma^2 I \). We assume that \( x_i (i = 1, 2, \ldots, p) \) are independent and identically distributed, and follow the distribution \( p(x) : \mathbb{R} \to \mathbb{R} \cup \{+\infty\} \). More generally, we can consider the coefficients of \( x \) under some orthonormal basis \( D: u_i = (D x)_i, (i = 1, 2, \ldots, p) \), to be independent and identically distributed, and follow some prior distribution. Once an estimate of \( u \) is obtained, \( x \) can be recovered by \( x = D^T u \). So we focus on the estimation of \( u \)
from
\[ z = AD^T u + \varepsilon. \]
Nevertheless, we can build the orthonormal matrix \( D \) into the design matrix \( A \) such that the form of (3.2) reduces to that of (3.1). So we only consider the problem (3.1) in the following.

The MMSE estimator for (3.1) is
\[ T_{\text{MMSE}}(z) = E_x(z|x) = \frac{\int x \prod_{j=1}^p p(x_j) \exp\left(-\frac{\|Ax-z\|^2}{2\sigma^2}\right) dx}{\int \prod_{j=1}^p p(x_j) \exp\left(-\frac{\|Ax-z\|^2}{2\sigma^2}\right) dx}. \]
It has an explicit form but is hard to compute because of the involvement of multiple integrals. To overcome this difficulty, we propose a cyclic coordinate minimization (CCM) algorithm to approximate (3.3) iteratively. The key to the success of our algorithm lies in reducing to the singleton case at each iteration and involving only single integrals, which is easy to compute.

The coordinate minimization algorithm approximates (3.3) by minimizing the mean square error with respect to a single coordinate in a sequential manner. More specifically, for the current estimate \( x^k \), an index \( i_k \in \{1, 2, \ldots, p\} \) is selected and then the \( i_k \)-th component of \( x^k \) is adjusted to minimize the mean square error:
\[ a_{i_k}^T x^k = \arg \min_{x \in \mathbb{R}} E_{(x_{i_k}, x|\varepsilon_{\neq i_k})} ((x_{i_k} - x)^2). \]
It is equivalent to finding the MMSE estimator for the following linear regression model
\[ z - \sum_{j \neq i_k} a_j x_j = a_{i_k} x_{i_k} + \varepsilon, \]
where \( a_i \) is the \( i \)-th column of \( A \). \( x_{i_k} \in \mathbb{R} \) is the parameter to be estimated and \( z - \sum_{j \neq i_k} a_j x_j \in \mathbb{R}^n \) is the vector of observations. The MMSE estimator for (3.5) is the posterior mean:
\[ E_{x_{i_k}} (x_{i_k}|z - \sum_{j \neq i_k} a_j x_j) = \frac{\int x_{i_k} p(x_{i_k}) \exp\left(-\frac{\|a_{i_k} x_{i_k} - (z - \sum_{j \neq i_k} a_j x_j)^2\|}{2\sigma^2}\right) dx_{i_k}}{\int p(x_{i_k}) \exp\left(-\frac{\|a_{i_k} x_{i_k} - (z - \sum_{j \neq i_k} a_j x_j)^2\|}{2\sigma^2}\right) dx_{i_k}}. \]
That is, the solution of (3.4) is obtained as
\[ x_{i_k}^{k+1} = S_{\text{MMSE}}^{a_{i_k}} (z - \sum_{j \neq i_k} a_j x_j). \]
At each iteration, the computation of (3.6) involves only single integration and is easy to perform. The coordinate minimization method can be deterministic or randomized depending on the choice of the update coordinates. Here, we select the coordinate index \( i_k \) in a cyclic fashion from the set \( \{1, 2, \ldots, p\} \) so that every coordinate is modified once in every cycle of \( p \) iterations. The resulting
The computation of the algorithm may be faster. Given an 
permutation \( \{i_0, i_1, \ldots, i_p-1\} \) of sequence 
\( \{1, 2, \ldots, p\} \), we choose the index \( i_k \) as
\[
(3.7) \quad i_k = \hat{i}_k, \text{ where } \hat{k} = [k \mod p].
\]

Finally, the CCM algorithm is summarized as follows.

**Algorithm 1** The Cyclic Coordinate Minimization (CCM) algorithm

Set \( k \leftarrow 0 \) and choose \( \mathbf{x}_0 \in \mathbb{R}^p \) such that \( H(\mathbf{x}_0) < +\infty \); while termination test is not satisfied do

- choose index \( i_k \) as \( (3.7) \)
- \[
x_{i_k}^{k+1} = S^\text{MMSE}_k(\mathbf{a}_{i_k}) \mathbf{x}_k^{k} - \sum_{j \neq i_k} \mathbf{a}_j \mathbf{x}_j^{k} / \|\mathbf{a}_{i_k}\|_2^2
\]
- \( k \leftarrow k + 1; \)
end while

We note that when Algorithm 1 converges, it converges to a fixed point of the iteration \( (3.4) \). Let \( \hat{x} \) be the fixed point of \( (3.4) \). It satisfies
\[
(3.8) \quad \hat{x}_{i_k} = \arg \min_{\mathbf{x} \in \mathbb{R}} \mathbb{E}_{(\mathbf{x}_{i_k}, \mathbf{x}_{i \neq i_k})} (\mathbf{x}_{i_k} - \mathbf{x})^2.
\]
This means that \( \hat{x} \) attains the minimum of the mean square error with respect to arbitrary single coordinate by fixing the remaining ones. We call \( \hat{x} \) the coordinatewise minimum mean square error (CMMSE) estimate.

Before giving the main theorem on the convergence of Algorithm 1, we need the following assumption.

