Optimal Bayesian Regression for Serially Dependent Training Observations

Samira Reihanian and Amin Zollanvari, *Senior Member, IEEE*

Electrical and Computer Engineering Department, School of Engineering and Digital Sciences,
Nazarbayev University, Kazakhstan
Emails: samira.reihanain@nu.edu.kz, amin.zollanvari@nu.edu.kz

***Abstract*—In this study, we construct, for the first time, the optimal Bayesian Regression (OBR) when the training data are serially dependent. To model the effect of dependency, we assume the training data are generated from VAR(p), which is a multi-dimensional vector autoregressive process of order p. Our numerical experiments show that when the training samples follow a VAR dependency model, the proposed regressor can outperform the usual regressor, which is developed under sample independence assumption.**

Optimal Bayesian Regression, Vector Autoregressive Processes, Serially Dependent Training Data

I. Introduction

Suppose $\mathbf{x}$ and y denote a $(q-1)$-dimensional feature vector (also known as predictors) and a target variable (or response), respectively. The goal in decision theory is to choose an operator $\psi(\mathbf{x})$ to estimate y for each input vector $\mathbf{x}$ [1, p.46]. In contrast with classical statistical estimation in which y is considered as a deterministic unknown constant, in Bayesian estimation y is considered as a random variable with a joint distribution $f(\mathbf{x}, y)$ with random vectors $\mathbf{x}$ [2]. To determine the optimal operator $\psi_{\text{o}}(\mathbf{x})$, it is common to minimize the mean square error (MSE) resulting in MMSE estimator given by

$$\begin{aligned}\psi_{\text{o}}(\mathbf{x}) &= \underset{\psi\in\mathcal{F}}{\operatorname{argmin}}\, \mathrm{E}_f\left[(y-\psi(\mathbf{x}))^2\right] \\ &= \underset{\psi\in\mathcal{F}}{\operatorname{argmin}} \int\int (y-\psi(\mathbf{x}))^2 f(\mathbf{x},y) d\mathbf{x}\, dy\,, \quad (1)\end{aligned}$$

where $\mathcal{F}$ denotes the class of all operators. The criterion (1) is also known as the Bayesian mean square error to distinguish that from the conventional (non-Bayesian) MSE [2, p.311]. It is well-known that the $\psi_{\text{o}}(\mathbf{x})$ is obtained by [2]

$$\psi_{\text{o}}(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}] = \int y\, f(y|\mathbf{x})\, dy\,. \quad (2)$$

However, if $f(\mathbf{x}, y)$ is a multivariate Gaussian density function with mean $\mathbf{m} = [\mathbf{m}_{\mathbf{x}}^T \;\; m_y]^T$ and covariance $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \mathbf{\Sigma}_{y\mathbf{x}}^T \\ \mathbf{\Sigma}_{y\mathbf{x}} & \mathbf{\Sigma}_{yy} \end{pmatrix}$ where $\mathbf{m}_{\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{xx}}$ are $(q-1)$ and $(q-1)\times(q-1)$ dimensional vector and matrix, respectively, then [2, p.324]

$$\psi_{\text{o}}(\mathbf{x}) = m_y + \mathbf{\Sigma}_{y\mathbf{x}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x}-\mathbf{m}_{\mathbf{x}})\,, \quad (3)$$

which is a linear function of $\mathbf{x}$. The goal in the inference stage is to learn $f(\mathbf{x}, y)$ given a training data $\mathcal{S}$. For (3), the inference stage includes replacing m_y, $\mathbf{\Sigma}_{y\mathbf{x}}$, and $\mathbf{\Sigma}_{\mathbf{xx}}^{-1}$ with their sample estimates to create the Bayes plug-in regression rule.

In recent years, a new framework known as optimal Bayesian regression (OBR) was proposed in [3], and was later extended to transfer regression in [4]. In contrast to Bayesian linear regression (BLR) framework [1, p.152], in OBR no linear mapping is imposed between y and $\mathbf{x}$; and (II) the uncertainty in OBR is directly considered on the joint distribution of y and $\mathbf{x}$. Let $f(\mathbf{x}, y, \boldsymbol{\theta})$ denote the joint distribution of y and $\mathbf{x}$ where the random vector $\boldsymbol{\theta}$ parameterizes the distribution and where the dependence of the distribution on $\boldsymbol{\theta}$ is made explicit by writing $f(., \boldsymbol{\theta})$. In OBR, it is assumed that $f(\mathbf{x}, y, \boldsymbol{\theta})$ belongs to an uncertainty class $\mathbf{\Theta}$ (parameter space) of joint distributions governed by a prior distribution $\pi(\boldsymbol{\theta})$, and we desire a regressor to minimize the expected *error* (defined by taking some criteria) over the uncertainty class. Given a training sample $\mathcal{S}$ and taking the MSE as the criterion, the OBR is defined as [3], [4],

