A quasi-stochastic gradient algorithm for variance-dependent component analysis

Aapo Hyvärinen[1], Shohei Shimizu[12]

[1] Helsinki Institute for Information Technology, University of Helsinki, Finland
[2] The Institute of Statistical Mathematics, Japan
http://www.cs.helsinki.fi/hiit_bru/index_neuro.html

Abstract. We discuss the blind source separation problem where the sources are not independent but are dependent only through their variances. Some estimation methods have been proposed on this line. However, most of them require some additional assumptions: a parametric model for their dependencies or a temporal structure of the sources, for example. In previous work, we have proposed a generalized least squares approach using fourth-order moments to the blind source separation problem in the general case where those additional assumptions do not hold. In this article, we develop a simple optimization algorithm for the least squares approach, or a quasi-stochastic gradient algorithm. The new algorithm is able to estimate variance-dependent components even when the number of variables is large and the number of moments is computationally prohibitive.

1 Introduction

In blind source separation methods, the observed signals $x_i(t)$ $(i = 1 \cdots m)$ are typically assumed to be linear mixtures of sources $s_j(t)$ $(j = 1 \cdots n)$. Let $\bar{a}_{ij}$ denote the coefficients in the linear mixing between the sources $s_j(t)$ and the observed signals $x_i(t)$. Then the mixing can be expressed as

$$x_i(t) = \sum_{j=1}^{n} \bar{a}_{ij} s_j(t). \tag{1}$$

The problem of blind source separation is now to estimate both the source signals $s_i(t)$ and the mixing coefficient $\bar{a}_{ij}$ based on observations of the $x_i(t)$ alone [1].

The model (1) is called independent component analysis (ICA) model if $s_j(t)$ are assumed to be non-gaussian and independent [2]. The ICA model has been extensively studied for last two decades, and many estimation techniques for the model are available [3].

Recently, many extensions of the ICA model have started to be considered [4–6]. A quite interesting extension among them is the case where the source signals are not independent but dependent only through their variances [6]. To model such dependencies, [7] assumed that each source signal $s_i(t)$ can be represented as a product of two random signals $v_i(t)$ and $y_i(t)$:

$$x_i(t) = \sum_{j=1}^{n} \bar{a}_{ij} v_j(t) y_j(t), \tag{2}$$

where $v_i(t)$ and $y_i(t)$ are independent, $y_i(t)$ are independent over time and are mutually independent of each other. No assumption on the distribution of $y_i(t)$ is made other than $y_i(t)$ have zero means. The variance signals $v_i(t)$ are nonnegative signals giving general activity levels and are allowed to be statistically dependent. Thus, the $v_i(t)$ could produce dependencies between sources $s_i(t) = v_i(t)y_i(t)$. No particular assumptions on the dependencies between $v_i(t)$ are made. This setting was called double-blind source separation problem because one neither observes the source signals $s_i(t)$ nor postulates a parametric model of their dependencies.

In [7], it was further assumed that the source signals have some time dependencies (autocorrelations) and a method was proposed that uses the time structure of the observed signals for separating the source signals. The time dependency assumption is the key to the method, and the method is not applicable to the case where the source signals are not temporally structured and has a more limited domain of applications, since many kinds of data do not have temporal structure in practice.

In [8], estimating functions for the model (2) was studied, and the quasi maximum likelihood estimation that requires no time dependencies was proposed. However, one has to appropriately choose the nonlinearity depending on whether the underlying independent signals $y_i(t)$ are supergaussian or subgaussian as in maximum likelihood methods for the ordinary ICA model. Moreover, they have to make certain extra assumptions on the signs of certain complicated nonlinear cross-moments of the sources, and it is not very clear when these are fulfilled.

In previous work [9], we proposed a generalized least squares approach using second- and fourth-order moment structures of observed signals in the general case where no temporal structure is available and it is unknown whether the underlying signals are supergaussian or subgaussian. However, its optimization using the ordinary gradient descent method is more difficult for larger variables since the number of moments increases enormously. In this paper, we provide a computationally efficient algorithm, or a *quasi*-stochastic gradient algorithm.