**Assumption 2.** Assume the prior distribution \( p(x) \) satisfies Assumption 1. Furthermore, there exist \( K > 0 \) and \( b > 0 \) such that
\[
(3.9) \quad p(x) = a(x) \exp(-b|x|), \quad |x| > K,
\]
where \( a(x) \geq 0 \) monotonously decreases to zero when \( |x| \) goes to infinity.

Note that Assumption 2 is not strong. Gaussian, Gaussian mixture and the Bernoulli-uniform sparse prior all satisfy Assumption 2. For example, since the Bernoulli-uniform sparse prior is compactly supported, it satisfies \( (3.9) \) as we can choose \( a(x) = 0 \) for a sufficiently large \( K \). Next we state the main theorem of this paper and its proof will be given in section 4.

**Theorem 3.1.** Assume the prior distribution \( p(x) \) satisfies Assumption 2. Let \( \{\mathbf{x}^k\} \) be the sequence generated by Algorithm 1. Then \( \{\mathbf{x}^k\} \) converges to the coordinatewise minimum mean square error (CMMSE) estimate as defined in \( (3.8) \).

**3.2. Numerical results.** Finally, we show the efficiency of our algorithm by the three examples discussed in section 2.3, i.e., Gaussian, Gaussian mixture and the Bernoulli-uniform sparse prior. The case where \( n = 100, p = 200 \) is tested. The columns of matrix \( \mathbf{A} \) in \( (3.1) \) are highly correlated, which are generated from Gaussian distribution with zero mean and variance \( \Sigma_{i,j} = 0.3^{|i-j|} \) (1 \( \leq i \leq n, \ 1 \leq j \leq p \)). We stop the iteration for Algorithm 1 if a maximum number of 300 iterations is reached, or the difference of the estimates between two successive cycles is smaller than \( 10^{-8} \), i.e., \( \|\mathbf{x}^{(l+1)p} - \mathbf{x}^{lp}\|_2 \leq 10^{-8} \). For comparison, the MMSE, LMMSE and MAP estimator are also computed when they are available and easy to compute.

The built-in singleton MMSE estimator in Algorithm 1 for each prior distribution has been given in section 2.3 accordingly. In the implementation, we can calculate the values of the singleton MMSE estimator, i.e. \( \mathbb{E}(x|z) \), at a given set of sample points of \( z \) in advance. Then the singleton MMSE estimator can be approximated by piecewise linear functions with arbitrary accuracy as long as the set of sample points is large enough. When iterating, we only need to call that piecewise linear function to update the sequence and avoid calculating the single integrals every time such that the computation of the algorithm may be faster.
Example 3.2. Gaussian prior. Under the Gaussian prior, all of the MMSE, LMMSE and MAP estimator for the problem (1.1) take the same form of

\[
\hat{x} = \alpha z + (1 - \alpha)\mu_s, \quad \alpha = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2}.
\]

Since (3.10) is separable, Algorithm 1 also converges to (3.10) after one cycle. That is, all these estimators have the same performance under the Gaussian prior.

Example 3.3. Gaussian Mixture prior. Similar to the singleton case, the posterior distribution with the Gaussian mixture prior and Gaussian noise is also the Gaussian mixture distribution for the general problem (3.1). In [14], the MMSE estimator for the general problem (3.1) under the Gaussian mixture prior is given by

\[
x_{\text{MMSE}} = \sum_{i=1}^{k} \beta_i(z)\left(\mu_i e + A^T(AA^T + \frac{\sigma^2_i}{\sigma^2}I)^{-1}(z - \mu_i A e)\right), \quad \beta_i(z) = \frac{\phi_i(z)}{\sum_{i=1}^{k} \phi_i(z)}
\]

where \(I\) is the identity matrix, \(e\) is the vector of all ones, and \(\phi_i(z)\) is the Gaussian probability density function (PDF) in \(z\) with mean

\[
\hat{\mu}_i = \mu_i A e,
\]

and variance

\[
C_i = \sigma^2_i A A^T + \sigma^2 I.
\]

From [16], the LMMSE estimator is

\[
x_{\text{LMMSE}} = \hat{\mu} e + A^T(AA^T + \frac{\sigma^2}{\sigma^2}I)^{-1}(z - \hat{\mu} A e).
\]

where

\[
\hat{\mu} = \sum_{i=1}^{k} p_i \mu_i, \quad \hat{\sigma}^2 = \sum_{i=1}^{k} p_i \sigma^2_i + \sum_{i=1}^{k} p_i \mu^2_i - \hat{\mu}^2.
\]

The MAP estimator is hard to compute due to the multi-mode issue of the Gaussian mixture posterior distribution, so we do not compare it here.

![Figure 3](image_url). The average \(\ell_2\) loss of the estimates for the general problem (3.1) under Gaussian mixture prior.