$$\psi_{\text{OBR}}(\mathbf{x}) = \int\int y f(y|\mathbf{x},\boldsymbol{\theta}) dy\, f(\boldsymbol{\theta}|\mathcal{S})\, d\boldsymbol{\theta} = \mathrm{E}_{\pi^*(\boldsymbol{\theta})}\left[\psi_{\text{o}}(\mathbf{x}|\boldsymbol{\theta})\right], \quad (4)$$

where in analogy to (2),

$$\psi_{\text{o}}(\mathbf{x}|\boldsymbol{\theta}) \triangleq \int y f(y|\mathbf{x},\boldsymbol{\theta}) dy\,, \quad (5)$$

and

$$\pi^*(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}|\mathcal{S})\,, \quad (6)$$

is the *posterior* probability density function. In other words, $\pi^*(\boldsymbol{\theta})$ characterizes the updated information about $\boldsymbol{\theta}$ after observing $\mathcal{S}$. From (4), we can also rewrite $\psi_{\text{OBR}}(\mathbf{x})$ as [3]

$$\psi_{\text{OBR}}(\mathbf{x}) = \mathrm{E}_{f_{\mathbf{\Theta}}}[y|\mathbf{x}] = \int y\, f_{\mathbf{\Theta}}(y|\mathbf{x})\, dy\,, \quad (7)$$

where $f_{\mathbf{\Theta}}(\mathbf{x}, y)$ is known as the *effective joint distribution* of $\mathbf{x}$ and y defined as

$$f_{\mathbf{\Theta}}(\mathbf{x}, y) = \int f(\mathbf{x}, y|\boldsymbol{\theta})\, \pi^*(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,. \quad (8)$$

One common aspect among the aforementioned machineries of Bayesian regression is the usual assumption of having independent and identically distributed data. While this assumption might be true under controlled conditions, in general the data might neither be independent nor identically distributed. In such applications, it would be beneficial to treat them as the output of some stochastic process [5], [6], [7], [8], [9], [10], [11].

Recently, we considered the problem of having sequentially dependent data in the context of optimal Bayesian classification (OBC) [9]. To model the effect of dependency, we assumed that the training observations are generated from a VAR(p) process. As in the case of OBC, we assumed the existence of an uncertainty class about the parameters governing the VAR$^y(p)$ model.

In the present investigation, we develop the OBR rule under a similar scenario where we assume the training samples are generated from a VAR(p) process. Similarly, we assume an uncertainty class about the parameters governing the VAR(p) model. To model this uncertainty, we assume that model parameters are random variables with a prior distribution.

Throughout this work, we use boldface lower case letters to denote a column vector. Two special cases are $\mathbf{0}_n$ and $\mathbf{1}_n$, which denote a column vector of length n with elements 0 and 1, respectively. A boldface upper case letter denotes a matrix. A special case is the identity matrix of size n denoted by $\mathbf{I}_n$. For a $n \times m$ matrix $\mathbf{A}$, vec($\mathbf{A}$) denotes the vectorization operator acting on $\mathbf{A}$ such that column of matrix $\mathbf{A}$ are concatenated to produce an nm-dimensional column vector. Conversely, for a nm-dimensional column vector $\mathbf{a}$, mat$_n(\mathbf{a})$ creates an $n \times m$ matrix $\mathbf{A}$ such that its i^{th} column consists of elements from $((i-1)\times n)+1$ to $i \times n$ in $\mathbf{a}$, $i = 1, ..., m$. Therefore, mat$_n$(vec($\mathbf{A}$)) $= \mathbf{A}$. Furthermore, tr(.) and $\otimes$ denote the trace and the Kronecker product, respectively. Furthermore, a multivariate Gaussian density function with mean $\mathbf{a}$ and precision matrix $\mathbf{A}$ (covariance matrix $\mathbf{A}^{-1}$) is denoted as $\mathcal{N}\left(\mathbf{a}, \mathbf{A}^{-1}\right)$.