2 Model

We shall define the following model, which we will refer to as variance-dependent component analysis (VDCA) here. Let us collect the source signals in a vector $\boldsymbol{s} = [s_1, \cdots, s_n]^T$, and also construct the observed signal vector $\boldsymbol{x}$ in the same manner. (We omit the time indices in the subsequent part since we do not consider time structures.) Let us further collect the mixing coefficients in a matrix

$\bar{\mathbf{A}} = [\bar{a}_{ij}]$. The VDCA model for the m-dimensional observed vector $\boldsymbol{x}$ is written as

$$\boldsymbol{x} = \bar{\mathbf{A}}\boldsymbol{s}, \tag{3}$$

where non-gaussian components s_i can be expressed as products of two signals v_i and y_i, $s_i = v_i y_i$, as in (2), where the y_i are zero-mean and mutually independent, and that the set of the y_i is independent from the set of the v_j. No assumptions on the dependencies of the v_j with each other are made. An important point in the VDCA model is that no temporal structure is assumed, which is different from [7]. Here, we further assume $\bar{\mathbf{A}}$ to be square, which is a typical assumption in blind source separation [3].

An illustrative example

To illustrate the VDCA model, let us consider two stereotypical signals for which ordinary ICA does not work but VDCA does work. Let us define v_1, v_2, y_1 and y_2 as follows:

$$v_1 = 0.2 + \exp\{-4(t-7)^2\} + 0.5\exp\{-4(t-4)^2\} \tag{4}$$
$$v_2 = 0.2 + \exp\{-4(t-6.8)^2\} + 0.5\exp\{-4(t-4.2)^2\} \tag{5}$$
$$y_1 = \sin(50t) \tag{6}$$
$$y_2 = \cos(37t) \tag{7}$$
$$(t = 0, 0.01, 0.02, \cdots, 10).$$

Then we define variance-dependent signals $s_1 = v_1 y_1$, $s_2 = v_2 y_2$. Here, the underlying signals y_i are subgaussian and variance signals v_i are highly correlated. See Figure 1 for the original source signals s_i, estimated sources s_i by VDCA and FastICA with the hyperbolic tangent nonlinearity [10].

The point is that ICA tries to find a maximally non-gaussian linear combination of the source signals. Now it finds two conflicting goals: in the source signals, the sinusoids y_i inside the envelopes are subgaussian, hence the original signals maximize subgaussianity inside the envelopes. In contrast, modulation by v_i make the signals s_i supergaussian, and hence an ICA algorithm should maximize supergaussianity to maximize non-gaussianity. This conflict between sub- and super-gaussianity makes ICA fail.

3 A generalized least squares approach

In previous work [9], we have proposed the generalized least squares approach (GLS) in estimation that utilizes higher-order moment to estimate $\bar{\mathbf{A}}$ in (3).

Let us denote by $\boldsymbol{\sigma}_2(\boldsymbol{\tau})$ the vector that consists of elements of the covariance matrix based on the model where any duplicates due to symmetry have been removed and by $\boldsymbol{\sigma}_4(\boldsymbol{\tau})$ the vector that consists of the tensor of fourth-order (cross-) moments where duplicate entries have been removed and by $\boldsymbol{\tau}$ the vector

