Estimation of the multi-index conditional volatility model

Hongfan Zhang
School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, P. R. China.

Email: zhfbeyond@126.com

Abstract. We concentrate on estimating of the multi-index conditional volatility model. The outer products of gradients (OPG) method is proposed to handle the estimation. We implement the optimization by minimizing a penalized loss function, which leads to the maximum quasi-likelihood estimate. Moreover, an adaptive version of cross-validation is employed to estimate the structure dimension. Simulation studies assesses the sample performance of our proposed methods.

1. Introduction
Dimension reduction is one of research focuses in quantitative sciences. Since [1] developed the sliced inverse regression (SIR) method, there have been various dimension reduction approaches to various dimension reduction models. Statistically speaking, a general dimension reduction model searches a parsimonious relationship between the high dimensional covariate \( X = (X_1,\ldots, X_p)^T \in \mathbb{R}^p \) with \( p \geq 3 \) and the response of interest \( Y \in \mathbb{R}^1 \), by which the curse of dimensionality can be avoided through projecting \( X \) along with one or two directions. That is \( Y \perp X \mid B^T X \) with \( B \in \mathbb{R}^{pd} \) being orthogonal, and \( d = 1 \) or 2, where \( \perp \) denotes conditional independence. The columns of \( B \) then span the well-known central subspace (CS) which is denoted by \( S_{MIX} \). The CS usually exists under some mild conditions [2]. The estimation methods mainly include [2, 3, 4, 5, 6], among which the first four methods are developed under the linearity condition of [1], which may be violated in many situations and hence in turn limit these methods.

The central mean subspace (CMS), a subspace of CS, summarizes all useful information of \( X \) about the conditional mean \( E(Y \mid X) \) through linear combinations \( B^T X \), that is to say, \( Y \perp E(Y \mid X) \mid B^T X \). The existence of CMS can also be guaranteed under some mild conditions. Various methods have been devoted to estimating CMS as [7, 8, 9]. When \( B \) is a vector with \( d = 1 \), the dimension reduction model \( Y = g(B^T X) + \varepsilon \), where \( \varepsilon \) is the error term, conditionally independent with \( X \), reduces to the single-index model; see [10, 11] for details. Particularly, the OPG and MAVE estimation methods proposed by [10] are free of the linear constraint and are very efficient. Another important subspace of CS is the central variance subspace (CVS) proposed by [12, 13, 14]. In this situation, dimension reduction aims to find an orthogonal matrix \( B \) which satisfies \( Y \perp \text{Var}(Y \mid X) \mid B^T X \). This is equivalent to assuming

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
\[ Y = g^{1/2}(B^T X) \times \varepsilon, \]

which is called the multi-index conditional volatility model with the function \( g \) being non-negative in this paper. Our methods will pay main attentions to estimating \( B \) and \( d \). In fact, for the general heteroscedastic regression model \( Y = m(A^T X) + v^{1/2}(B^T X)\varepsilon \), to estimate \( B \) we only need to replace the response by the fitted residuals provided that \( A \) and \( m \) can be estimated efficiently. For estimations of the universal dimensional regression model \( Y = m(X) + v^{1/2}(X)\varepsilon \) where \( X \in \mathbb{R}^1 \), one can refer to [15, 16, 17, 18]. Note that [12] proposes a square of residuals based OPG method to recover \( B \). Our estimation method although is also based on OP, is very different from [12]. In this paper, a more adaptive and a log penalized loss function is minimized, rather than the usual least squares loss function of [12]. Our method can be deemed as a generalization of [19] which only focuses on the corresponding single-index version. In [19], the quasi-likelihood OP and MAVE methods are developed to recover the single-index parameter vector in the conditional variance function. The MAVE based procedure in [19] relies on nonlinear optimization with respect to the single-index parameter, and our experience suggests MAVE may not be appropriate for estimating \( B \) in this situation because it is difficult to deal with matrix parameter nonlinear optimization. Thus, in this paper, we only consider the OPG based method, to see whether it could efficiently recover the multi-index parameters. In addition, we develop an adaptive delete-one cross-validation criterion to estimate the structure dimension \( d \).