The paper is organized as follows. In Section II, we formulate the problem of training data generation from VAR processes. Section III presents the main result of the present work. There, we obtain the optimal Bayesian regression rule when the training observations follow VAR(p). In Section IV, the numerical experiments comparing the performance of the developed regression rule are presented. Section V concludes the paper and discusses some of the future directions that can be pursued.

II. Vector Autoregressive Dependency in Training Data

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be n available training (column) feature vectors with corresponding (univariate) response values $y_1, ..., y_n$ taken from a q-dimensional VAR(p) process and defined as,

$$\mathbf{v}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{v}_{t-i} + \mathbf{u}_t, \tag{9}$$

where

$$\mathbf{v}_t = [x_{t,1}, x_{t,2}, ..., x_{t,q-1}, y_t]^T, \tag{10}$$

for $t = 1, ..., n$, $x_{t,i}$ is the i^{th} dimension of vector $\mathbf{x}_t$, $\mathbf{u}_t$ is q-dimensional Gaussian white-noise with $\mathrm{E}[\mathbf{u}_t] = \mathbf{0}$, $\mathrm{E}[\mathbf{u}_t {\mathbf{u}_t}^T] = \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \mathbf{\Sigma}_{y\mathbf{x}}^T \\ \mathbf{\Sigma}_{y\mathbf{x}} & \mathbf{\Sigma}_{yy} \end{pmatrix}$, $\mathrm{E}[\mathbf{u}_t {\mathbf{u}_r}^T] = \mathbf{0}$ for $r \neq t$, $\mathbf{A}_i$ is a $q \times q$ matrix of model parameters, and $\mathbf{c}$ is the vector of q intercept terms. Note that here we assumed $(q-1)$-dimensional feature vectors (rather than q-dimensional) to simplify notation. We can equivalently write (9) as

$$\mathbf{V}_{\mathcal{S}} = \mathbf{A}\mathbf{Z} + \mathbf{U}, \tag{11}$$

where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_n]$, $\mathbf{V}_{\mathcal{S}} = [\mathbf{v}_1, ..., \mathbf{v}_n]$ is a $q \times n$ dimensional matrix, $\mathbf{A} = [\mathbf{c}, \mathbf{A}_1, ..., \mathbf{A}_p]$ is a $q \times (1+pq)$ dimensional matrix of unknown model parameters (similar to [12, p. 224] and [9] we assume known $\mathbf{\Sigma}$), $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_n]$ is a $(1+pq) \times n$ matrix with

$$\mathbf{z}_t = [1, {\mathbf{v}_{t-1}}^T, ..., {\mathbf{v}_{t-p}}^T]^T, \tag{12}$$

where $\mathbf{v}_0 = \mathbf{v}_{-1} = ... = \mathbf{v}_{1-p} = \mathbf{0}_q$.

Let $\mathbf{v}_{\mathcal{S}} = \mathrm{vec}(\mathbf{V}_{\mathcal{S}})$, $\mathbf{a} = \mathrm{vec}(\mathbf{A})$, $\mathbf{L} \triangleq \mathbf{\Sigma}^{-1}$, and $\mathbf{L}^{1/2}$ denote the square root of $\mathbf{L}$ such that $\mathbf{L}^{1/2}\mathbf{L}^{1/2} = \mathbf{L}$. To derive the VAR-OBR, we assume the prior distribution of $\mathbf{a}$ is a multivariate Gaussian with known $(1+pq)q$ dimensional mean vector $\mathbf{m}$ and $(1+pq)q \times (1+pq)q$ precision $\mathbf{\Lambda}$:

$$\pi(\mathbf{a}) \propto |\mathbf{\Lambda}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{a}-\mathbf{m})^T \mathbf{\Lambda} (\mathbf{a}-\mathbf{m})\right). \tag{13}$$