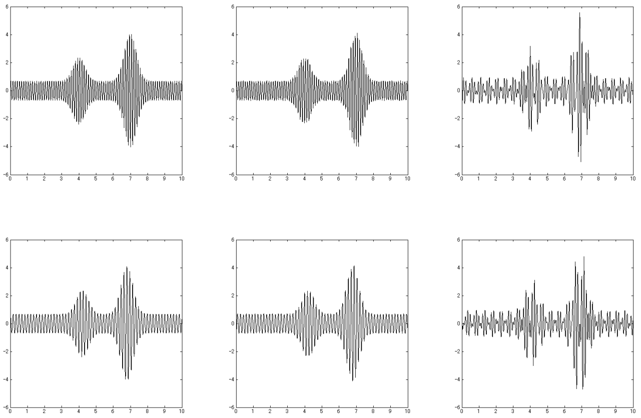

Fig. 1. Top left and bottom left: the original sources. Top center and bottom center: the estimated sources by VDCA. Top right and bottom right: the estimated sources by FastICA (tanh). The quasi-stochastic gradient algorithm with the stepsize 0.01 was run 10 times, and the estimates with the smallest value of the objective function were taken to avoid getting stuck in local minimum. Then our algorithm separated 100% of the sources (100 replications), whereas FastICA worked poorly (2%).

of source statistics and mixing coefficients that uniquely determines the second- and fourth-order moment structures of the model $\boldsymbol{\sigma}_2(\boldsymbol{\tau})$ and $\boldsymbol{\sigma}_4(\boldsymbol{\tau})$. Then the $\boldsymbol{\sigma}_2(\boldsymbol{\tau})$, $\boldsymbol{\sigma}_4(\boldsymbol{\tau})$ and $\boldsymbol{\tau}$ can be written as

$$\boldsymbol{\sigma}_i(\boldsymbol{\tau}) = \mathbf{H}_i E[\, \overbrace{\boldsymbol{x} \otimes \cdots \otimes \boldsymbol{x}}^{i \text{ times}} \,] \quad (i = 2, 4), \tag{8}$$

where the symbol $\otimes$ denotes the Kronecker product[3] and $\mathbf{H}_i$ is a selection matrix of order $\binom{m+i-1}{i} \times m^i$ $(i = 2, 4)$ that selects non-duplicated elements. The parameter vector $\boldsymbol{\tau}$ consists of $\bar{\mathbf{A}}$ and $E(s_p^2 s_q^2)$.

In [9], we proposed that the model is estimated using the principle of generalized least-square estimation. This is a method of matching the moments of the observed data $\boldsymbol{m}_i$ and those based on the model $\boldsymbol{\sigma}_i(\boldsymbol{\tau})$ in a weighted least-squares sense $(i = 2, 4)$.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be a random sample from the VDCA model as defined in Section 2, and define the sample counterparts to the moments in (8) as

$$\boldsymbol{m}_i = \frac{1}{N} \mathbf{H}_i \sum_{t=1}^{N} \overbrace{\boldsymbol{x}_t \otimes \cdots \otimes \boldsymbol{x}_t}^{i \text{ times}} \quad (i = 2, 4). \tag{9}$$

[3] The Kronecker product $\mathbf{X} \otimes \mathbf{Y}$ of matrices $\mathbf{X}$ and $\mathbf{Y}$ is defined as a partitioned matrix with (i, j)-th block equal to $x_{ij}\mathbf{Y}$.

Let us denote by $\boldsymbol{\tau}_0$ the true parameter vector. The $\boldsymbol{\sigma}_i(\boldsymbol{\tau}_0)$ can be estimated by the $\boldsymbol{m}_i$ when N is enough large: $\boldsymbol{\sigma}_i(\boldsymbol{\tau}_0) \approx \boldsymbol{m}_i$ $(i = 2, 4)$.

The GLS estimator of $\boldsymbol{\tau}$ is obtained as

$$\widehat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \left\| \begin{bmatrix} \boldsymbol{m}_2 \\ \boldsymbol{m}_4 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\sigma}_2(\boldsymbol{\tau}) \\ \boldsymbol{\sigma}_4(\boldsymbol{\tau}) \end{bmatrix} \right\|^2_{\widehat{\mathbf{U}}^{-1}} . \tag{10}$$

(For simplicity, the norm $\boldsymbol{y}^T \mathbf{M} \boldsymbol{y}$ of a vector $\boldsymbol{y}$ associated with a nonnegative definite matrix $\mathbf{M}$ is here expressed as $\|\boldsymbol{y}\|^2_{\mathbf{M}}$.) Here $\widehat{\mathbf{U}}$ is a weight matrix in GLS estimation and converges in probability to a certain positive definite matrix $\mathbf{U}$. The resultant GLS estimator $\widehat{\boldsymbol{\tau}}$ determined by (10) is then consistent and asymptotic normal [11]. We simply take the identity matrix as $\widehat{\mathbf{U}}$ in the following.

