Online Robust Reduced-Rank Regression

Yangzhuoran Fin Yang
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia
fin.yang@monash.edu

Ziping Zhao
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
zhaoziping@shanghaitech.edu.cn

***Abstract*—The reduced-rank regression (RRR) model is widely used in data analytics where the response variables are believed to depend on a few linear combinations of the predictor variables, or when such linear combinations are of special interest. In this paper, we will address the RRR model estimation problem by considering two targets which are popular especially in big data applications: i) the estimation should be robust to heavy-tailed data distribution or outliers; ii) the estimation should be amenable to large-scale data sets or data streams. In this paper, we address the robustness via the robust maximum likelihood estimation procedure based on Cauchy distribution and a stochastic estimation procedure is further adopted to deal with the large-scale data sets. An efficient algorithm leveraging on the stochastic majorization minimization method is proposed for problem-solving. The proposed model and algorithm is validated numerically by comparing with the state-of-the-art methods.**

***Index Terms*—Multivariate regression, low-rank, heavy-tails, outliers, stochastic optimization, majorization minimization, large-scale optimization, adaptive algorithm.**

I. Introduction

The reduced-rank regression (RRR) [1], [2] model is a multivariate linear regression model where the coefficient matrices can be reduced-rank (a.k.a. low-rank). The concept of RRR was first brought up in [3]. For a vector of dependent variables $\mathbf{y} \in \mathbb{R}^P$ and the vectors of independent variables $\mathbf{x} \in \mathbb{R}^Q$ and $\mathbf{z} \in \mathbb{R}^R$, a RRR model is written as follows:

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\mu} + \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{z} + \boldsymbol{\epsilon} \\ &= \boldsymbol{\mu} + \mathbf{A}\mathbf{B}^T\mathbf{x} + \mathbf{D}\mathbf{z} + \boldsymbol{\epsilon},\end{aligned} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^P$ is the constant intercept, $\mathbf{C} \in \mathbb{R}^{P\times Q}$ is the low-rank coefficient matrix for $\mathbf{x}$ with $\operatorname{rank}(\mathbf{C}) = r \leq \min\{P, Q\}$, $\mathbf{D} \in \mathbb{R}^{P\times R}$ is the coefficient matrix for $\mathbf{z}$, and $\boldsymbol{\epsilon}$ is the innovation with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. Since $\mathbf{C}$ is low-rank, we have $\mathbf{C} = \mathbf{A}\mathbf{B}^T$ with $\mathbf{A} \in \mathbb{R}^{P\times r}$ and $\mathbf{B} \in \mathbb{R}^{Q\times r}$, which offers effective dimension reduction and improves the model interpretability. Matrix $\mathbf{A}$ is commonly named the exposure matrix and $\mathbf{B}$ is called the factor matrix with the linear combinations $\mathbf{B}^T\mathbf{x}$ being the latent factors. When RRR is used in autoregressive time series modeling, it is also referred to as the vector error correction model (VECM) [4]. RRR is widely used in many fields related to data analytics like wireless systems [5]–[8], financial econometrics [9]–[13], computer vision [14], environmental engineering [15], etc.

The classical methods for RRR/VECM estimation are the ordinary least squares estimation (LSE) [3], [4] and Gaussian maximum likelihood estimation (MLE) [16], where simple closed-form solutions can be attained. In many applications, the data to analyze often exhibit features of heavy-tails or outliers [17]. Such features contradict the data distribution assumption typically made for the theoretical analysis and estimation procedures in ordinary LSE and Gaussian MLE, hence leading to serious consequences in the estimated models [18], [19]. For example, in finance the common market behaviour and the proper portfolio design may be easily masked or misrepresented by the outlier data (e.g., bankruptcy of big corporations or financial crisis). In [20], a robust RRR (RRRR) estimation procedure against outliers was put forward based on nonconvex loss functions and mean-shift modelling. In [21], the authors proposed to estimate a robust VECM via an MLE procedure based on the Cauchy distribution, which is a conservative representative of the heavy-tailed distributions to better fit the heavy-tails and dampen the influence of outliers. In this paper, inspired by [21], we will tackle the heavy-tailedness via robust estimation based on Cauchy MLE.

In the literature of model estimation, a path of samples are always assumed available and deterministic batch estimation will be employed. When processing large-scale data sets or data streams, however, the deterministic estimation scheme becomes impractical owing to the requirement that the whole data set would be available at each iteration of the algorithm (the data collection process commonly spans over a long period) and dealing with a large-scale data set will make the algorithm computationally expensive. In view of this, a natural research question to ask is: *Can we create a continuously updated scheme for the model parameter estimation?* There has been a strong interest for online estimation procedure, which makes it possible to estimate the parameters of a data model without storing the whole data set. An online method was proposed in [22] using ordinary LSE for VECM but it lacks robustness. To deal with online estimation for the RRRR problem, the stochastic estimation techniques [23], [24] targeting the objective function with an expectation over a random variable will be adopted in this paper.