The rest of the article is organized as the follows. Section 2 introduces the penalized OPG estimation method. Section 3 gives an adaptive delete-one cross-validation criterion to estimate the structure dimension. We evaluate the finite sample performance of our methods through simulated data sets in Section 4. Finally, some conclusions and relevant concerns are given in Section 5.

2. Penalized OPG method

In this section, we tentatively assume the structure dimension \( d \) is known. We first consider the gradient of the conditional variance (volatility) function \( g(B^T x) \) with respect to \( x \), which is the basis of Outer Products of Gradients. Suppose that \( \varphi \) is a known differentiable transformation on its domain of definition, a subspace of \( \mathbb{R}^1 \), then it can be easily seen that

\[
\frac{\partial \varphi(g(B^T x))}{\partial x} = B \times \nabla \varphi(g(B^T x)) \times \nabla g(B^T x),
\]

where

\[
\nabla \varphi(v) = \frac{\partial \varphi(v)}{\partial v}, \quad \text{and} \quad \nabla g(v_1, \ldots, v_d) = (\nabla_1 g(v_1, \ldots, v_d), \ldots, \nabla_d g(v_1, \ldots, v_d))^T
\]

with

\[
\nabla_k g(v_1, \ldots, v_d) = \frac{\partial}{\partial v_k} g(v_1, \ldots, v_d), \quad k = 1, 2, \ldots, d.
\]

Thus, by taking the expectation of the outer product of the gradients, we have

\[
E \left\{ \left( \frac{\partial \varphi(g(B^T X))}{\partial x} \right) \left( \frac{\partial \varphi(g(B^T X))}{\partial x} \right)^T \right\} = BE \left\{ [\nabla \varphi(g(B^T X))]^T \nabla g(B^T X) \nabla^T g(B^T X) \right\} B^T.
\]

Then the estimate of \( B \) can be available by calculating the first \( d \) eigenvectors of the sample estimate of the expectation of the outer products, see [6, 10] for detailed discussion. By defining \( G(x) = g(B^T x) \), it suffices to acquire the estimator of the gradient \( \frac{\partial \varphi(G(x))}{\partial x} \). Now suppose we have \( n \) i.i.d. observations \( \{(X_i, Y_i), i = 1, 2, \ldots, n\} \) from the population \( (X, Y) \). To estimate the gradient, we use local kernel smoothing techniques of [20] to approximate the function \( \varphi(G(X_i)) \) by
\( \varphi(G(X_i)) \approx a_x + b_x^T X_{ix} \), where \( X_{ix} = X_i - x \), \( a_x \) and \( b_x \) are approximations of \( \varphi(G(x)) \) and \( \hat{\varphi}(G(x)) \) respectively. The next task is to seek a proper loss function from which the estimate \( b_x \) can be derived.

The reference [12] uses the usual least squares loss function by noting that \( E(Y^2 \mid X) = g(B^T X) \). Separately, our method uses an entirely different, but a more adaptive penalized loss function. Specifically, we minimize \( L(a_x, b_x) \) with respect to \( (a_x, b_x^T) \) as follows,

\[
(\hat{a}_x, \hat{b}_x) = \arg\min_{a_x,b_x} n^{-1} \sum_i \left\{ \exp[-(a_x + b_x^T X_{ix})] Y_i^2 + a_x + b_x^T X_{ix} \right\} w_{ix} = L(a_x, b_x).
\]