4 A quasi-stochastic gradient algorithm

In this section, we propose a simple optimization algorithm for the least squares approach above. We assume that the data is prewhitened by a whitening matrix $\mathbf{V}$ in an ordinary way [3] and denote by $\boldsymbol{z} = \mathbf{V}\boldsymbol{x}$ the prewhitened signals. Then we can constrain $\mathbf{A} = \mathbf{V}\bar{\mathbf{A}}$ to be orthogonal, which stabilizes the algorithm below.

The total objective function (10) in the GLS approach becomes monstrous because we have sum over all the moments[4], and it only works for small dimensions. Denote by $\widetilde{\boldsymbol{a}}_i$ the i-th row of $\mathbf{A}$ and by $\mathbf{C}$ a symmetric matrix whose (i, j)-th element is $E(s_i^2 s_j^2)$. (Note that $\boldsymbol{\tau}$ consists of the elements of $\mathbf{A}$ and the lower triangular elements of $\mathbf{C}$.) A simple way to solve this problem would be to consider the objective function as a sum over the variable indices i, j, k, l:

$$\sum_{i,j,k,l} J_{ijkl}(\mathbf{A}, \mathbf{C}), \tag{11}$$

where

$$J_{ijkl}(\mathbf{A}, \mathbf{C}) = \left\{ \frac{1}{N} \sum_{t=1}^{N} z_{it} z_{jt} z_{kt} z_{lt} - E(\widetilde{\boldsymbol{a}}_i \boldsymbol{s}, \widetilde{\boldsymbol{a}}_j \boldsymbol{s}, \widetilde{\boldsymbol{a}}_k \boldsymbol{s}, \widetilde{\boldsymbol{a}}_l \boldsymbol{s}) \right\}^2 . \tag{12}$$

Let us compute the gradient of J_{ijkl} with respect to $\mathbf{A}$ and $\mathbf{C}$, denoted by $\nabla_{\mathbf{A}} J_{ijkl}$ and $\nabla_{\mathbf{C}} J_{ijkl}$, respectively (see Appendix A for the complete formulas). We can now update the estimate of $\mathbf{A}$ and $\mathbf{C}$ by taking *random* indices i, j, k, l at each iteration and using a simple gradient descent for J_{ijkl} (see Step 4 in the algorithm below). At each gradient step, we take new random indices i, j, k, l. This kind of a stochastic gradient descent finds the minimum of the sum of the J_{ijkl} that we wanted to minimize in the first place, because the gradient is *on the average* the same as the gradient of the whole sum.

[4] The number of fourth-order moments is of order n^4.

To improve the convergence, it is quite useful to perform a projection of the gradient on the tangent surface of the set of orthogonal matrices [12]. This means replacing the gradient $\nabla_{\mathbf{A}} J_{ijkl}$ by

$$\nabla^{ort}_{\mathbf{A}} J_{ijkl} = \nabla_{\mathbf{A}} J_{ijkl} - \mathbf{A}(\nabla_{\mathbf{A}} J_{ijkl})^T \mathbf{A}. \tag{13}$$

Thus, the estimation consists of the following steps:

0. Remove the mean from the data and whiten it. Choose (random) initial values for the matrices $\mathbf{A}$ and $\mathbf{C}$.
1. Randomly choose four indices i, j, k, l.
2. Compute the gradients with respect to $\mathbf{A}$ and $\mathbf{C}$ as given in Appendix A.
3. Compute the projected gradient with respect to $\mathbf{A}$ by (13).
4. Do a gradient step

$$\mathbf{A} \leftarrow \mathbf{A} - \mu \nabla^{ort}_{\mathbf{A}} J_{ijkl} \tag{14}$$

$$\mathbf{C} \leftarrow \mathbf{C} - \mu \nabla_{\mathbf{C}} J_{ijkl}, \tag{15}$$

   where μ is a small stepsize constant.
5. Orthogonalize $\mathbf{A}$ by

$$\mathbf{A} \leftarrow (\mathbf{A}\mathbf{A}^T)^{-1/2}\mathbf{A}. \tag{16}$$

The five steps 1-5 are repeated until $\mathbf{A}$ and $\mathbf{C}$ have converged. Then we obtain the estimate of $\bar{\mathbf{A}}$ by $\mathbf{V}^{-1}\mathbf{A}$.

5 Simulations

We conducted simulations to study the empirical performance of the algorithm above. The simulation consisted of 100 source separation trials with three different methods: 1) the quasi-stochastic gradient algorithm proposed in the paper; 2) FastICA using kurtosis and 3) FastICA using hyperbolic tangent function [10]. For the two FastICA, the symmetric orthogonalization was made. (The FastICA with the symmetric orthogonalization using hyperbolic tangent function as the nonlinearity is basically the same as the quasi-maximum likelihood estimation [13].) We took 0.1 as the stepsize and stopped the quasi-stochastic gradient iteration when the *average* change of orthogonalized mixing matrices measured by $1 - \min\{\mathrm{diag}(\mathbf{A}^T_{old}\mathbf{A}_{new})\}$ over the last 100 iterations is smaller than 0.0001[5].

In each trial, we generated 10 sources that were dependent through their variances and created observed signals following the VDCA model as defined in Section 2. First, we created a random signal v_0 with several sample sizes (3,000, 5,000, 10,000, 30,000) where their components were independently distributed according to the gaussian distribution with zero mean and unit variance. Outliers, defined as values larger than a threshold of 3 times the standard deviation,

[5] Here, the quasi-stochastic gradient algorithm was run once for each data.

were eliminated from the resulting signals by reducing their values to the above-mentioned threshold. The variance signals v_i were then defined as the absolute values of the signal, that is, $v_i = |v_0|$ $(i = 1, \cdots, 10)$. The variance signals were completely dependent on each other since they were identical, but they were independent over time. (Therefore, the double-blind method [7] that used temporal correlations was not applicable to this case.)

Next the source signals s_i were created by multiplying the variance signals v_i by ten-dimensional random signals y_i, that is, $s_i = v_i y_i$. Here, the ten underlying signals y_i were i.i.d. (white) zero-mean subgaussian random processes to create enough variance dependencies [7]. (The subgaussian signals were signed fourth root of zero mean-uniform variables.) The source signals were normalized to have zero means and unit variances. Finally, a random mixing matrix $\bar{\mathbf{A}}$ was created, and the signals were mixed to provide the observed signals $x_i, i = 1, \cdots, 10$.

The three methods were then applied on the data after prewhitening it. The performance of each method was assessed as follows. Denoting by $\mathbf{W}$ the transpose of the obtained estimate of the orthogonalized mixing matrix $\mathbf{A}$ (with permutation and sign indeterminacies), we looked at the matrix $\mathbf{WV}\bar{\mathbf{A}}$. We computed how many elements in this matrix had an absolute value that was larger than 0.90. First of all, it must be noted that the matrix $\mathbf{WV}\bar{\mathbf{A}}$ is rather exactly orthogonal (up to insignificant errors occurred in the estimation of the whitening matrix), so there can be no more than 10 such elements in the matrix, and no row or column can contain more than one such element. In the ideal case where $\mathbf{WV}\bar{\mathbf{A}}$ is a signed permutation matrix, there would be exactly 10 such elements. Thus, this gave a measure of how many source signals had been separated.

The results are shown in Table 1. Our method separated more than 97.0% of the components for the reasonable sample sizes (5,000, 10,000, 30,000). On the other hand, both FastICAs could not separate the components at all (0%) since FastICA is based on independence of sources. Thus, our method was quite good, while not being perfect.