The classical approach to solve the stochastic optimization problem is the sample average approximation (SAA) method [25]–[28], which has played an important role in machine learning and signal processing. However, this approach could be computationally costly since it requires an iterative procedure at each iteration especially when the SAA subproblem is nonconvex. To overcome this problem, the stochastic majorization minimization (SMM) (a.k.a. stochastic upper-bound minimization) method [29] was proposed to solve the subproblem in each iteration by minimizing a well-chosen surrogate function and convergence to stationary points can be guaranteed. In [30], a penalized least squares estimation problem was studied based on the SMM method. In this paper, to deal with the nonconvex RRRR estimation and to tackle the large-scale data sets or data streams we design an algorithm based on the SMM method.[1] The efficiency of the proposed algorithm is demonstrated numerically by comparing to the state-of-the-art methods.

II. Problem Formulation

A. Robustness Pursuit by Cauchy Log-likelihood Loss

To take into account the heavy-tailed property and to mitigate outliers from the underlying data generating process, we adopt the multivariate Cauchy distribution for robustness pursuit [21]. Assume the innovation $\boldsymbol{\epsilon} \in \mathbb{R}^P$ in (1) follows a multivariate Cauchy distribution, i.e., $\boldsymbol{\epsilon} \sim \mathcal{C}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \in \mathbb{S}^P_{++}$, then its probability density function (PDF) is given as follows:

$$f(\boldsymbol{\epsilon}) = \frac{\Gamma[(1+P)/2]}{\Gamma(1/2)\pi^{P/2}\det(\boldsymbol{\Sigma})^{1/2}} \left(1 + \boldsymbol{\epsilon}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\epsilon}\right)^{-\frac{1+P}{2}}, \tag{2}$$

where $\Gamma(\cdot)$ denotes the gamma function and $\det(\cdot)$ is the determinant.

Considering the RRR model (1), the negative log-likelihood loss function for one sample $\boldsymbol{\xi} \triangleq \{\mathbf{y}, \mathbf{x}, \mathbf{z}\}$ is given by

$$\begin{aligned} &\ell(\boldsymbol{\theta}, \boldsymbol{\xi}) \\ &\triangleq \tfrac{1}{2} \log\det(\boldsymbol{\Sigma}) + \tfrac{1+P}{2} \log\big[1 + (\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\mathbf{B}^T\mathbf{x} \\ &\quad -\mathbf{D}\mathbf{z})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\mathbf{B}^T\mathbf{x} - \mathbf{D}\mathbf{z})\big], \end{aligned} \tag{3}$$

where $\boldsymbol{\theta} \triangleq \{\boldsymbol{\mu}, \mathbf{A}, \mathbf{B}, \mathbf{D}, \boldsymbol{\Sigma}\}$ is the parameter set to be estimated[2], $\boldsymbol{\theta} \in \boldsymbol{\Theta} \triangleq \{\boldsymbol{\Sigma} \succeq \mathbf{0}\}$, and the constant factors from the Cauchy PDF were removed.

B. Problem Formulation for Online RRRR

Based on the specified log-likelihood loss above, the online RRRR problem is readily given as follows:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\mathsf{minimize}} & \left[L(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\xi}}[\ell(\boldsymbol{\theta}, \boldsymbol{\xi})]\right] \\ \mathsf{subject\ to} & \boldsymbol{\theta} \in \boldsymbol{\Theta}, \end{array} \tag{4}$$

which is a constrained nonconvex stochastic optimization problem. To efficiently solve this problem, we will resort to an iterative numerical optimization method called stochastic majorization minimization [29] to be discussed in the next section. It is also worth mentioning that $\ell(\boldsymbol{\theta}, \boldsymbol{\xi})$ can be generalized to other robust loss functions like the Huber loss [31] and the algorithm described later still apply, but due to space limitation details will not be provided in this paper.

III. Solving The Online RRRR Problem via SMM

A. The Stochastic Majorization Minimization (SMM) Method

In stochastic optimization, people are not interested in the minimization of an empirical loss on a finite data set, but instead in minimizing an expected loss [24]. A general stochastic optimization problem of function $f(x)$ is given by

$$\begin{array}{ll} \underset{x}{\mathsf{minimize}} & \left[f(x) \triangleq \mathbb{E}_{\xi}[g(x, \xi)]\right] \\ \mathsf{subject\ to} & x \in \mathcal{X}, \end{array}$$

where $\mathcal{X}$ is a bounded closed convex set; ξ is a random vector drawn from a set $\varXi$, and $g(x)$ is a real-valued function in x. A classical approach for solving the above optimization problem is the SAA method [25], [26]. At each iteration of the SAA method, a new realization of the random vector ξ is obtained and the optimization variable x is updated by solving