Where we have used \( \varphi(v) = \ln(v) \), \( a_x \) and \( b_x \) are approximations of the functions \( \ln(G(x)) \) and \( G^{-1}(x) \times \ln(G(x)) \), respectively, and \( w_{ix} = K_h(X_{ix}) = h^{-p} K(h^{-1} X_{ix}) \) where \( K \) is a \( p \) dimensional kernel function. Note that \( L(a_x, b_x) \) is a nonlinear function, hence optimizing it involves nonlinear iterative algorithms, such as the Newton method and the Steepest Descent method. We next explain why we use \( L \) rather than the usual least squares loss function as [12]. In fact, \( L \) consists of two parts. If we further define the function \( \sigma^2(B^T X) = \text{Var}(\varepsilon \mid B^T X = B^T x) \), then from the fact that \( \sigma^2(B^T X) = E(G^{-1}(X) Y_i^2 \mid B^T X) \) and \( \exp(a_x + b_x^T X_{ix}) \) essentially approximates \( G(X_i) \), we can know the first part \( n^{-1} \sum_i \left\{ \exp[-(a_x + b_x^T X_{ix})] Y_i^2 \right\} w_{ix} \) estimates \( E\{\sigma^2(B^T X)\} \), which is more adaptive in this situation than the usual least squares criterion. However, if we just minimize the first part, it will give a trivial solution. The second part, parametrized by \( n^{-1} \sum_i \left\{ a_x + b_x^T X_{ix} \right\} w_{ix} \), penalizes the parameters we have used in the first part, not only to circumvent trivial solutions but also to make the final function estimator positive, namely, \( \hat{G}(x) = \exp(\hat{a}_x) > 0 \). In this sense, our loss function \( L \) is more adaptive to the target model.

The loss function \( L \) can also be derived from the assumption that \( \varepsilon \mid X \sim N(0,1) \). If it holds, by model (1), we have \( Y \mid X = x \sim N(0, G(x)) \), which leads to the sample log-likelihood function

\[
-\sum_i \left\{ G^{-1}(X_i) Y_i^2 + \ln G(X_i) \right\}.
\]

By approximating the function \( \ln G(X_i) \) by the local linear smoother, and maximizing the corresponding quasi-likelihood function, it is easy to see, the resulting maximizer is just the minimizer of (5). Hence our penalized OPG method can be deemed as a maximum quasi-likelihood estimate.

With \( \{ \hat{b}_j, j = 1, ..., n \} \) available, we consider a weighted average of the outer products,

\[
\hat{\Sigma} = n^{-1} \sum_{j=1}^n \hat{\rho}_j \hat{b}_j \hat{b}_j^T,
\]

where \( \hat{\rho}_j \) is a trimming function introduced to handle the boundary points. Then the space spanned by the first \( d \) eigenvectors of \( \hat{\Sigma} \), denoted by a matrix, say \( \hat{B} \), is the estimator of the space spanned by the columns of \( B \). Let \( \rho(\cdot) \) be a bounded function on \( \mathbb{R}^1 \) such that \( \rho(v) = v > 0 \) if \( v > \alpha_0 \), and
\( \rho(v) = 0 \) if \( v < a_b \) for some small \( a_b > 0 \), see [6, 10] for details. In this paper, we consider 
\[ \hat{\rho}_j = \rho(\hat{f}(X_j)) \]
where \( \hat{f}(X_j) = n^{-1} \sum_{i=1}^n K_h(X_{ij}) \), a kernel density estimator.

To allow the estimation to be more adaptive to the dimension reduction model (1), we may replace
\( w_{ij} = K_h(\hat{B}^T X_{ij}) \), where \( K \) now is a \( d \) dimensional kernel function, and \( \hat{B} \) is the latest estimator. Similarly, we can replace \( \hat{\rho}_j \) in (6) with 
\[ \hat{\rho}_j = \rho(\hat{f}(\hat{B}^T X_j)) \]
where \( \hat{f}(\hat{B}^T x) \) is a kernel estimator of the density function \( f(B^T x) \). To implement the whole estimation, we suggest the following OPG based algorithm as if the dimension \( d \) was known.

**Step(1).** Set initial matrix \( \hat{B} = I_p \).

**Step(2).** With \( \hat{w}_{ij} = K_h(\hat{B}^T X_{ij}) \), employ the Newton method and obtain the solution
\( (\hat{a}_j, \hat{b}_j) = \arg\min L(a_j, b_j) \).

**Step(3).** With \( \hat{\rho}_j = \rho(\hat{f}(\hat{B}^T X_j)) \), calculate the weighted average of outer products, \( \hat{\Sigma} \), as (6).

**Step(4).** Calculate the first \( d \) eigenvectors of \( \hat{\Sigma} \), and denote them by a new matrix \( \hat{B} \). Set \( \hat{B} = \hat{B} \) and repeat Steps (1) - (4) till convergence.