We select the Gaussian mixture prior with five components. The component means are

\((-1, -0.5, 0, 1, 3),\)

and the variances are all 1. We vary the logarithm of the noise variance from -5 to 1 equally. Algorithm 1 is initialized with the LMMSE estimate. The average \(\ell_2\) loss of the obtained estimates over 50 independent trials are plotted in Figure 3. Clearly, Algorithm 1 provides a better approximation of the MMSE estimator than the LMMSE estimator.
Example 3.4. The Bernoulli-uniform sparse prior. Under the Bernoulli-uniform sparse prior, the MAP models with the $\ell_1$-norm, SCAD and MCP penalty for the general problem (3.1) are solved to compare with our algorithm. We choose $\gamma = 3.7$ in the SCAD penalty and $\gamma = 2$ in the MCP penalty. The $\ell_1$-norm penalized model is computed by the MATLAB package Glmnet\(^1\), while the SCAD or MCP penalized model is solved by the LLA algorithm proposed by [13]. The LLA algorithm computes a sequential weighted $\ell_1$-norm penalized problems. It is initialized with zero and terminated if a maximum number of 300 iterations is reached or the $\ell_2$ norm of the difference between two successive updates is smaller than $10^{-8}$. In addition, we compute the oracle solution for comparison. The oracle solution is the solution obtained by knowing the true support:

$$x_{\text{oracle}} = \arg \min_{x \in \mathbb{R}^p : x_{\text{oracle}} = 0} \frac{1}{2} \| Ax - z \|_2^2,$$

where $S$ is the support set of the truth $x$ and $S^c$ is the complementary set of $S$. The LMMSE estimator is not compared here because it fails to enhance sparsity. Our algorithm is initialized with the solution of the the $\ell_1$-norm penalized model.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\ell_1$ loss</th>
<th>$\ell_2$ loss</th>
<th>#FP</th>
<th>#FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>5.4407</td>
<td>1.4980</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0.9543)</td>
<td>(0.2411)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>$\ell_1$-norm penalty</td>
<td>25.4007</td>
<td>4.6593</td>
<td>35.88</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(7.3596)</td>
<td>(1.2588)</td>
<td>(4.58)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>SCAD</td>
<td>5.7361</td>
<td>1.6021</td>
<td>0.36</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(1.1306)</td>
<td>(0.3327)</td>
<td>(0.66)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>MCP</td>
<td>5.4622</td>
<td>1.5075</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.9834)</td>
<td>(0.2568)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Algorithm 1</td>
<td><strong>4.7092</strong></td>
<td><strong>1.3060</strong></td>
<td><strong>0.00</strong></td>
<td><strong>0.00</strong></td>
</tr>
<tr>
<td></td>
<td><strong>(0.7872)</strong></td>
<td><strong>(0.2076)</strong></td>
<td><strong>(0.00)</strong></td>
<td><strong>(0.00)</strong></td>
</tr>
</tbody>
</table>

The estimation accuracy is evaluated by the average $\ell_1$ loss and $\ell_2$ loss. We treat the estimated parameter as zero if its absolute value is less than $10^{-6}$. The selection accuracy of zero parameters is measured by the average counts of false positive (#FP) and false negative (#FN). The average performance over 50 independent trials is reported in Table 2 with the standard error shown in the parenthesis. The experiment results show that our method performs the best in any way of evaluation.

4. Proof of Theorem 3.1

It is hard to analyze the convergence of Algorithm 1 from the aspect of minimizing the mean square error sequentially. Inspired by the connection between the singleton MMSE estimator and the singleton variational model, we consider the following variational model for the general problem:

$$\min_{x \in \mathbb{R}^p} H(x) = \frac{1}{2} \| Ax - z \|_2^2 + \sum_{i=1}^p \| a_i \|_2^2 \phi_{\text{MMSE}}(x_i) \sigma / \| a_i \|_2,$$

which has an $\ell_2$ fidelity for Gaussian noise and a separable MMSE penalty. Applying the coordinate minimization algorithm to (4.1), we update a single coordinate and fix the remaining ones to

minimize $H(x)$ at each iteration. Specifically, at iteration $k$, we only update $x_{ik}$ as follows

$$x_{ik}^{k+1} = \arg \min_{x_{ik} \in \mathbb{R}} H(x_{ik})$$

(4.2)

$$= \arg \min_{x_{ik} \in \mathbb{R}} \frac{1}{2} \|a_{ik}\|^2 (x_{ik} - \left( a_{ik}^T (z - \sum_{j \neq ik} a_{jk} x_j^k) \right)^2 + \|a_{ik}\|^2 \varphi_{\sigma}^{\text{MMSE}}(x_{ik})).$$

Since the values of $x_{ik}^k$ where $j \neq ik$ are fixed, (4.2) reduces to the singleton variational model with respect to $x_{ik}$. By Lemma 2.3, the solution is

$$x_{ik}^{k+1} = S_{\sigma/\|a_{ik}\|}^{\text{MMSE}}(a_{ik}^T (z - \sum_{j \neq ik} a_{jk} x_j^k)/\|a_{ik}\|^2),$$

which is the same as the iteration (3.6). To conclude, we have the following lemma.

**Lemma 4.1.** Applying the coordinate minimization algorithm to the the variational model (4.1), we obtain the iteration (3.6).

Moreover, choosing the index $i_k$ as (3.7) in a cyclic manner, we obtain the cyclic coordinate minimization algorithm for solving the variational model (4.1) that is the same as Algorithm 1. Thus we can prove the convergence of Algorithm 1 by analysing the properties of the objective function $H(x)$ of the variational model (4.1) and employing the tools developed in optimization.

In order to prove the convergence of Algorithm 1, we require the objective function $H(x)$ to satisfy the Kurdyka-Lojasiewicz property and to be coercive. Once these two requirements are satisfied, Theorem 3.1 can be proved by applying the theorem developed by [1].

### 4.1. The Kurdyka-Lojasiewicz property and coerciveness of $H(x)$

We start by giving the definition of the limiting subdifferential (or simply subdifferential) [22] as it plays an important role in the following analysis.