$$x^{(k)} \leftarrow \arg\min\nolimits_{x \in \mathcal{X}} \tfrac{1}{N^{(k)}} \textstyle\sum_{i=1}^{N^{(k)}} g(x, \xi_i),$$

where $N^{(k)}$ is the considered sample size in the kth iteration. The SAA method is essentially an online optimization scheme since the sampling data is continuously incorporated into the problem and the optimization variables are hence updated in each iteration. However, the above SAA subproblem can be computationally expensive, say, when $g(x, \xi)$ is highly nonconvex. The SMM method [29] overcomes this difficulty through replacing $g(x, \xi)$ by a well-chosen majorizing surrogate function $\bar{g}(x, x^{(k)}, \xi)$ at $x^{(k)}$ which can be much easier to optimize. Specifically, the update step is summarized as

$$x^{(k)} \leftarrow \arg\min\nolimits_{x \in \mathcal{X}} \tfrac{1}{N^{(k)}} \textstyle\sum_{i=1}^{N^{(k)}} \bar{g}(x, x^{(k)}, \xi_i),$$

where the surrogate function $\bar{g}(x, x^{(k)}, \xi^i)$ satisfies

$$\begin{aligned} \bar{g}(x^{(k)}, x^{(k)}, \xi) = g(x^{(k)}, \xi), &\quad \forall x^{(k)} \in \mathcal{X},\ \forall \xi \in \varXi, \\ \bar{g}(x, x^{(k)}, \xi) \geq g(x, \xi), &\quad \forall x, x^{(k)} \in \mathcal{X},\ \forall \xi \in \varXi. \end{aligned}$$

To ensure convergence, the surrogate function $\bar{g}(x, x^{(k)}, \xi)$ is commonly required to be chosen such that the global optimal solution can be attained by solving the SMM subproblem, for instance, $\bar{g}(x, x^{(k)}, \xi)$ should be strongly convex in x.

B. Solving the Online RRRR Problem via SMM

In this section, we will solve the problem in (4) based on the SMM method. It is easy to see that the key of using SMM is to find a good majorizing function $\bar{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\xi})$ for the loss $\ell(\boldsymbol{\theta}, \boldsymbol{\xi})$ in each iteration, which will be detailed in the following. The loss function $\ell(\boldsymbol{\theta}, \boldsymbol{\xi})$ in (3) is nonconvex in $\boldsymbol{\theta}$. To find a surrogate function for it via the majorization minimization technique, we first introduce the following result.

Lemma 1 (Linear Majorization [21]): For a given point $x_0 \in \mathbb{R}$, the function $\log(1 + x)$ can be linearly majorized in the following way

$$\log(1+x) \leq \tfrac{1}{1+x_0} x + \log(1 + x_0) - \tfrac{x_0}{1+x_0},$$

where the equality is attained if and only if $x = x_0$.

Based on Lemma 1, at iterate $\boldsymbol{\theta}^{(k)}$ the second term in $\ell(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$ (given the sample $\boldsymbol{\xi}_i = \{\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\}$) can be majorized as in (5). Then combining the first term in $\ell(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$ with (5), we have the majorizing function as follows:

$$\begin{aligned} &\bar{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\xi}_i) \triangleq \\ &\tfrac{1}{2} \log\det(\boldsymbol{\Sigma}) + \tfrac{1+P}{2} \big(\bar{\mathbf{y}}_i^{(k)} - \sqrt{w_i^{(k)}}\boldsymbol{\mu} - \mathbf{A}\mathbf{B}^T\bar{\mathbf{x}}_i^{(k)} - \mathbf{D}\bar{\mathbf{z}}_i^{(k)}\big)^T \\ &\quad \times \boldsymbol{\Sigma}^{-1} \big(\bar{\mathbf{y}}_i^{(k)} - \sqrt{w_i^{(k)}}\boldsymbol{\mu} - \mathbf{A}\mathbf{B}^T\bar{\mathbf{x}}_i^{(k)} - \mathbf{D}\bar{\mathbf{z}}_i^{(k)}\big) + const., \end{aligned}$$

where we have defined $\bar{\mathbf{y}}_i^{(k)} \triangleq \sqrt{w_i^{(k)}}\mathbf{y}_i$, $\bar{\mathbf{x}}_i^{(k)} \triangleq \sqrt{w_i^{(k)}}\mathbf{x}_i$, and $\bar{\mathbf{z}}_i^{(k)} \triangleq \sqrt{w_i^{(k)}}\mathbf{z}_i$. Considering all the $N^{(k)}$ available samples in the kth iteration, the objective of the SMM subproblem is accordingly given as in (6).

[1] It should be noted that the proposed algorithm is also applicable to the non-robust regression case, where [22] is a special case.