### 3. Determining the structure dimension

In practice, the dimension of \( B \) is unknown, thus there is a need to estimate \( d \). In this section, we propose an adaptive delete-one cross-validation criterion. The meaning of “adaptive” is that, instead of the commonly used conditional mean function \( E(Y \mid B^T X = B^T x) \), we use the estimate of the function \( E(Y^2 \mid B^T X = B^T x) \approx m(B^T x) \), because the former equals zero in our model (1).

With a working dimension \( q \) and its corresponding delete-one estimate \( \hat{B}_{q}^{(j)} \), we first calculate the delete-one fitted value of \( m(B^T x) \) for each observation. For each \( j = 1, 2, \ldots, n \), with the delete-one sample \( \{(X_i, Y_i), i \neq j\} \) and a newly introduced bandwidth \( h > 0 \), calculate the local constant estimator
\[ \hat{\gamma}_{i,j}^{(j)} = \frac{\sum_{i \neq j} K_h(X_{ij}^T \hat{B}_{q}^{(j)})Y_i^2}{\sum_{i \neq j} K_h(X_{ij}^T \hat{B}_{q}^{(j)})}. \]  

Then the CV value for model (1) is defined as
\[ CV(q) = n^{-1} \sum_{j=1}^n \left( \gamma_j - \hat{\gamma}_{i,j}^{(j)} \right) \hat{\rho}_j, \]

where \( \hat{\rho}_j = \rho(\hat{f}(\hat{B}_{q}^{(j)} X_j)) \). Finally, the dimension for model (1) is selected as
\[ \hat{d} = \arg\min_{1 \leq q \leq p} CV(q). \]  

### 4. Simulations

We now demonstrate the performance of our proposed method by simulations. The estimation error for the estimator \( \hat{B} \) is defined as the maximum singular value of \( \hat{B} \hat{B}^T - B^T B \); see [6] for details. We
use the Epanechnikov kernel and $q_b = 0.01$. In our simulation studies, we set $h = 1.06 n^{-1/(q+4)}$ and $h = 1.06 n^{-1/(q+4)}$ where $q$ is the working dimension. We consider the following model,

$$Y = \left\{ (b_i X)^2 e^{-2(b_i X)^2} \right\}^{1/2} \times \varepsilon,$$

(10)

where $\varepsilon$ is independent with $X$, and $\varepsilon \sim N(0,1)$. We set $b_1 = (1,2,2,0,0)^T / 3$ and $b_2 = (0,0,0,3,4)^T / 5$. Two designs of $X$ are considered: (A) each element $X_i, 1 \leq i \leq p$, is independently distributed from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$; (B) $X \sim N\left(0, (\sigma_y)^2 I_{si,j\leq p}\right)$, where $\sigma_y = 0.25^{l-j}$. Note that the structure dimension equals 2. The results for different sample sizes based on 200 replications are reported in table 1.

### Table 1. Estimation errors (deviations in bracket) and relative frequencies of correct selection of the structure dimension.

| Design | A       | Correct frequency | Estimation error | Correct frequency |
|--------|---------|-------------------|------------------|-------------------|
|        | Estimation error |                   |                  |                   |
| n=600  | 0.0740  (0.0048) | 99.50%            | 0.1622           (0.0014) | 96.00%            |
| n=800  | 0.0621  (0.0034) | 100.00%           | 0.1513           (0.0011) | 99.50%            |
| n=1000 | 0.0509  (0.0012) | 100.00%           | 0.1426           (0.0008) | 100.00%           |

As is well known, estimating the central variance subspace is much more difficult than estimating the central mean subspace, thus we here use relatively huge sample sizes to investigate the performance of our proposed methods. It can be concluded from table 1 that, the performance for design A is better than that for design B, no matter in terms of the estimation error or the frequency of correct selection of the structure dimension. As the sample size grows, the estimation error declines, and the correct frequency increases to 100%, which respectively verifies the consistency of the penalized OPG estimation method and the delete-one cross-validation criterion.