**Definition 4.2** (Subdifferentials). Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous function.

(i) The *regular subdifferential* of $f$ at $\bar{z} \in \text{dom } f = \{z \in \mathbb{R}^p : f(z) < +\infty\}$ is defined as

$$\hat{\partial} f(\bar{z}) := \left\{ v \in \mathbb{R}^p : \liminf_{\bar{z} \to \bar{z}, \bar{z} \neq \bar{z}} \frac{f(z) - f(\bar{z}) - \langle v, z - \bar{z} \rangle}{\|z - \bar{z}\|} \geq 0 \right\};$$

(ii) The (limiting) *subdifferential* of $f$ at $\bar{z} \in \text{dom } h$ is defined as

$$\partial f(\bar{z}) := \left\{ v \in \mathbb{R}^p : \exists z^{(k)} \to \bar{z}, f(z^{(k)}) \to f(z), v^{(k)} \in \hat{\partial} f(z^{(k)}), v^{(k)} \to v \right\}.$$

From the above definition, we have the following remark.

**Remark 4.3.**

- If $z^k \to z$, $f(z^k) \to f(z)$, $v^k \in \partial f(z^k)$ and $v^k \to v$, then $v \in \partial f(z)$.
- If $x$ is on the effective domain of $\varphi_{\sigma}^{\text{MMSE}}$, i.e. the domain of $r_{\sigma}^{\text{MMSE}}$, the subdifferential of $\varphi_{\sigma}^{\text{MMSE}}$ is

$$\partial \varphi_{\sigma}^{\text{MMSE}}(x) = r_{\sigma}^{\text{MMSE}}(x) - x;$$

otherwise, it is empty.

Then we give the definition of the Kurdyka-Lojasiewicz (KL) property. We say a real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper if it is not positive infinity identically.

**Definition 4.4** (Kurdyka-Lojasiewicz Property). A proper real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is said to have the Kurdyka-Lojasiewicz property at $\bar{z} \in \text{dom } \partial f = \{z \in \mathbb{R}^p : \partial f(z) \neq \emptyset\}$ if there exist $\zeta \in (0, +\infty]$, a neighborhood $U$ of $\bar{z}$, and a continuous concave function $\psi : [0, \zeta) \to \mathbb{R}_+$ such that

(i) $\psi(0) = 0$;
(ii) $\psi(0)$ is $C^1$ on $(0, \zeta)$;
(iii) for all $s \in (0, \zeta)$, $\psi'(s) > 0$;
(iv) for all $z \in U$, $|\partial f(z)| \leq \zeta \psi(|z - \bar{z}|)$.

Finally, we give the definition of coerciveness.

**Definition 4.5** (Coerciveness). A proper real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is said to be coercive if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.

Then we combine the Kurdyka-Lojasiewicz property and coerciveness to conclude the convergence of Algorithm 1.
is definable in an o-minimal structure $O$ where (4.4) $E$

**Definition 4.5** (o-minimal structure). Let $O = \{O_n\}_{n \in \mathbb{N}}$ be such that each $O_n$ is a collection of subsets of $\mathbb{R}^n$. The family $O$ is o-minimal structure over $\mathbb{R}$, if it satisfies the following axioms:

(i) Each $O_n$ is a boolean algebra. Namely $\emptyset \in O_n$ and for each $A, B$ in $O_n$, $A \cup B, A \cap B$ and $\mathbb{R}^n \setminus A$ belong to $O_n$.

(ii) For all $A$ in $O_n, A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $O_{n+1}$.

(iii) For all $A$ in $O_{n+1}, \prod(A) := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : (x_1, \ldots, x_n, x_{n+1}) \in A\}$ belongs to $O_n$.

(iv) For all $i \neq j$ in $\{1, \ldots, n\}, \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i = x_j\}$ in $O_n$.

(v) The set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 < x_2\}$ belongs to $O_2$.

(vi) The elements of $O_1$ are exactly finite unions of intervals.

Let $O$ be an o-minimal structure. A set $S \subset \mathbb{R}^p$ is said to be definable (in $O$), if $S$ belongs to $O_p$. A real extended valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is said to be definable if its graph is a definable subset of $\mathbb{R}^p \times \mathbb{R}$.

The content of definable functions is very rich. Most of the functions arising in real optimization problems are definable. Here we list some known examples and properties of definable functions which help identify $H(x)$ [see 2, 21, for more information].

- indicator functions of real intervals are definable;
- real polynomial functions are definable;
- The restrictions of analytic functions to real bounded intervals are definable;
- finite sums and compositions of definable functions are definable;
- functions of the type $f(x) = \sup_{g \in S} g(x, y)$ (resp. $f(x) = \inf_{g \in S} g(x, y)$) where the function $g$ and the set $S$ are definable, are definable.

Next we present the theorem in [5] that states the definable functions which are proper and lower semi-continuous satisfy the KL property. We say that an extended real-valued function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous if

$$\lim_{x \to x_0} \inf f(x) \geq f(x_0),$$

at every point $x_0$ in its domain. It is obvious that the singleton MMSE penalty $\varphi^{\text{MMSE}}_\sigma(x)$ is semi-continuous and so is $H(x)$.

**Theorem 4.6.** [5] Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a proper and lower semi-continuous function. If $f$ is definable in an o-minimal structure $O$, then it has the KL property at each point of its domain.

We define the effective domain of $H(x)$ as

$$(4.4) \quad \mathcal{E} = \prod_{i=1}^m (r_i, u_i),$$

where $(r_i, u_i)$ is the effective domain of $\varphi^{\text{MMSE}}_{\sigma/\|a_i\|_2}$, i.e. the domain of $\varphi^{\text{MMSE}}_{\sigma/\|a_i\|_2}(x)$. Next we prove $H(x)$ is definable and thus satisfies the KL property at each point of $\mathcal{E}$.

**Lemma 4.7.** Assume $p(x)$ satisfies Assumption 1. The objective function $H(x)$ of the variational model (4.1) has the KL property at each point of $\mathcal{E}$.