We can express the likelihood as (see [12, p. 223]):

$$f(\mathbf{v}_{\mathcal{S}}|\mathbf{a}) \propto |\mathbf{L}|^{\frac{n}{2}} \exp\Bigg(-\frac{1}{2}\Big((\mathbf{a}-\boldsymbol{\mu})^T \mathbf{\Gamma} (\mathbf{a}-\boldsymbol{\mu}) + (\mathbf{s}-\mathbf{W}\boldsymbol{\mu})^T (\mathbf{s}-\mathbf{W}\boldsymbol{\mu}) \Big) \Bigg), \tag{14}$$

where

$$\mathbf{s} = \left(\mathbf{I}_n \otimes \mathbf{L}^{1/2}\right) \mathbf{v}_{\mathcal{S}}, \ \mathbf{W} = \mathbf{Z}^T \otimes \mathbf{L}^{1/2}, \\ \mathbf{\Gamma} = \mathbf{Z}\mathbf{Z}^T \otimes \mathbf{L}, \ \ \boldsymbol{\mu} = \mathbf{\Gamma}^{-1} (\mathbf{Z} \otimes \mathbf{L}) \mathbf{v}_{\mathcal{S}}. \tag{15}$$

The posterior distribution $\pi^*(\mathbf{a})$ is obtained as

$$\pi^*(\mathbf{a}) = \frac{\pi(\mathbf{a}) f(\mathbf{v}_{\mathcal{S}}|\mathbf{a})}{\int_{\mathcal{R}^{(1+pq)q}} \pi(\mathbf{a}) f(\mathbf{v}_{\mathcal{S}}|\mathbf{a}) d\mathbf{a}}. \tag{16}$$

Similarly to the proof presented in Appendix B of [9], we can write

$$\pi^*(\mathbf{a}) = (2\pi)^{\frac{-(1+pq)q}{2}} |\overline{\mathbf{\Lambda}}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{a}-\overline{\mathbf{m}})^T \overline{\mathbf{\Lambda}} (\mathbf{a}-\overline{\mathbf{m}})\right), \tag{17}$$

where

$$\overline{\mathbf{m}} = \overline{\mathbf{\Lambda}}^{-1} (\mathbf{\Lambda}\,\mathbf{m} + \mathbf{\Gamma}\,\boldsymbol{\mu}), \tag{18}$$

$$\overline{\mathbf{\Lambda}} = \mathbf{\Lambda} + \mathbf{\Gamma} = \begin{pmatrix} \overline{\mathbf{\Lambda}}_1 & \overline{\mathbf{\Lambda}}_2^T \\ \overline{\mathbf{\Lambda}}_2 & \overline{\mathbf{\Lambda}}_3 \end{pmatrix}, \tag{19}$$

in which $\mathbf{m}$ and $\mathbf{\Lambda}$ are the parameters of the prior defined in (13), $\boldsymbol{\mu}$ and $\mathbf{\Gamma}$ are defined in (15), and as shown $\overline{\mathbf{\Lambda}}_1$, $\overline{\mathbf{\Lambda}}_2$, and $\overline{\mathbf{\Lambda}}_3$ partition $\overline{\mathbf{\Lambda}}$ where $\overline{\mathbf{\Lambda}}_1$ is a $q \times q$ matrix.

III. Optimal Bayesian Regression with Vector Autoregressive Data Dependency of Training Data

We consider the problem of estimating the target variable y for a given test feature vector $\mathbf{x}$ when the training data is generated from (9). Although this assumption captures a VAR dependency model between training data, it assumes the test data is collected independently from training data. Similar to [9], we assume the distribution of the test observation is obtained by nullifying $\mathbf{A}_i$ in (9) for $i = 1, ..., p$. Let $\mathbf{v} = [\mathbf{x}^T, y]^T$ denote a test observation. Therefore, we can write

$$\mathbf{v} = \mathbf{A}\mathbf{e}_1 + \mathbf{u}, \tag{20}$$

where $\mathbf{e}_1$ is a $(1 + pq)$-dimensional column vector with 1 at the first position and 0 otherwise. Therefore, we can write

$$\begin{aligned} f(\mathbf{v}|\mathbf{a}, \mathcal{S}) &= \mathcal{N}\left(\mathbf{A}\mathbf{e}_1, \mathbf{L}^{-1}\right) \\ &= \mathcal{N}\left((\mathbf{e}_1{}^T \otimes \mathbf{I}_q)\mathbf{a}, \mathbf{L}^{-1}\right). \end{aligned} \tag{21}$$