To give a better illustration, figure 1 plots a typical data set for design A with $n = 600$. The top two panels (a) and (b) plot the observed values $Y_i$ and the true volatility surface $g(B^T X_i)$ against the true projected directions $X_i^T b_1$ and $X_i^T b_2$, where $B = (b_1, b_2)$, with two different angles. In the middle left panel (c), we plot the observed values $Y_i$ and the true volatility values $g(B^T X_i)$ against $X_i^T b_1$. In the middle right panel (d), we plot the observed values $Y_i$ and the estimated volatility values $\hat{g}(X_i)$ against the estimated projected direction $X_i^T \hat{b}_1$, as if the structure dimension was known to be 2. The bottom two panels (e) and (f) depict the same plots as the middle panels, however with respect to the projected directions $X_i^T b_2$ and $X_i^T b_2$. We can see that our estimated directions are accurate, at least to the naked eye.
Figure 1. A typical data set from design A with \( n = 600 \) and its penalized OPG estimation. In the top panels (a) and (b), “o” denotes plots of the observed values \( Y \) against the true projected directions.
In the middle left panel (c), “o” denotes plots of \( Y \) against the true projected direction \( X^\top b_1 \), “.” denotes plots of the true volatility function \( g(B^\top X) \) against \( X^\top b_1 \). In the middle right panel (d), “o” denotes plots of \( Y \) against the estimated projected direction \( X^\top \hat{b}_1 \), “.” denotes plots of the estimated volatility function \( \hat{G}(X) \) against \( X^\top \hat{b}_1 \). The bottom panels (e) and (f) concern similar plots against to the other projected direction as the middle panels.

5. Conclusions and discussions

In this paper, we proposed the penalized outer products of gradients method to handle the estimation of the multi-index conditional volatility model, and an adaptive delete-one cross-validation criterion to estimate the structure dimension. Limited simulations verified the efficiency of the OPG based estimation method and the consistency of the cross-validation procedure.

There are some related concerns to be discussed. By requiring some conditions, the asymptotic normality of the penalized OPG estimator \( \hat{B} \) can be established; see [6, 12] for detailed discussion. Another concern is the consistency of our proposed cross-validation criterion. We believe in that there should be \( \hat{d} \rightarrow d \) in probability. Despite that we don’t give any rigorous theoretic arguments here, the theoretic research will be on-going.

References

[1] Li K C 1991 Journal of the American Statistical Association 86 316-327
[2] Cook R D 1998 Journal of the American Statistical Association 93 84-94
[3] Cook R D and Weisberg S 1991 Journal of the American Statistical Association 86 28-33
[4] Cook R D and Li B 2002 The Annals of Statistics 32 455-474
[5] Li B and Wang S L 2007 Journal of the American Statistical Association 102 997-1008
[6] Xia Y 2007 The Annals of Statistics 35 2654-2690
[7] Xia Y, Tong H, Li W K and Zhu L 2002 Journal of Royal Statistical Society, Ser. B 64 363-410
[8] Xia Y 2008 Journal of the American Statistical Association 103 1631-1640
[9] Ma Y and Zhu L 2012 Journal of the American Statistical Association 107 168-179
[10] Xia Y 2006 Econometric Theory 22 1112-1137
[11] Hardle W and Stoker T M 1989 Journal of the American Statistical Association 84 986-995
[12] Zhu L P and Zhu L X 2009 Statistica Sinica 19 869-883
[13] Yin X and Cook R D 2002 Journal of Royal Statistical Society, Ser. B 64 159-175
[14] Xia Y, Tong H and Li W K 2002 Statistica Sinica 12 785-799
[15] Fan J and Yao Q 1998 Biometrika 85 645-660
[16] Ruppert D, Wand M P, Holst U and Hossjer O 1997 Local polynomial variance function estimation Technometrics 39 1257-1270
[17] Yu K and Jones M C 2004 Likelihood-based local linear estimation of the conditional variance function Journal of the American Statistical Association 99 139-144
[18] Ziegelmann F 2002 Nonparametric estimation of volatility functions: the local exponential estimator Econometric Theory 18 985-991
[19] Zhang H F 2017 Quasi-likelihood estimation of the single-index conditional variance model Computational Statistics and Data Analysis (Submitted)
[20] Fan J and Gijbels I 1996 Local polynomial modelling and its applications (New York: Chapman & Hall) 3 57-105