**Proof.** To prove $H(x)$ has the KL property at each point of $\mathcal{E}$, it is sufficient to show it is definable on $\mathcal{E}$ by Theorem 4.6. Since $H(x)$ is the sum of an $\ell_2$ fidelity and singleton MMSE penalty for each coordinate, we only need to prove the singleton MMSE penalty $\varphi^{\text{MMSE}}_\sigma(x)$ is definable on its effective domain.
Without loss of generality, we only prove the case where the effective domain of \( \sigma_{\text{MMSE}}(x) \), i.e. the domain of \( \sigma_{\text{MMSE}} \) is \( \mathbb{R} \). Let \( t(x) = \int_{0}^{x} r_{\sigma_{\text{MMSE}}}(u) du \). As \( \sigma_{\text{MMSE}}(x) = t(x) - \frac{1}{2}x^2 \), it is enough to prove \( t(x) \) is definable. Since \( r_{\sigma_{\text{MMSE}}}(x) \) is analytic, \( t(x) \) is also analytic. Moreover, \( r_{\sigma_{\text{MMSE}}}(x) \) is strictly increasing by Proposition 2.2 and its range is the whole real domain \( \mathbb{R} \). Then, there exists a positive value \( K \) such that \( r_{\sigma_{\text{MMSE}}}(x) > 0 \) (resp. \( r_{\sigma_{\text{MMSE}}}(x) < 0 \)) when \( x > K \) (resp. \( x < -K \)). So \( t(x) \) increases when \(|x|\) is greater than \( K \) and goes to infinity. Hence, \( t(x) \) can be represented as

\[
t(x) = \sup_{y_1 \in (K, +\infty), y_2 \in (-\infty, -K)} \left\{ t(y_1) - t(y_2) \right\}.
\]

By the given arguments about definable functions, we can easily get that \( t(x) \) is definable. 

The coerciveness of the objective function guarantees the boundedness of the generated sequence if the objective function value is finite. We give the definition of coerciveness as follows.

**Definition 4.8 (Coerciveness).** A real extended valued function \( f : \mathbb{R}^p \to \mathbb{R} \cup \{\pm \infty\} \) is called coercive iff \( f(x) \to +\infty \) as \( \|x\| \to +\infty \).

Finally, we prove \( H(x) \) is coercive with Assumption 2 in the following lemma.

**Lemma 4.9.** Assume the prior \( p(x) \) satisfies Assumption 2. The objective function \( H(x) \) of the variational model (4.1) is coercive.

**Proof.** Note that \( H(x) \) includes a separable MMSE penalty which is the sum of the singleton MMSE penalty for every coordinate. Once the coerciveness of the singleton MMSE penalty \( \sigma_{\text{MMSE}}(x) \) is proved, we can obtain that \( H(x) \) is coercive. Next, we focus on the proof of the coerciveness of \( \sigma_{\text{MMSE}}(x) \). When the domain of \( r_{\sigma_{\text{MMSE}}} \) is not \( \mathbb{R} \), \( \sigma_{\text{MMSE}}(x) \) is defined to be positive infinity outside. So we only need to prove the case where the domain of \( r_{\sigma_{\text{MMSE}}} \) is \( \mathbb{R} \) and

\[
\sigma_{\text{MMSE}}(x) = \int_{0}^{x} (r_{\sigma_{\text{MMSE}}}(u) - u) du, \quad x \in \mathbb{R}.
\]

We only show that \( \sigma_{\text{MMSE}}(x) \to +\infty \) as \( x \to +\infty \), because the case is similar when \( x \) goes to \( -\infty \).

Supposing \( r_{\sigma_{\text{MMSE}}}(x) > x + \xi \) for some \( \xi > 0 \) when \( x \) is sufficiently large, immediately we get that \( \sigma_{\text{MMSE}}(x) \to +\infty \) as \( x \to +\infty \). Since \( r_{\sigma_{\text{MMSE}}}(x) \) is the inverse function of \( \sigma_{\text{MMSE}} \), the condition \( r_{\sigma_{\text{MMSE}}}(x) > x + \xi \) for sufficiently large \( x \) can be guaranteed by showing that \( \sigma_{\text{MMSE}}(z) < z - \xi \) for sufficiently large \( z \). In the following, we will show there exist \( \xi > 0 \) and \( K > 0 \) such that \( \sigma_{\text{MMSE}}(z) < z - \xi \) when \( z > K \).

Recall

\[
\sigma_{\text{MMSE}}(z) = \frac{\int_{-\infty}^{x} xp(x) \phi_{\sigma}(x - z) dx}{\int_{-\infty}^{x} p(x) \phi_{\sigma}(x - z) dx},
\]

where \( \phi_{\sigma} \) is the probability density function of the Gaussian distribution with mean zero and variance \( \sigma^2 \). We separate the involved integrals in (4.5) into two parts

\[
\sigma_{\text{MMSE}}(z) = \frac{\int_{-\infty}^{\tilde{z}} xp(x) \phi_{\sigma}(x - z) dx}{\int_{-\infty}^{x} p(x) \phi_{\sigma}(x - z) dx} + \frac{\int_{\tilde{z}}^{x} xp(x) \phi_{\sigma}(x - z) dx}{\int_{-\infty}^{x} p(x) \phi_{\sigma}(x - z) dx}.
\]

