FAST ICA FOR NOISY DATA USING GAUSSIAN MOMENTS

Aapo Hyvärinen

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, FIN-02015 HUT, Finland
aapo.hyvarinen@hut.fi
http://www.cis.hut.fi/~aapo/

ABSTRACT

A novel approach for the problem of estimating the data model of independent component analysis (or blind source separation) in the presence of gaussian noise is introduced. We define the gaussian moments of a random variable as the expectations of the gaussian function (and some related functions) with different scale parameters, and show how the gaussian moments of a random variable can be estimated from noisy observations. This enables us to use gaussian moments as one-unit contrast functions that have no asymptotic bias even in the presence of noise, and that are robust against outliers. To implement efficiently the maximization of the contrast functions based on gaussian moments, a modification of our FastICA algorithm is introduced.

1. INTRODUCTION

Indendent component analysis [5, 17] is a statistical model where the observed data is expressed as a linear transformation of latent variables ('independent components') that are nongaussian and mutually independent. Important applications of ICA are e.g. blind source separation and feature extraction [17, 18]. We may express the model as

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)$ is the vector of the latent variables called the independent components, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The vector $\mathbf{n}$ is noise, and is often omitted; most research has concentrated on the problem of estimating the noise-free model [1, 2, 3, 5, 15, 9, 17, 22]. For simplicity, we make in this paper some assumptions that are not strictly necessary: 1) the dimension of $\mathbf{s}$ equals the dimension of $\mathbf{x}$, i.e. $n = m$, 2) the noise $\mathbf{n}$ is gaussian and 3) the noise covariance matrix $\mathbf{\Sigma}$ is known. In practice, this essentially means that we use some additional method or prior knowledge to estimate the order of the model and the covariance matrix; such methods are somewhat independent of the method for estimating the mixing matrix $\mathbf{A}$.

In this paper, we introduce a novel approach to the estimation of the noisy ICA model in (1). Using the concept of gaussian moments, we show how it is possible to estimate some higher-order statistics of the original (noise-free) data using only noisy observations. This property is somewhat similar to the property that higher-order cumulants are immune to gaussian noise, but using gaussian moments, we have a larger repertoire of nonquadratic functions that can be used on the algorithms. In particular, we may choose functions that are robust against outliers and/or reduce asymptotic (mean-square) error. The simplest way to use the gaussian moments is in the form of one-unit contrast functions [9, 13]. As a practical method for optimizing the contrast functions, a modification of the FastICA algorithm, which is based on a fixed-point iteration scheme [15, 9, 13, 8], is introduced.

2. ONE-UNIT ALGORITHMS FOR NOISY DATA USING KURTOSIS

Before introducing gaussian moments, we first show how to estimate the noisy ICA model using higher-order cumulants, especially kurtosis. Our approach is based on the one-unit (or deflation) methods for noise-free ICA [6, 9], which are closely related to projection pursuit. Let us denote the noise-free data in the following by

$$\mathbf{y} = \mathbf{A}\mathbf{s}. \tag{2}$$

The basic idea in the one-unit approach is to take some measure of nongaussianity and then find projections, say $\mathbf{w}^T\mathbf{y}$, in which this is locally maximized for sphered (whitened) data, with constraint $\|\mathbf{w}\| = 1$. Projections in such directions give consistent estimates of the independent components, if the measure of nongaussianity is well chosen. This approach could be used for noisy ICA as well, if only we had measures of nongaussianity which are immune to gaussian noise, or at least, whose values for the original data can be easily estimated from noisy observations.

If the measure of nongaussianity is kurtosis [6] (the fourth-order cumulant), it is almost trivial to construct one-unit methods for noisy ICA, because kurtosis is immune to gaussian noise. It must be noted, however, that in the preliminary whitening, the effect of noise must be taken into account; this is quite simple if the noise covariance matrix is known. Denoting by $\mathbf{C} = E\{\mathbf{x}\mathbf{x}^T\}$ the covariance matrix of the observed noisy data, the ordinary whitening should be replaced by the operation

$$\tilde{\mathbf{x}} = (\mathbf{C} - \mathbf{\Sigma})^{-1/2}\mathbf{x}. \tag{3}$$

In other words, the covariance matrix $\mathbf{C}-\mathbf{\Sigma}$ of the noise-free data should be used in whitening instead of the covariance matrix $\mathbf{C}$ of the noisy data. In the following, we call this operation 'quasi-whitening'. After this operation, the quasi-whitened data $\tilde{\mathbf{x}}$ follows a noisy ICA model as well:

$$\tilde{\mathbf{x}} = \mathbf{B}\mathbf{s} + \tilde{\mathbf{n}} \tag{4}$$

where $\mathbf{B}$ is *orthogonal*, and $\tilde{\mathbf{n}}$ is a linear transform of the original noise in (1). Thus, the theorem in [6] is valid for $\tilde{\mathbf{x}}$, and finding local maxima of the absolute value of kurtosis is a valid method for estimating the independent components.