At this stage, we use the properties of marginal and conditional Gaussian density function presented in [1, p.93] for deriving the effective joint distribution of $\mathbf{v}$, i.e. $f_{\mathbf{\Theta}}(\mathbf{v}|\mathcal{S})$ using the integral (8). By having $\pi^*(\mathbf{a})$ and $f(\mathbf{v}|\mathbf{a}, \mathcal{S})$, we can write

$$f_{\mathbf{\Theta}}(\mathbf{v}|\mathcal{S}) = f_{\mathbf{\Theta}}(\mathbf{v}) = f_{\mathbf{\Theta}}(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\mathbf{m}^{\text{eff}}, \mathbf{\Sigma}^{\text{eff}}\right), \tag{22}$$

where

$$\mathbf{m}^{\text{eff}} = \left((\mathbf{e}_1{}^T \otimes \mathbf{I}_q)\,\overline{\mathbf{m}} = \overline{\mathbf{M}}\,\mathbf{e}_1\right), \tag{23}$$

$$\mathbf{\Sigma}^{\text{eff}} = \mathbf{\Sigma} + (\mathbf{e}_1{}^T \otimes \mathbf{I}_q)\,\overline{\mathbf{\Lambda}}^{-1}(\mathbf{e}_1 \otimes \mathbf{I}_q), \tag{24}$$

such that

$$\overline{\mathbf{M}} \triangleq \text{mat}_q(\overline{\mathbf{m}}), \tag{25}$$

with $\overline{\mathbf{m}}$ being defined in (18). We can rewrite (23) and (24) as

$$\mathbf{m}^{\text{eff}} = \overline{\mathbf{m}}_{1:q}, \tag{26}$$

$$\mathbf{\Sigma}^{\text{eff}} = \mathbf{\Sigma} + \left(\overline{\mathbf{\Lambda}}_1 - \overline{\mathbf{\Lambda}}_2^T \overline{\mathbf{\Lambda}}_3^{-1} \overline{\mathbf{\Lambda}}_2\right)^{-1}. \tag{27}$$

Replacing (22) in (8) and using standard properties of conditional Gaussian density function [13, p.35] yield

$$\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x}) = m_y^{\text{eff}} + \mathbf{\Sigma}_{y\mathbf{x}}^{\text{eff}} (\mathbf{\Sigma}_{\mathbf{xx}}^{\text{eff}})^{-1} (\mathbf{x} - \mathbf{m}_{\mathbf{x}}^{\text{eff}}), \tag{28}$$

where $\mathbf{m}_{\mathbf{x}}^{\text{eff}}$ and m_y^{eff} are obtained by partitioning $\mathbf{m}^{\text{eff}}$ defined in (26) as

$$\mathbf{m}^{\text{eff}} = [(\mathbf{m}_{\mathbf{x}}^{\text{eff}})^T\ m_y^{\text{eff}}]^T, \tag{29}$$

where $\mathbf{m}_{\mathbf{x}}^{\text{eff}}$ is a $q-1$ dimensional vector (implying m_y^{eff} is the last element of vector $\mathbf{m}^{\text{eff}}$), and in analogy with the block partitioned structure of $\mathbf{\Sigma}$, $\mathbf{\Sigma}^{\text{eff}}$ determined from (27) is partitioned as

$$\mathbf{\Sigma}^{\text{eff}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}}^{\text{eff}} & (\mathbf{\Sigma}_{y\mathbf{x}}^{\text{eff}})^T \\ \mathbf{\Sigma}_{y\mathbf{x}}^{\text{eff}} & \mathbf{\Sigma}_{yy}^{\text{eff}} \end{pmatrix}, \tag{30}$$

in which $\mathbf{\Sigma}_{\mathbf{xx}}^{\text{eff}}$ is a $(q-1) \times (q-1)$ matrix.

IV. Evaluation of the Proposed OBR

In this section, we evaluate the ability of the proposed OBR (i.e. $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ in (28)) to capture the dependence among the samples. To this end, we compare the performance of $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ with respect to $\psi_o(\mathbf{x})$ obtained from (3), in case the training samples are generated from a VAR(p) process. Note that, $\psi_o(\mathbf{x})$ is the result of considering generation of training samples from VAR(0) process (i.e., $p = 0$). By considering $p = 0$ in (9), $f(\mathbf{v}) = \mathcal{N}(\mathbf{c}, \mathbf{\Sigma})$ and therefore the Bayes plug-in regression rule in (3) is applicable.