If we can prove there exist \( \xi > 0 \), \( K > 0 \) and \( \tilde{z} \) such that

\[
\frac{\int_{-\infty}^{\tilde{z}} xp(x) \phi_{\sigma}(x - z) dx}{\int_{-\infty}^{x} p(x) \phi_{\sigma}(x - z) dx} < z - 2\xi, \quad z > K',
\]

and

\[
\frac{\int_{\tilde{z}}^{x} xp(x) \phi_{\sigma}(x - z) dx}{\int_{-\infty}^{x} p(x) \phi_{\sigma}(x - z) dx} < \xi, \quad z > K'','
\]

we obtain \( \sigma_{\text{MMSE}}(z) < z - \xi \) when \( z > K = \max\{K', K''\} \). Next we focus on the proof of the inequality (4.6) and (4.7).
To show the inequality (4.6), we need to employ Assumption 2. It says that there exist $K' > 0$ and $b > 0$ such that
\[ p(x) = a(x)\exp(-bx), \quad x > K', \]
where $a(x) \geq 0$ monotonously decreases to zero when $|x|$ goes to infinity. We choose
\[ \xi = \frac{b\sigma^2}{2}, \quad \tau = \exp(-\frac{z^2 - (z - 2\xi)^2}{2\sigma^2}), \quad \tilde{z} = 2(z - 2\xi) - K', \]
and assume $\tilde{z} > K'$. Then we have
\[
\int_{K'}^{\tilde{z}} (x - z + 2\xi) p(x) \phi_\sigma(x - z) dx \\
= \int_{K'}^{\tilde{z}} (x - z + 2\xi) a(x) \exp(-bx) \phi_\sigma(x - z) dx \\
= \tau \int_{K'}^{\tilde{z}} (x - z + 2\xi) a(x) \phi_\sigma(x - z + 2\xi) dx \\
= \tau \int_{K' - (z - 2\xi)}^{\tilde{z} - (z - 2\xi)} x a(x + z - 2\xi) \phi_\sigma(x) dx \\
= \tau \int_{0} x (a(x + z - 2\xi) - a(-x + z - 2\xi)) \phi_\sigma(x) dx \\
\leq 0.
\]
That is,
\[
\int_{K'}^{\tilde{z}} x p(x) \phi_\sigma(x - z) \leq (z - 2\xi) \int_{K'}^{\tilde{z}} p(x) \phi_\sigma(x - z).
\]
It is obvious that
\[
\int_{-\infty}^{K'} x p(x) \phi_\sigma(x - z) \leq K' \int_{-\infty}^{K'} p(x) \phi_\sigma(x - z) < (z - 2\xi) \int_{-\infty}^{K'} p(x) \phi_\sigma(x - z).
\]
Combining the above two inequalities, we get that
\[
\frac{\int_{-\infty}^{\tilde{z}} x p(x) \phi_\sigma(x - z) dx}{\int_{-\infty}^{\tilde{z}} p(x) \phi_\sigma(x - z) dx} = \frac{\int_{-\infty}^{K'} K' x p(x) \phi_\sigma(x - z) dx + \int_{K'}^{\tilde{z}} p(x) \phi_\sigma(x - z) dx}{\int_{-\infty}^{K'} p(x) \phi_\sigma(x - z) dx + \int_{K'}^{\tilde{z}} p(x) \phi_\sigma(x - z) dx} < z - 2\xi.
\]
Then, it is easy to obtain the inequality (4.6):
\[
\frac{\int_{-\infty}^{\tilde{z}} x p(x) \phi_\sigma(x - z) dx}{\int_{-\infty}^{\tilde{z}} p(x) \phi_\sigma(x - z) dx} \leq \min \left\{ 0, \frac{\int_{-\infty}^{\tilde{z}} x p(x) \phi_\sigma(x - z) dx}{\int_{-\infty}^{\tilde{z}} p(x) \phi_\sigma(x - z) dx} \right\} < z - 2\xi.
\]
Next we prove the inequality (4.7). If $a(\tilde{z}) = 0$, which means that $a(x) = 0$ for all $x \in (\tilde{z}, +\infty)$, the proof is done; otherwise, since $a(x)$ decreases when $x > K'$, we have
\[
\frac{\int_{-\tilde{z}}^{\tilde{z}} x a(x) \phi_\sigma(x - z + 2\xi) dx}{\int a(x) \phi_\sigma(x - z + 2\xi) dx} \leq \frac{\int_{-\tilde{z}}^{\tilde{z}} x a(x) \phi_\sigma(x - z + 2\xi) dx}{\int_{-\tilde{z}}^{\tilde{z}} a(x) \phi_\sigma(x - z + 2\xi) dx} \\
\leq \frac{\int_{-\tilde{z}}^{\tilde{z}} x \phi_\sigma(x - z + 2\xi) dx}{\int_{-\tilde{z}}^{\tilde{z}} \phi_\sigma(x - z + 2\xi) dx} \\
= \frac{\sigma^2 \phi_\sigma((z - 2\xi) - K) + (z - 2\xi) \int_{(z - 2\xi) - K}^{\infty} \phi_\sigma(x) dx}{\int_{(z - 2\xi) - K}^{\infty} \phi_\sigma(x) dx}.
\]
By L'Hôpital's rule, we obtain
\[
\lim_{z \to +\infty} (z - 2\xi) \int_{(z-2\xi)-K}^{\infty} \phi_\sigma(x) dx = \lim_{z \to +\infty} (z - 2\xi)^2 \phi_\sigma((z - 2\xi) - K) = 0.
\]
Moreover, it is easy to see that
\[
\lim_{z \to +\infty} \sigma^2 \phi_\sigma((z - 2\xi) - K) = 0, \quad \text{and} \quad \lim_{z \to +\infty} \int_{(z-2\xi)-K}^{\infty} \phi_\sigma(x) dx = 1.
\]
Then there exists \( K'' > 0 \) such that
\[
\frac{\sigma^2 \phi_\sigma((z - 2\xi) - K) + (z - 2\xi) \int_{(z-2\xi)-K}^{\infty} \phi_\sigma(x) dx}{\int_{(z-2\xi)-K}^{\infty} \phi_\sigma(x) dx} < \xi, \quad \text{when } z > K''.
\]
So we complete the proof of the inequality (4.7).