Table 1. Percentage of components recovered (100 replications)

	Sample size			
	3,000	5,000	10,000	30,000
Stoc. grad. alg.	87.4	97.6	97.8	97.6
FastICA (kurtosis)	0	0	0	0
FastICA (tanh)	0	0	0	0

6 Conclusions

We proposed a quasi-stochastic gradient algorithm for the GLS approach using second- and fourth-order moment structures of observed signals to the blind source separation of sources that are dependent only through their variances. In the approach, we do not have to assume that the sources have some temporal structures nor postulate any parametric models for their dependencies. This could be a big advantage of our approach over the conventional methods.

Although our method works well in simulations, moment-based methods often suffer from sensitivity to outliers when applied on certain kinds of real data. An important question for future research is to investigate how serious this problem is and, eventually, how it can be alleviated.

Acknowledgements

This work was partially carried out at Division of Mathematical Science, Osaka University and Transdisciplinary Research Integration Center, Research Organization of Information and Systems. A.H. was supported by the Academy of Finland through an Academy Research Fellow Position and project #203344. S.S. was supported by Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science.

References

1. Jutten, C., Hérault, J.: Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing **24** (1991) 1–10
2. Comon, P.: Independent component analysis. a new concept? Signal Processing **36** (1994) 62–83
3. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley, New York (2001)
4. Bach, F.R., Jordan, M.I.: Tree-dependent component analysis. In: Proc. the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002). (2002)
5. Hyvärinen, A.: A unifying model for blind separation of independent sources. Signal Processing **85** (2005) 1419–1427
6. Hyvärinen, A., Hoyer, P.O., Inki, M.: Topographic independent component analysis. Neural Computation **13** (2001) 1525–1558
7. Hyvärinen, A., Hurri, J.: Blind separation of sources that have spatiotemporal dependencies. Signal Processing **84** (2004) 247–254
8. Kawanabe, M., Müller, K.R.: Estimating functions for blind separation when sources have variance-dependencies. In: Proc. 5th International Conference on ICA and Blind Source Separation, Granada, Spain. (2004) 136–143
9. Shimizu, S., Hyvärinen, A., Kano, Y.: A generalized least squares approach to blind separation of sources which have variance dependency. In: Proc. IEEE Workshop on Statistical Signal Processing (SSP2005). (2005)
10. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Networks **10** (1999) 626–634
11. Ferguson, T.S.: A method of generating best asymptotically normal estimates with application to estimation of bacterial densities. Annals of Mathematical Statistics **29** (1958) 1046–1062
12. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications **20** (1998) 303–353

13. Hyvärinen, A.: The fixed-point algorithm and maximum likelihood estimation for independent component analysis. Neural Processing Letters **10** (1999) 1–5

A Gradient of the objective function

The gradients of the objective function in (12) are as follows:

$$\nabla_{\mathbf{A}} J_{ijkl} = -2\left\{\frac{1}{N}\sum_{t=1}^{N} z_{it}z_{jt}z_{kt}z_{lt} - E(z_i z_j z_k z_l)\right\}\frac{\partial E(z_i z_j z_k z_l)}{\partial \mathbf{A}} \tag{17}$$

$$\nabla_{\mathbf{C}} J_{ijkl} = -2\left\{\frac{1}{N}\sum_{t=1}^{N} z_{it}z_{jt}z_{kt}z_{lt} - E(z_i z_j z_k z_l)\right\}\frac{\partial E(z_i z_j z_k z_l)}{\partial \mathbf{C}}. \tag{18}$$

In what follows, we provide $E(z_i z_j z_k z_l)$ that were given by the VDCA model and their first derivatives with respect to $\mathbf{A}$ and $\mathbf{C}$ to compute $\nabla_{\mathbf{A}} J_{ijkl}$ and $\nabla_{\mathbf{C}} J_{ijkl}$ above.