[2] The rank of $\mathbf{A}$ can be determined based on prior knowledge or statistical analysis, however discussion on this topic is beyond the scope of this paper.

$$\begin{aligned} &\tfrac{1+P}{2}\log\Big[1+(\mathbf{y}_i-\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\mathbf{x}_i-\mathbf{D}\mathbf{z}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\mathbf{x}_i-\mathbf{D}\mathbf{z}_i)\Big]\\ &\leq \tfrac{1+P}{2}w_i^{(k)}(\mathbf{y}_i-\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\mathbf{x}_i-\mathbf{D}\mathbf{z}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\mathbf{x}_i-\mathbf{D}\mathbf{z}_i)+\mathit{const.}, \end{aligned} \tag{5}$$

where $w_i^{(k)} \triangleq \Big[1+(\mathbf{y}_i-\boldsymbol{\mu}^{(k)}-\mathbf{A}^{(k)}\mathbf{B}^{T(k)}\mathbf{x}_i-\mathbf{D}^{(k)}\mathbf{z}_i)^T\boldsymbol{\Sigma}^{-(k)}(\mathbf{y}_i-\boldsymbol{\mu}^{(k)}-\mathbf{A}^{(k)}\mathbf{B}^{T(k)}\mathbf{x}_i-\mathbf{D}^{(k)}\mathbf{z}_i)\Big]^{-1}$.

$$\begin{aligned} &\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}) \triangleq \tfrac{1}{N^{(k)}}\textstyle\sum_{i=1}^{N^{(k)}}\bar{\ell}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)},\boldsymbol{\xi}_i)\\ &=\tfrac{1}{2}\log\det(\boldsymbol{\Sigma})+\tfrac{1+P}{2}\textstyle\sum_{i=1}^{N^{(k)}}\big(\bar{\mathbf{y}}_i^{(k)}-\sqrt{w_i^{(k)}}\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{x}}_i^{(k)}-\mathbf{D}\bar{\mathbf{z}}_i^{(k)}\big)^T\boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{y}}_i^{(k)}-\sqrt{w_i^{(k)}}\boldsymbol{\mu}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{x}}_i^{(k)}-\mathbf{D}\bar{\mathbf{z}}_i^{(k)}\big)+\mathit{const.}\\ &=\tfrac{1}{2}\log\det(\boldsymbol{\Sigma})+\tfrac{1+P}{2}\mathrm{tr}\Big[\big(\bar{\mathbf{Y}}^{(k)}-\boldsymbol{\mu}\sqrt{\mathbf{w}^{(k)}}^T-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}-\mathbf{D}\bar{\mathbf{Z}}^{(k)}\big)^T\boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{Y}}^{(k)}-\boldsymbol{\mu}\sqrt{\mathbf{w}^{(k)}}^T-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}-\mathbf{D}\bar{\mathbf{Z}}^{(k)}\big)\Big]+\mathit{const.}, \end{aligned} \tag{6}$$

where $\mathbf{w}^{(k)} \triangleq [w_1^{(k)},\ldots,w_{N^{(k)}}^{(k)}]^T \in \mathbb{R}^{N^{(k)}}$, $\sqrt{(\cdot)}$ is the squared-root operator and is applied elementwise, $\bar{\mathbf{Y}} \triangleq [\bar{\mathbf{y}}_1,\ldots,\bar{\mathbf{y}}_{N^{(k)}}] \in \mathbb{R}^{P\times N^{(k)}}$, $\bar{\mathbf{X}} \triangleq [\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_{N^{(k)}}] \in \mathbb{R}^{Q\times N^{(k)}}$, and $\bar{\mathbf{Z}} \triangleq [\bar{\mathbf{z}}_1,\ldots,\bar{\mathbf{z}}_{N^{(k)}}] \in \mathbb{R}^{R\times N^{(k)}}$.

Finally, with the objective function $\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)})$, we get the SMM subproblem to be solved in each iteration as follows:

$$\begin{array}{ll} \underset{\boldsymbol{\theta}}{\mathsf{minimize}} & \bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}) \\ \mathsf{subject\ to} & \boldsymbol{\Sigma}\succeq\mathbf{0}, \end{array} \tag{7}$$

which is still highly nonconvex in $\boldsymbol{\theta}$; however, by carefully examining the problem structure, the global optimal solution can be attained in closed form and we give the details in the next section.