4.2. Proof of the main theorem. In order to prove the convergence of Algorithm 1, we need to employ the following theorem that is given in [1].

**Theorem 4.10** ([1]). Let \( f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) be a semi-continuous function. Consider a sequence \( \{x^k\}_{k \in \mathbb{N}} \) that satisfies

1. **(Sufficient decrease condition).** For each \( k \in \mathbb{N} \),
   \[
   f(x^{k+1}) + a\|x^{k+1} - x^k\|^2 \leq f(x^k), \quad a > 0;
   \]
   2. **(Relative error condition).** For each \( k \in \mathbb{N} \), there exists \( w^{k+1} \in \partial f(x^{k+1}) \) such that
   \[
   \|w^{k+1}\|_2 \leq b\|x^{k+1} - x^k\|_2, \quad b > 0;
   \]
   3. **(Continuity condition).** There exists a subsequence \( \{x^{k_j}\}_{j \in \mathbb{N}} \) and such that
   \[
   x^{k_j} \to \hat{x} \text{ and } f(x^{k_j}) \to f(\hat{x}), \quad \text{as } j \to +\infty.
   \]

If \( f \) has the KL property at \( \hat{x} \), then the sequence \( \{x^k\}_{k \in \mathbb{N}} \) converges to \( \hat{x} \) as \( k \) goes to infinity, and \( \hat{x} \) is a stationary point of \( f \).

To apply Theorem 4.10, we need to show our objective function \( H(x) \) has the KL property and satisfies the conditions C1-C3. The KL property of \( H(x) \) is proven in Lemma 4.7 and the coerciveness in Lemma 4.9. The condition C3 follows from the coerciveness and continuity of \( H(x) \). The rest of the proof of Theorem 3.1 is given below.

**Proof.** Let \( \{x^k\}_{k \in \mathbb{N}} \) be the sequence generated by Algorithm 1. As only one coordinate is updated between two successive iterates of \( \{x^k\}_{k \in \mathbb{N}} \), we consider the subsequence \( \{x^{lp}\}_{l \in \mathbb{N}} \) in every cycle of \( p \) iterations such that all coordinates is updated once after one cycle. Then for each \( k' \in \{1,2,\ldots,p\} \), we have
\[
\|x^{lp+k'} - x^{lp}\|_2 \leq \|x^{(l+1)p}\|_2 - x^{lp}\|_2.
\]
Thus the convergence of \( \{x^k\}_{k \in \mathbb{N}} \) can be guaranteed by that of the subsequence \( \{x^{lp}\}_{l \in \mathbb{N}} \). In the following, we focus on the proof of the convergence of the subsequence \( \{x^{lp}\}_{l \in \mathbb{N}} \). Since \( H(x) \) is semi-continuous and satisfies the KL property (by Lemma 4.7) at each point of \( \mathcal{E} \), to obtain the convergence of \( \{x^{lp}\}_{l \in \mathbb{N}} \) by Theorem 4.10, we only need to prove it satisfies C1 (sufficient decrease condition), C2 (relative error condition), C3 (Continuity condition) and the cluster point \( \hat{x} \) in C3 belongs to \( \mathcal{E} \).

Firstly, we show \( \{x^{lp}\}_{l \in \mathbb{N}} \) satisfies the condition C1. As the coordinate minimization algorithm only updates one coordinate once, we consider the singleton objective function:
\[
H_{ik}^k(x) = \|a_{ik}\|^2 \|x - \frac{1}{\|a_{ik}\|^2} a_{ik}^T(z - \sum_{j \neq i_k} a_j x_j^k)\|^2 + \|a_{ik}\|^2 \sigma_\text{MMSE}^2\|a_{ik}\|^2\|x - a_{ik}^\text{MMSE}(x).
\]
Denote $\Delta = \min_{i \in \{1, 2, \ldots, p\}} \inf_{x \in \mathbb{R}} \|a_i\|_2^2 / \|dS^{MMSE}_{\sigma/a_i}(x)\|$. By Proposition 2.4, we know that $H_{ik}^k(x)$ is $\Delta$-strongly convex on the range of $S^{MMSE}_{\sigma/a_i}$ for all $i_k \in \{1, 2, \ldots, p\}$. Moreover, because $x_{ik}^{k+1}$ is the minimizer of $H_{ik}^k(x)$, it yields that

$$0 \in \partial H_{ik}^k(x_{ik}^{k+1}).$$

Since the components of $\{x^k\}_{k \in \mathbb{N}}$ is computed by singleton MMSE estimators, they always belong to the range of their corresponding singleton MMSE estimators. Then by the strong convexity of $H_{ik}^k(x)$, we have

$$H_{ik}^k(x_{ik}^{k+1}) + \frac{\Delta}{2} \|x_{ik}^{k+1} - x_{ik}^k\|^2 \leq H_{ik}^k(x_{ik}^k).$$