We first provide the model-based expectations $E(z_i z_j z_k z_l)$:

$$\begin{aligned}
E(z_i^4) &= \sum_p a_{ip}^4 E(s_p^4) + 6\sum_{p<q} a_{ip}^2 a_{iq}^2 E(s_p^2 s_q^2) \\
E(z_i^3 z_j) &= \sum_p a_{ip}^3 a_{jp} E(s_p^4) + 3\sum_{p<q}(a_{ip}^2 a_{iq} a_{jq} + a_{ip} a_{jp} a_{iq}^2) E(s_p^2 s_q^2) \\
E(z_i^2 z_j z_k) &= \sum_p a_{ip}^2 a_{jp} a_{kp} E(s_p^4) + \sum_{p<q}(a_{ip}^2 a_{jq} a_{kq} + 2a_{ip} a_{jp} a_{iq} a_{kq} \\
&\quad +2a_{ip} a_{kp} a_{iq} a_{jq} + a_{iq}^2 a_{jp} a_{kp}) E(s_p^2 s_q^2) \\
E(z_i^2 z_j^2) &= \sum_p a_{ip}^2 a_{jp}^2 E(s_p^4) + \sum_{p<q}(a_{ip}^2 a_{jq}^2 + a_{iq}^2 a_{jp}^2 + 4a_{ip} a_{jp} a_{iq} a_{jq}) E(s_p^2 s_q^2) \\
E(z_i z_j z_k z_l) &= \sum_p a_{ip} a_{jp} a_{kp} a_{lp} E(s_p^4) + \sum_{p<q}(a_{ip} a_{jp} a_{kq} a_{lq} + a_{ip} a_{jq} a_{kp} a_{lq} \\
&\quad +a_{ip} a_{jq} a_{kq} a_{lp} + a_{iq} a_{jq} a_{kp} a_{lp} + a_{iq} a_{jp} a_{kq} a_{lp} + a_{iq} a_{jp} a_{kp} a_{lq}) E(s_p^2 s_q^2).
\end{aligned}$$

Next, we give the first derivatives:

$$\begin{aligned}
\frac{\partial E(z_i^4)}{\partial a_{ip}} &= 4a_{ip}^3 E(s_p^4) + 12\sum_{q\neq p} a_{ip} a_{iq}^2 E(s_p^2 s_q^2), \\
\frac{\partial E(z_i^4)}{\partial E(a_{rp})} &= 0\ (r \neq i, l),\ \frac{\partial E(z_i^4)}{\partial E(s_p^4)} = a_{ip}^4,\ \frac{\partial E(z_i^4)}{\partial E(s_p^2 s_q^2)} = 6a_{ip}^2 a_{iq}^2 \\
\frac{\partial E(z_i^3 z_j)}{\partial a_{ip}} &= 3a_{ip}^2 a_{jp} E(s_p^4) + 3\sum_{q\neq p}(2a_{ip} a_{iq} a_{jq} + a_{jp} a_{iq}^2) E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^3 z_j)}{\partial a_{jp}} &= a_{ip}^3 E(s_p^4) + 3\sum_{q\neq p} a_{ip} a_{iq}^2 E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^3 z_j)}{\partial a_{rp}} &= 0\ (r \neq i, j),\ \frac{\partial E(z_i^3 z_j)}{\partial E(s_p^4)} = a_{ip}^3 a_{jp} \\
\frac{\partial E(z_i^3 z_j)}{\partial E(s_p^2 s_q^2)} &= 3(a_{ip}^2 a_{iq} a_{jq} + a_{ip} a_{jp} a_{iq}^2)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial E(z_i^2 z_j z_k)}{\partial a_{ip}} &= 2a_{ip}a_{jp}a_{kp}E(s_p^4) + \sum_{q\neq p}(2a_{ip}a_{jq}a_{kq} + 2a_{jp}a_{iq}a_{kq} + 2a_{kp}a_{iq}a_{jq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^2 z_j z_k)}{\partial a_{jp}} &= a_{ip}^2 a_{kp}E(s_p^4) + \sum_{q\neq p}(2a_{ip}a_{iq}a_{kq} + a_{iq}^2 a_{kp})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^2 z_j z_k)}{\partial a_{kp}} &= a_{ip}^2 a_{jp}E(s_p^4) + \sum_{q\neq p}(2a_{ip}a_{iq}a_{jq} + a_{iq}^2 a_{jp})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^2 z_j z_k)}{\partial a_{rp}} &= 0 \ (r \neq i,j,k), \ \ \frac{\partial E(z_i^2 z_j z_k)}{\partial E(s_p^4)} = a_{ip}^2 a_{jp} a_{kp} \\
\frac{\partial E(z_i^2 z_j z_k)}{\partial E(s_p^2 s_q^2)} &= a_{ip}^2 a_{jq}a_{kq} + 2a_{ip}a_{jp}a_{iq}a_{kq} + 2a_{ip}a_{kp}a_{iq}a_{jq} + a_{iq}^2 a_{jp}a_{kp} \\
\frac{\partial E(z_i^2 z_j^2)}{\partial a_{ip}} &= 2a_{ip}a_{jp}^2 E(s_p^4) + \sum_{q\neq p}(2a_{ip}a_{jq}^2 + 4a_{jp}a_{iq}a_{jq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^2 z_j^2)}{\partial a_{jp}} &= 2a_{ip}^2 a_{jp}E(s_p^4) + \sum_{q\neq p}(2a_{iq}^2 a_{jp} + 4a_{ip}a_{iq}a_{jq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i^2 z_j^2)}{\partial a_{rp}} &= 0 \ (r \neq i,j), \ \ \frac{\partial E(z_i^2 z_j^2)}{\partial E(s_p^4)} = a_{ip}^2 a_{jp}^2 \\
\frac{\partial E(z_i^2 z_j^2)}{\partial E(s_p^2 s_q^2)} &= a_{ip}^2 a_{jq}^2 + a_{iq}^2 a_{jp}^2 + 4a_{ip}a_{jp}a_{iq}a_{jq} \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial a_{ip}} &= a_{jp}a_{kp}a_{lp}E(s_p^4) + \sum_{q\neq p}(a_{jp}a_{kq}a_{lq} + a_{jq}a_{kp}a_{lq} + a_{jq}a_{kq}a_{lp})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial a_{jp}} &= a_{ip}a_{kp}a_{lp}E(s_p^4) + \sum_{q\neq p}(a_{ip}a_{kq}a_{lq} + a_{iq}a_{kq}a_{lp} + a_{iq}a_{kp}a_{lq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial a_{kp}} &= a_{ip}a_{jp}a_{lp}E(s_p^4) + \sum_{q\neq p}(a_{ip}a_{jq}a_{lq} + a_{iq}a_{jq}a_{lp} + a_{iq}a_{jp}a_{lq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial a_{lp}} &= a_{ip}a_{jp}a_{kp}E(s_p^4) + \sum_{q\neq p}(a_{ip}a_{jq}a_{kq} + a_{iq}a_{jq}a_{kp} + a_{iq}a_{jp}a_{kq})E(s_p^2 s_q^2) \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial a_{rp}} &= 0 \ (r \neq i,j,k,l), \ \ \frac{\partial E(z_i z_j z_k z_l)}{\partial E(s_p^4)} = a_{ip}a_{jp}a_{kp}a_{lp} \\
\frac{\partial E(z_i z_j z_k z_l)}{\partial E(s_p^2 s_q^2)} &= a_{ip}a_{jp}a_{kq}a_{lq} + a_{ip}a_{jq}a_{kp}a_{lq} \\
&\quad + a_{ip}a_{jq}a_{kq}a_{lp} + a_{iq}a_{jq}a_{kp}a_{lp} + a_{iq}a_{jp}a_{kq}a_{lp} + a_{iq}a_{jp}a_{kp}a_{lq}.
\end{aligned}
$$