C. Solving the Subproblem in SMM

We first examine the first-order optimality conditions for variables $[\boldsymbol{\mu},\mathbf{D}]$ and $\boldsymbol{\Sigma}$. The partial derivative with respect to $[\boldsymbol{\mu},\mathbf{D}]$ is given as follows:

$$\begin{aligned} \nabla_{[\boldsymbol{\mu},\mathbf{D}]}\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}) = -(1+P)\boldsymbol{\Sigma}^{-1}\big(&\bar{\mathbf{Y}}^{(k)}-\boldsymbol{\mu}\sqrt{\mathbf{w}^{(k)}}^T\\ &-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}-\mathbf{D}\bar{\mathbf{Z}}^{(k)}\big)\big[\sqrt{\mathbf{w}^{(k)}},\ \bar{\mathbf{Z}}^{(k)T}\big]. \end{aligned}$$

For notational simplicity, we will denote $\mathbf{Q}^{(k)} \triangleq \big[\sqrt{\mathbf{w}^{(k)}},\ \bar{\mathbf{Z}}^{(k)T}\big]^T$ hereafter. By setting the above equation to be zero, for fixed $\mathbf{A}$ and $\mathbf{B}$ the optimal value for $[\boldsymbol{\mu},\mathbf{D}]$ is

$$\big[\boldsymbol{\mu},\mathbf{D}\big](\mathbf{A},\mathbf{B}) = \big(\bar{\mathbf{Y}}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\big)\mathbf{Q}^{(k)}\big(\mathbf{Q}^{(k)}\mathbf{Q}^{(k)T}\big)^{-1}. \tag{8}$$

Then we can have the following relation

$$\begin{aligned} &\bar{\mathbf{Y}}^{(k)}-\boldsymbol{\mu}\sqrt{\mathbf{w}^{(k)}}^T-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}-\mathbf{D}\bar{\mathbf{Z}}^{(k)}\\ &=\bar{\mathbf{Y}}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\\ &\quad-\big(\bar{\mathbf{Y}}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\big)\mathbf{Q}^{(k)T}\big(\mathbf{Q}^{(k)}\mathbf{Q}^{(k)T}\big)^{-1}\mathbf{Q}^{(k)}\\ &=\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}, \end{aligned} \tag{9}$$

where we have defined the projection matrix $\mathbf{P}^{(k)} \triangleq \mathbf{I}_N - \mathbf{Q}^{(k)T}\big(\mathbf{Q}^{(k)}\mathbf{Q}^{(k)T}\big)^{-1}\mathbf{Q}^{(k)}$ with $\mathbf{I}_N$ to be the identity matrix.

Considering the relation defined in (9), the partial derivative with respect to $\boldsymbol{\Sigma}$ is given as follows:

$$\begin{aligned} &\nabla_{\boldsymbol{\Sigma}}\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}) = \tfrac{N^{(k)}}{2}\boldsymbol{\Sigma}^{-1}-\tfrac{1+P}{2}\boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}\\ &-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)^T\boldsymbol{\Sigma}^{-1}, \end{aligned}$$

and then the optimal $\boldsymbol{\Sigma}$ is accordingly computed as

$$\begin{aligned} &\boldsymbol{\Sigma}(\boldsymbol{\mu},\mathbf{A},\mathbf{B},\mathbf{D})\\ &=\tfrac{1+P}{N^{(k)}}\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)\\ &\qquad\times\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)^T. \end{aligned} \tag{10}$$

Substituting (8) and (10) into (6), the loss function becomes

$$\begin{aligned} &\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)})\\ &=\tfrac{N^{(k)}}{2}\log\det\Big[\tfrac{1+P}{N^{(k)}}\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)\\ &\quad\times\big(\bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}-\mathbf{A}\mathbf{B}^T\bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}\big)^T\Big]+\mathit{const.} \end{aligned} \tag{11}$$

To simplify the notation, we further denote $\mathbf{N}^{(k)} \triangleq \bar{\mathbf{Y}}^{(k)}\mathbf{P}^{(k)}$ and $\mathbf{M}^{(k)} \triangleq \bar{\mathbf{X}}^{(k)}\mathbf{P}^{(k)}$ hereafter. Finally, the SMM subproblem becomes the minimization of $\bar{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)})$ in (11) with respect to $\mathbf{A}$ and $\mathbf{B}$. Since $\log(x)$ is a monotonically increasing function in x, the minimization problem is equivalent to

$$\underset{\mathbf{A},\,\mathbf{B}}{\mathsf{minimize}}\ \det\big[\big(\mathbf{N}^{(k)}-\mathbf{A}\mathbf{B}^T\mathbf{M}^{(k)}\big)\big(\mathbf{N}^{(k)}-\mathbf{A}\mathbf{B}^T\mathbf{M}^{(k)}\big)^T\big],$$

which is a matrix factorization problem with the determinant loss function. This problem is still nonconvex, but an analytical solution can be attained through singular value decomposition.