3. A FAMILY OF ONE-UNIT CONTRAST FUNCTIONS

It has been argued e.g. in [9, 10] that kurtosis may be a rather poor measure of nongaussianity (contrast function) in many applications. This is because it gives estimators that are very sensitive to outliers, and have large mean-square errors (at least for supergaussian data). Therefore, in [9, 13, 12] an approach was developed in which the higher-order statistics of the projection $\mathbf{w}^T\mathbf{y}$ are taken into account through general contrast functions of the form

$$J_G(\mathbf{w}^T\mathbf{y}) = |E\{G(\mathbf{w}^T\mathbf{y})\} - E\{G(\nu)\}|^p \tag{5}$$

where $p = 1, 2$, the function G is a sufficiently regular non-quadratic function, and ν is a standardized gaussian variable. It has been proven [16] that finding maxima of (5) for whitened data, under the constraint $\|\mathbf{w}\| = 1$ allows for the estimation of the noise-free ICA model under certain assumptions. Moreover, such contrast functions can be interpreted as approximations of differential entropy [12].

These one-unit contrast functions enable estimation of independent components one-by-one, thus without prior knowledge on the number of the independent components. Moreover, the several one-unit contrasts can be used in parallel, which is approximately equivalent to ML estimation [14]. It must be noted, however, that our contrast functions do not require prior knowledge on the nature of the distributions of the independent components, either.

4. UNBIASED CONTRASTS USING GAUSSIAN MOMENTS

The approach of the preceding section could be used for noisy data as well, if only we were able to estimate $J_G(\mathbf{w}^T\mathbf{y})$ of the noise-free data from the noisy observations $\mathbf{x}$. Denoting by z a nongaussian random variable, and by n a gaussian noise variable of variance σ^2, we should be able to express the relation between $E\{G(z)\}$ and $E\{G(z+n)\}$ in simple algebraic terms. In general, this relation seems quite complicated, and can be computed only using numerical integration. The main point of this paper is to show that for certain choices of G, a similar relation becomes very simple. The basic idea is to choose G to be the density function of a zero-mean gaussian random variable, or a related function.

Denote by

$$\varphi_c(x) = \frac{1}{c}\varphi(\frac{x}{c}) = \frac{1}{\sqrt{2\pi}c}\exp(-\frac{x^2}{2c^2}) \tag{6}$$

the gaussian density function of variance c^2, and by $\varphi_c^{(k)}(x)$ the k-th ($k > 0$) derivative of $\varphi_c(x)$. Denote further by $\varphi_c^{(-k)}$ the k-th integral function of $\varphi_c(x)$, obtained by $\varphi_c^{(-k)}(x) = \int_0^x \varphi_c^{(-k+1)}(\xi)d\xi$, where we define $\varphi_c^{(0)}(x) = \varphi_c(x)$. (The lower integration limit 0 is here quite arbitrary, but has to be fixed.) Then we have the following theorem (proven in Appendix A):

Theorem 1 *Let z be any nongaussian random variable, and denote by n an independent gaussian noise variable of variance σ^2. Define the gaussian function φ as in (6). Then for any constant $c > \sigma^2$ we have*

$$E\{\varphi_c(z)\} = E\{\varphi_d(z+n)\} \tag{7}$$

with $d = \sqrt{c^2 - \sigma^2}$. Moreover, (7) still holds when φ is replaced by $\varphi^{(k)}$ for any integer index k.

The theorem means that we can estimate the independent components from noisy observations by maximizing a general contrast function of the form (5), where the direct estimation of the statistics $E\{G(\mathbf{w}^T\mathbf{y})\}$ of the noise-free data is made possible by using $G(u) = \varphi_c^{(k)}(u)$. We call the statistics of the form $E\{\varphi_c^{(k)}(\mathbf{w}^T\mathbf{y})\}$ the *gaussian moments* of the data. Thus we maximize, for quasi-whitened data $\tilde{\mathbf{x}}$, the following contrast function:

$$\max_{\|\mathbf{w}\|=1} |E\{\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\tilde{\mathbf{x}})\} - E\{\varphi_c^{(k)}(\nu)\}|^p \tag{8}$$

with $d(\mathbf{w}) = \sqrt{c^2 - \mathbf{w}^T\tilde{\mathbf{\Sigma}}\mathbf{w}}$. This gives a consistent (i.e. convergent) method of estimating the noisy ICA model due to the theorem in [16].