One practical consideration about the implementation of the proposed OBR regressor in (28) is that it depends on the covariance matrix of the white-noise process. As a result, in our numerical experiments, we consider two cases: I) $\mathbf{\Sigma}$ is known; and II) it is estimated by its least squares estimator (LSE) given by (see p. 75 [12])

$$\hat{\mathbf{\Sigma}} = \frac{1}{(n - pq - 1)} \mathbf{V}_s \left(\mathbf{I}_n - \mathbf{Z}^T (\mathbf{Z}\mathbf{Z}^T)^{-1} \mathbf{Z}\right) \mathbf{V}_s{}^T. \tag{31}$$

In our numerical experiments, we examine:

I. the ability of $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ to capture the dependence among samples, compared to $\psi_o(\mathbf{x})$, which was obtained under the assumption of sample independence;

II. the effect of estimating the white-noise covariance matrix using (31) on regressors performance; and

III. the effect of the order of the underlying VAR(p) processes on the performance of the regressors.

In this regard, we consider two scenarios:

Scenario (a): $q = 3$, $p = 1$, $\mathbf{c} = \mathbf{0}_q$ and

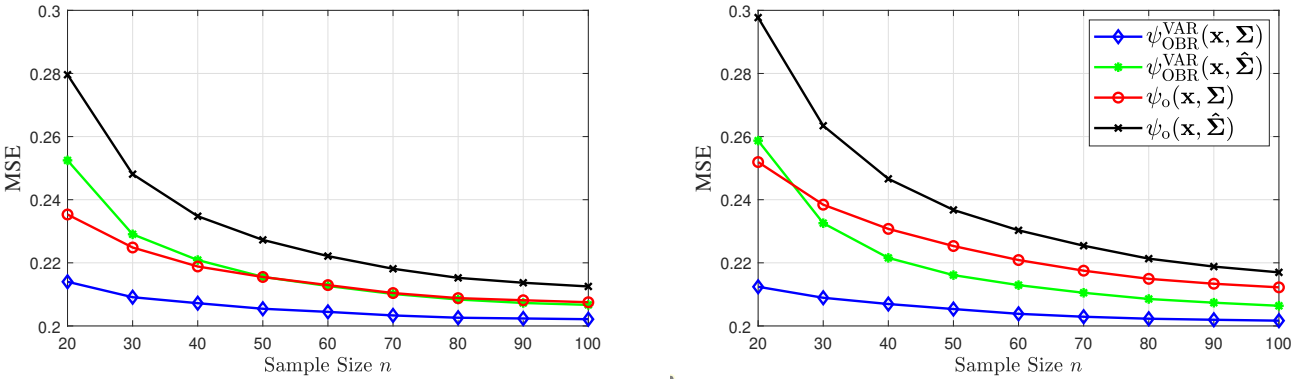
$$\mathbf{A}_1 = \begin{bmatrix} 0.8 & 0 & 0 \\ 0.2 & 0.4 & 0.2 \\ 0 & 0 & 0.5 \end{bmatrix}. \tag{32}$$

Scenario (b): $p = 2$,

$$\mathbf{A}_2 = \begin{bmatrix} -0.9 & 0 & 0 \\ 0 & 0.1 & 0 \\ -0.2 & 0.2 & 0 \end{bmatrix}.$$

and all other experimental parameters similar to scenario (a). We set the prior means (i.e. $\mathbf{m}$) using parameters of $\mathbf{A}_i$. In particular, $\mathbf{m} = \mathbf{a} = \text{vec}(\mathbf{A})$, where $\mathbf{A} = [\mathbf{c}, \mathbf{A}_1, ..., \mathbf{A}_p]$. Then, as for the covariance matrix of the prior and the white-noise process, we use a diagonal matrix with 0.1 and 0.2 as diagonal elements, respectively. In order to estimate the mean squared error of each regressor, we generate $n \in [20, 100]$ training samples from VAR(p) process and 1,000 independent test observations and determine the prediction error rate. The procedure of generating training observations and independent test observations is repeated 500 times to estimate the mean square error of each regressor for each n. Figure 1 shows the performances of $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ and $\psi_o(\mathbf{x})$ as a function of training sample size n in scenario (a) and (b), when the actual covariance matrix of the white-noise process is known (i.e. $\mathbf{\Sigma}$) or estimated (i.e. $\hat{\mathbf{\Sigma}}$). In particular the results show:

I. in case of existence of sample dependence, in both scenarios, the proposed $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ leads to smaller mean squared error



(a) (b)

Fig. 1: Mean squared error of the regressors with and without sample dependence assumptions (i.e. $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ and $\psi_o(\mathbf{x})$) as a function of training sample size n, when the actual covariance matrix of the white-noise process is known (i.e. $\boldsymbol{\Sigma}$) or estimated (i.e. $\hat{\boldsymbol{\Sigma}}$). (a)-(b) depicts the results of Scenario (a)-(b), respectively.

rate compared to $\psi_o(\mathbf{x})$, which assumes sample independence among observations.