Proposition 1 ([16], [32]): Let $\mathbf{N}\in\mathbb{R}^{P\times N}$ be of rank P and $\mathbf{M}\in\mathbb{R}^{Q\times N}$ be of rank Q. Let $\mathbf{A}\in\mathbb{R}^{P\times r}$ and $\mathbf{B}\in\mathbb{R}^{Q\times r}$ be of rank r. Define $\mathbf{R}_{mm}\triangleq\mathbf{M}\mathbf{M}^T$, $\mathbf{R}_{mn}\triangleq\mathbf{M}\mathbf{N}^T$, $\mathbf{R}_{nm}\triangleq\mathbf{N}\mathbf{M}^T=\mathbf{R}_{mn}^T$, and $\mathbf{R}_{nn}\triangleq\mathbf{N}\mathbf{N}^T$. Then the minimum of $\det\big[\big(\mathbf{N}-\mathbf{A}\mathbf{B}^T\mathbf{M}\big)\big(\mathbf{N}-\mathbf{A}\mathbf{B}^T\mathbf{M}\big)^T\big]$ with respect to $\mathbf{A}$ and $\mathbf{B}$ is obtained for

$$\mathbf{A}^\star=\mathbf{R}_{nm}\mathbf{R}_{mm}^{-\frac{1}{2}}\mathbf{U}_r \quad\text{and}\quad \mathbf{B}^\star=\mathbf{R}_{mm}^{-\frac{1}{2}}\mathbf{U}_r, \tag{12}$$

where $\mathbf{U}_r\in\mathbb{R}^{Q\times r}$ contains the left singular vectors corresponding to the r largest singular values of matrix $\mathbf{R}_{mm}^{-\frac{1}{2}}\mathbf{R}_{mn}\mathbf{R}_{nn}^{-\frac{1}{2}}$ sorted in nonincreasing order. (The $\mathbf{R}_{mm}^{-\frac{1}{2}}$ is some matrix satisfying $\mathbf{R}_{mm}^{-\frac{1}{2}}\mathbf{R}_{mm}\mathbf{R}_{mm}^{-\frac{1}{2}T}=\mathbf{I}_Q$ and the same applies to $\mathbf{R}_{nn}^{-\frac{1}{2}}$.)

Finally, based on Proposition 1 we get the optimal solutions $\mathbf{A}^{(k+1)}$ and $\mathbf{B}^{(k+1)}$ and furthermore the updates for $\big[\boldsymbol{\mu}^{(k+1)},\mathbf{D}^{(k+1)}\big]$ and $\boldsymbol{\Sigma}^{(k+1)}$ can be obtained from (8) and (10), respectively.

D. The Overall SMM-based Algorithm

To solve the original online RRRR problem in (4), based on SMM it suffices to solve the subproblem in (7) iteratively with a closed-form solution update in each iteration.

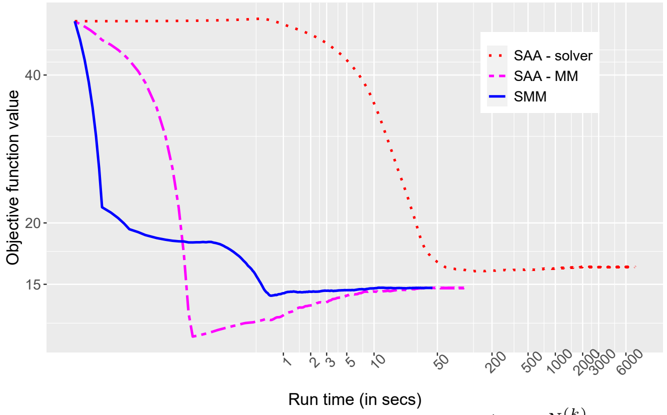

Fig. 1. Convergence comparisons for objective value $\frac{1}{N^{(k)}}\sum_{i=1}^{N^{(k)}} \ell(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$.

Algorithm 1 summarizes the whole procedure.[3]

Algorithm 1: Online RRRR via SMM

```
**Input:** Training data $\{\boldsymbol{\xi}_i\}_{i=1}^{\infty}$, the initial parameter $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$, and $k = 1$;
**for** $i = 1, \dots$ **do**
    Calculate $\{\mathbf{w}^{(k)}, \bar{\mathbf{Y}}^{(k)}, \bar{\mathbf{X}}^{(k)}, \bar{\mathbf{Z}}^{(k)}, \mathbf{Q}^{(k)}, \mathbf{P}^{(k)}\}$ based on the parameter $\boldsymbol{\theta}^{(k)}$ and data $\{\boldsymbol{\xi}_i\}_{i=1}^{N^{(k)}}$;
    Calculate $\{\mathbf{M}^{(k)}, \mathbf{N}^{(k)}, \mathbf{R}_{mm}^{(k)}, \mathbf{R}_{mn}^{(k)}, \mathbf{R}_{nn}^{(k)}, \}$;
    Compute $r$ left singular vectors of $\mathbf{R}_{mm}^{-\frac{1}{2}}\mathbf{R}_{mn}\mathbf{R}_{nn}^{-\frac{1}{2}}$;
    Update $\boldsymbol{\theta}^{(k+1)}$;
    $k \leftarrow k + 1$;
**Output:** $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\mu}^{(k)}, \mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{D}^{(k)}, \boldsymbol{\Sigma}^{(k)}\}$.
```

Notice that although in this online estimation algorithm the update of the parameters depends on all the past realizations, all the required information can be encoded into several matrices, which can be updated recursively.