As the remaining coordinates except $i_k$ are fixed, it also gives that

$$H(x^{k+1}) + \frac{\Delta}{2} \|x_{ik}^{k+1} - x_{ik}^k\|^2 \leq H(x^k).$$

Summing up the inequality (4.10) from $k = lp$ to $k = (l+1)p$, we obtain that $\{x^{lp}\}_{l \in \mathbb{N}}$ satisfies the condition C1:

$$\frac{\Delta}{2} \|x^{(l+1)p} - x^{lp}\|^2 = \frac{\Delta}{2} \sum_{k=0}^{p-1} \|x_{ik}^{lp+k+1} - x_{ik}^{lp+k}\|^2 \leq \sum_{k=0}^{p-1} (H(x^{lp+k}) - H(x^{lp+k+1})) = H(x^{lp}) - H(x^{(l+1)p}).$$

Next we show the subsequence $\{x^{lp}\}_{l \in \mathbb{N}}$ also satisfies C2. Since

$$x_{ij}^{lp+k+1} = \begin{cases} x_{ij}^{(l+1)p}, & \text{if } j \leq k; \\ x_{ij}^{lp}, & \text{if } j > k, \end{cases}$$

we can get that

$$\partial x_{ik} H(x^{(l+1)p}) = a_{ik}^T(Ax^{(l+1)p} - z) + \partial x_{ik} \epsilon_{\sigma/\|a_i\|_2}(x_{ik}^{(l+1)p})$$

$$= a_{ik}^T A(x^{(l+1)p} - x^{lp+k+1}) + \partial x_{ik} H_{ik}^{lp+k}(x_{ik}^{lp+k+1})$$

$$= a_{ik}^T \sum_{j > k} a_{ij} (x_{ij}^{(l+1)p} - x_{ij}^{lp}) + \partial x_{ik} H_{ik}^{lp+k}(x_{ik}^{lp+k+1}).$$

Combining with (4.9), we have

$$a_{ik}^T \sum_{j > k} a_{ij} (x_{ij}^{(l+1)p} - x_{ij}^{lp}) \in \partial x_{ik} H(x^{(l+1)p}),$$

and thus

$$w^{(l+1)p} = S(x^{(l+1)p} - x^{lp}) \in \partial H(x^{(l+1)p}),$$

where the matrix $S$ takes the form of

$$S_{st} = \begin{cases} (A^T A)_{st}, & \text{if } i_t > i_s; \\ 0, & \text{otherwise}. \end{cases}$$

Taking $\ell_2$-norm on both sides of (4.12), we get that

$$||w^{(l+1)p}||_2 \leq \rho_{\text{max}}(S)||x^{(l+1)p} - x^{lp}||_2.$$

That is, the subsequence $\{x^{lp}\}_{l \in \mathbb{N}}$ satisfies C2.

Lastly, we prove the condition C3 and $\hat{x}$ belongs to $E$. Since $H(x_0) < +\infty$ and $H(x^{lp})$ is decreasing, we have $H(x^{lp})$ is always finite. Together with the fact $H(x)$ is coercive by Lemma 4.9, we can get that the sequence $\{x^{lp}\}_{l \in \mathbb{N}}$ is bounded and thus there exists a convergent subsequence of $\{x^{lp}\}_{l \in \mathbb{N}}$, e.g. $\{x^k\}_{k \in \mathbb{N}}$ that converges to $\hat{x}$. Because the coordinates of the sequence $\{x^{lp}\}_{l \in \mathbb{N}}$ are obtained by the singleton MMSE estimator, $\{x^{lp}\}_{l \in \mathbb{N}}$ always belongs to $E$. Thus, $\hat{x}$ belongs
to the closure of $E$. Moreover, $H(x)$ is continuous on the closure of $E$ if $H(x) < +\infty$, so
$$\lim_{j \to +\infty} H(x^k_j) = H(\hat{x}).$$
Then the condition C3 is satisfied. Taking limits on (4.13), we have
$$\lim_{j \to +\infty} w^{k_j} = 0 \in \partial H(\hat{x}).$$
By Remark 4.3, we obtain $\hat{x} \in E$ and $H(x)$ has the KL property at $\hat{x}$.

Finally, by Theorem 4.10, we can conclude that the subsequence $\{x^{p_l}\}_{l \in \mathbb{N}}$ converges to $\hat{x}$, and $\hat{x}$ is a stationary point of $H(x)$. Thus, the whole sequence $\{x^k\}_{k \in \mathbb{N}}$ also converges to $\hat{x}$. Furthermore, since $H(x)$ is strongly convex along arbitrary single coordinate by fixing the remaining ones on $E$, $\hat{x} \in E$ always attains the minimum of $H(x)$ with respect to arbitrary single coordinate, which means that it attains the minimum of the mean square error with respect to arbitrary single coordinate as well. 

\section{Conclusion}
For Bayesian method, the minimum mean square error estimator makes a good balance between bias and variance and has the posterior mean as its explicit form. However, it is hard to compute numerically, since it involves multiple integrals of many variables. This paper proposes a simple iterative algorithm to approximate the MMSE estimator, which is easy to implement and efficient in numerical experiments. A complete convergence analysis is given. The analysis and algorithm developed here are then applied to a few given prior distributions with Gaussian noise. In particular, we give a complete analysis and implementation details for estimation under the Bernoulli-uniform sparse prior assumption. We also compare with other available approaches, e.g. the MAP method. Among many properties stated in this paper, our approach, for example, gives a stable estimator balancing unbiasedness and sparsity.

\section*{References}
REFERENCES


DEPARTMENT OF MATHEMATICS, NATIONAL UNIVERSITY OF SINGAPORE, SINGAPORE 117543, SINGAPORE
E-mail address: pangtongyao@nus.edu.sg

DEPARTMENT OF MATHEMATICS, NATIONAL UNIVERSITY OF SINGAPORE, SINGAPORE 117543, SINGAPORE
E-mail address: matzuows@nus.edu.sg