II. as we expected, estimating the white-noise covariance matrix using (31) generally degrades performance of both regressors (i.e. $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ and $\psi_o(\mathbf{x})$) with respect to the use of the actual covariance matrix in the structure of the regressors. This degradation is more pronounced when the number of the training samples is small. However, the effect of estimating this covariance matrix on the performance of the regressors is mediated as the sample size increases.

III. by increasing the order of the underlying VAR(p), difference between the performance of the regressors with and without sample dependence assumptions (i.e. $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ and $\psi_o(\mathbf{x})$) is more obvious. Such that, the performances of the $\psi_o(\mathbf{x}, \boldsymbol{\Sigma})$ and $\psi_o(\mathbf{x}, \hat{\boldsymbol{\Sigma}})$ compared to the $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x}, \boldsymbol{\Sigma})$ and $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x}, \hat{\boldsymbol{\Sigma}})$ in scenario (b) are much more weaker than scenario (a).

V. Discussion

In this work, we developed $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$, which incorporates prior knowledge into regressor construction when the training observations are serially dependent. Our initial numerical experiments presented in this study confirms the capability of the proposed regressor in capturing VAR dependency among training samples and outperforms $\psi_o(\mathbf{x})$ regressor obtained under the assumption of sample independence. As with any Bayesian technique, the performance of $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ highly depends on the strength of the prior knowledge, which is integrated into the regressor through the prior distribution. In this paper, to construct the $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$, we assumed that we have the underlying actual model and we use knowledge of the underlying parameters of VAR(p) processes used to generate the data. However, such knowledge is not generally available in practice. In practical applications, we have little or no prior knowledge about the process and the parameters. Therefore, the future work will focus on developing $\psi_{\text{OBR}}^{\text{VAR}}(\mathbf{x})$ when we have no or little knowledge about the prior and examine the performance of the developed regressor on real data sets.

Acknowledgment

This work is supported by the Faculty Development Competitive Research Grants Program of Nazarbayev University under grant number 021220FD1151 (A. Zollanvari).

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

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

[3] X. Qian and E. R. Dougherty, "Bayesian regression with network prior," *IEEE Trans. Sig. Proc.*, vol. 64, pp. 6243–6253, 2016.

[4] A. Karbalayghareh, X. Qian, and E. R. Dougherty, "Optimal bayesian transfer regression," *IEEE Signal Processing Letters*, vol. 25, pp. 1655–1659, 2018.

[5] C. R. O. Lawoko and G. J. McLachlan, "Asymptotic error rates of the w and z statistics when the training observations are dependent," *Pattern Recogn.*, vol. 19, pp. 467–471, 1986.

[6] A. Zollanvari and E. R. Dougherty, "Analytical study of performance of linear discriminant analysis in stochastic settings detection," *Pattern Recogn*, vol. 46, p. 30173029, 2013.

[7] C. R. O. Lawoko and G. J. McLachlan, "Some asymptotic results on the effect of autocorrelation on the error rates of the sample linear discriminant function," *Pattern Recogn.*, vol. 16, pp. 119–121, 1983.

[8] J. D. Tubbs, "Effect ofautocorrelated training samples on bayes' probability of misclasslficatlon," *Pattern Recogn.*, vol. 12, pp. 351–354, 1980.

[9] A. Zollanvari and E. R. Dougherty, "Optimal bayesian classification with vector autoregressive data dependency," *IEEE Trans. Sig. Proc.*, vol. 67, no. 12, pp. 3073–3086, 2019.

[10] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Prentice Hall, 2013.

[11] A. Zollanvari and E. R. Dougherty, "Optimal bayesian classification when the training observations are serially dependent," in *In IEEE 2018 New York Scientific Data Summit (NYSDS)*, 2018, pp. 1–3.

[12] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, 2005.

[13] H. Tong, *Non-linear time series: a dynamical system approach*. Oxford university press, 1990.