IV. Numerical Simulations

In this section, we numerically evaluate the performance of our proposed model and algorithm. The simulation is conducted on a server with Intel(R) Xeon(R) CPU E5-2643 v4 (6x 3.40 GHz) and 128 GB RAM. A RRR model is specified with $P = Q = 10$ and $r = R = 1$. A path of 1000 samples is generated where innovations follow a Student's t-distribution with degree of freedom of 3 mimicking the real data scenarios. In the online estimation, we start with 25 samples and 1 sample is added in each iteration.

We first compare our proposed SMM-based algorithm with the benchmark SAA-based algorithm for solving the RRRR problem (4). Since the SAA subproblem is nonconvex, we solve it using both the embedded general-purpose solver in R "optim" [33] and a problem-tailored deterministic majorization minimization (MM) algorithm with 10 subiterations. The convergence comparisons on the average objection function value of 30 Monte Carlo runs is shown in Fig. 1. We can see that SMM-based algorithm converges faster than the SAA-based algorithms ("SAA - solver" and "SAA - MM"), and this can be explained by the double-loop nature of the SAA-based algorithms. In particular, "SAA - MM" tends to overfit at the beginning and takes more time to converge to the same level as "SMM".

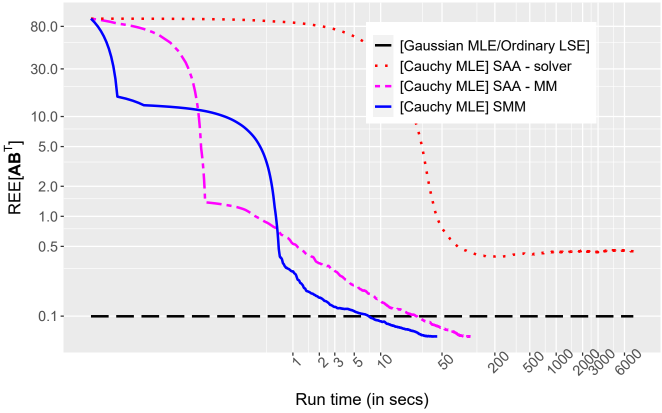

Fig. 2. Convergence comparisons for the $\mathsf{REE}[\mathbf{AB}^T]$.

TABLE I
Comparisons on average run time (in secs)

$(P,\ Q)$	$(5,\ 5)$	$(10,\ 10)$	$(20,\ 20)$	$(30,\ 30)$
SAA - MM	48.0 (16.8)	69.9 (17.7)	153.2 (27.6)	264.9 (27.2)
SMM	17.2 (6.85)	25.4 (7.49)	54.0 (11.04)	88.5 (11.06)

We further compare the estimation error between the true parameters and the estimated ones measured by the relative estimation error (REE) (for parameter $\mathbf{AB}^T$) defined as

$$\mathsf{REE}[\mathbf{AB}^T] \triangleq \frac{||\mathbf{A}^{(k)}\mathbf{B}^{(k)T} - [\mathbf{AB}^T]_{\mathsf{TRUE}}||_F^2}{||[\mathbf{AB}^T]_{\mathsf{TRUE}}||_F^2}.$$

We show the average REE of 30 Monte Carlo runs for $\mathbf{AB}^T$ in Fig. 2. The estimation result form non-robust Gaussian MLE, which is the same as the ordinary LSE, is also reported. It can be seen that the Cauchy assumption can attain a better estimation result in comparison to the non-robust estimation procedures, although the deficiency of "SAA - solver" prevents it from converging to the same level as other algorithms. A crossover of the convergence curves of "SAA - MM" and "SMM" can be found at the beginning, indicating a larger improvement over REE by "SAA - MM" than "SMM" in the first several iterations. However, the "SMM" algorithm eventually converges faster than both "SAA - MM" and "SAA - solver".

To show the computational efficiency of the proposed SMM-based algorithm, we compare the estimation time with varying the parameter dimensions where $P = Q$ and $r = R = 1$ based on 100 Monte Carlo simulations. In Table I, the average run time measured in seconds is presented with the standard error presented in parentheses. For all specified problem cases, the SMM algorithm consistently runs faster and more stable than the SAA method and scales well with the dimension.

V. Conclusions

In this paper, we have discussed the online robust reduced-rank regression problem. An efficient algorithm based on the stochastic majorization minimization method has been proposed. The effectiveness of the model and algorithm has been demonstrated through simulation simulations.

[3] To promote reproducible research, an implementation for this algorithm in R [33] is publicly available in the package RRRR [34].

References

[1] T. Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 327–351, 1951.

[2] T. Anderson, *An introduction to multivariate statistical analysis (Wiley series in probability and statistics)*. Wiley, 2003.

[3] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[4] R. F. Engle and C. W. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Applied Econometrics*, vol. 39, no. 3, pp. 107–135, 2015.

[5] Y. Hua, M. Nikpour, and P. Stoica, "Optimal reduced-rank estimation and filtering," *IEEE Transactions on Signal Processing*, vol. 49, no. 3, pp. 457–469, 2001.

[6] T. Gustafsson and B. D. Rao, "Statistical analysis of subspace-based estimation of reduced-rank linear regressions," *IEEE Transactions on Signal Processing*, vol. 50, no. 1, pp. 151–159, 2002.

[7] M. Nicoli and U. Spagnolini, "Reduced-rank channel estimation for time-slotted mobile communication systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 926–944, 2005.

[8] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.

[9] Z. Zhao and D. P. Palomar, "Sparse reduced rank regression with nonconvex regularization," in *2018 IEEE Statistical Signal Processing Workshop, SSP 2018*, (Freiburg), pp. 618–622, IEEE, jun 2018.

[10] S. Johansen and K. Juselius, "Some structural hypotheses in a multivariate cointegration analysis of the purchasing power parity and the uncovered interest parity for UK," *Journal of Econometrics*, vol. 53, no. 1-3, pp. 211–244, 1992.

[11] E. Bernardini and G. Cubadda, "Macroeconomic forecasting and structural analysis through regularized reduced-rank regression," *International Journal of Forecasting*, vol. 31, no. 3, pp. 682–691, 2015.

[12] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2342–2357, 2018.

[13] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Transactions on Signal Processing*, vol. 67, no. 7, pp. 1681–1695, 2019.

[14] D. Huang and F. De la Torre, "Bilinear kernel reduced rank regression for facial expression synthesis," in *European Conference on Computer Vision*, pp. 364–377, Springer, 2010.

[15] C. A. Glasbey, "A reduced rank regression model for local variation in solar radiation," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 41, no. 2, p. 381, 1992.

[16] S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models," *Econometrica*, vol. 59, p. 1551, nov 1991.

[17] G. George, M. R. Haas, and A. Pentland, "Big data and management," 2014.

[18] J. Law, F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, "Robust statistics: The approach based on influence functions.," *The Statistician*, vol. 35, no. 5, p. 565, 1986.

[19] R. D. Martin and V. J. Yohai, "Influence functionals for time series," *The Annals of Statistics*, vol. 14, no. 3, pp. 781–818, 1986.

[20] Y. She and K. Chen, "Robust reduced-rank regression," *Biometrika*, vol. 104, no. 3, pp. 633–647, 2017.

[21] Z. Zhao and D. P. Palomar, "Robust maximum likelihood estimation of sparse vector error correction model," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 913—-917, IEEE, 2017.

[22] P. Arce, J. Antognini, W. Kristjanpoller, and L. Salinas, "An online vector error correction model for exchange rates forecasting," in *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, vol. 2, pp. 193–200, 2015.

[23] L. Bottou, "Online learning and stochastic approximations," *On-line learning in neural networks*, vol. 17, no. 9, p. 142, 1998.

[24] L. A. Hannah, "Stochastic optimization," *International Encyclopedia of the Social \& Behavioral Sciences*, vol. 2, pp. 437–481, 2015.

[25] E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri, "Sample-path optimization of convex stochastic performance functions," *Mathematical Programming, Series B*, vol. 75, no. 2, pp. 137–176, 1996.

[26] K. Healy and L. W. Schruben, "Retrospective simulation response optimization," in *Winter Simulation Conference Proceedings*, vol. 1, (Los Alamitos), pp. 901–906, IEEE Computer Society, 1991.

[27] R. Y. Rubinstein, "Optimization of computer simulation models with rare events," *European Journal of Operational Research*, vol. 99, no. 1, pp. 89–112, 1997.

[28] A. Shapiro, "Monte carlo sampling methods," *Handbooks in Operations Research and Management Science*, vol. 10, no. C, pp. 353–425, 2003.

[29] M. Razaviyayn, M. Sanjabi, and Z. Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Mathematical Programming*, vol. 157, pp. 515–545, jun 2016.

[30] E. Chouzenoux and J. C. Pesquet, "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation," in *IEEE Transactions on Signal Processing*, vol. 65, pp. 4770–4783, dec 2017.

[31] P. J. Huber, *Robust statistics*. Springer, 2011.

[32] H. Lütkepohl, *New introduction to multiple time series analysis*. New York, 2005.

[33] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[34] Y. F. Yang and Z. Zhao, *RRRR: Online robust reduced-rank regression estimation*, 2020. R package version 1.0.0. https://CRAN.R-project.org/package=RRRR.