To use these results in practice, we need to choose some values for c and k. (The value of p is of little consequence.) The choice of c is in fact avoided by the developments in the next section. Two indices k for the gaussian moments seem to be of particular interest: $k = 0$ and $k = -2$. The first corresponds to the gaussian density function; such a contrast function has been used succesfully in noise-free ICA [9] and its use can be justified from the viewpoint of robust statistics [10]. The case $k = -2$ is interesting because the contrast function is then of the form of a (negative) log-density of a supergaussian variable. In fact, $\varphi^{(-2)}(u)$ can be very accurately approximated by $G(u) = 1/2 \log\cosh u$, which has been widely used in ICA [2, 9, 22].

5. FAST OPTIMIZATION FOR GAUSSIAN MOMENTS

To perform the optimization in (8), we can derive a modification of the FastICA algorithm [9, 8, 13]. Fast ICA is a computationally optimized algorithm that optimizes one-unit contrast functions considerably faster than ordinary gradient methods. It can also be used to fast maximization of the likelihood [14]. Speed-up factors of in the range of 10 to 100 are often observed.

First, we derive the FastICA algorithm for noisy data usin kurtosis as the contrast function. Modifying slightly the derivation in [15], we obtain the following form for the FastICA algorithm:

$$\mathbf{w}^* = E\{\tilde{\mathbf{x}}(\mathbf{w}^T\tilde{\mathbf{x}})^3\} - 3(\mathbf{I} + \tilde{\mathbf{\Sigma}})\mathbf{w}\mathbf{w}^T(\mathbf{I} + \tilde{\mathbf{\Sigma}})\mathbf{w} \tag{9}$$

where $\mathbf{w}^*$, the new value of $\mathbf{w}$, is normalized to unit norm after every iteration, and

$$\tilde{\mathbf{\Sigma}} = E\{\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\} = (\mathbf{C} - \mathbf{\Sigma})^{-1/2}\mathbf{\Sigma}(\mathbf{C} - \mathbf{\Sigma})^{-1/2} \tag{10}$$

is the covariance of the noise after quasi-whitening. The convergence proof in [15] is valid for (9) as well, showing that the algorithm has global, cubic convergence. Several independent components can be found using different orthogonalization schemes, exactly as in the noise-free case [15].

In the (more interesting) case where we use gaussian moments in the contrast functions, we can replace in the noise-free version all the expectations by gaussian moments that give consistent estimates of the corresponding gaussian moments of the original data. A detailed derivation is given in Appendix B. Thus we obtain the following preliminary form of the fixed-point iteration for quasi-whitened data:

$$\mathbf{w}^* = E\{\tilde{\mathbf{x}}\varphi_{d(\mathbf{w})}^{(k+1)}(\mathbf{w}^T\tilde{\mathbf{x}})\} - (\mathbf{I} + \tilde{\mathbf{\Sigma}})\mathbf{w}E\{\varphi_{d(\mathbf{w})}^{(k+2)}(\mathbf{w}^T\tilde{\mathbf{x}})\} \tag{11}$$

where $\mathbf{w}^*$, the new value of $\mathbf{w}$, is normalized to unit norm after every iteration, and $\tilde{\mathbf{\Sigma}}$ is given by (10).

The fixed-point algorithm in (11) can be considerably simplified by adapting the value of c at every iteration. At the same time, this solves the problem of choosing values for the parameter c. Such an adaptation of c is justified by the fact that the function G needs only to be of a given shape, so that the signs of certain non-polynomial cumulants do not change [3, 16]. Moderate changes of c do not thus change the convergence of the algorithm. For example, one could adapt c before every step so that $d(\mathbf{w}) = \sqrt{c^2 - \mathbf{w}^T\tilde{\mathbf{\Sigma}}\mathbf{w}} = 1$.

This gives finally the following *FastICA algorithm with bias removal* for quasi-whitened data:

$$\boxed{\mathbf{w}^* = E\{\tilde{\mathbf{x}}g(\mathbf{w}^T\tilde{\mathbf{x}})\} - (\mathbf{I} + \tilde{\mathbf{\Sigma}})\mathbf{w}E\{g'(\mathbf{w}^T\tilde{\mathbf{x}})\}} \tag{12}$$

where $\mathbf{w}^*$, the new value of $\mathbf{w}$, is normalized to unit norm after every iteration, and $\tilde{\mathbf{\Sigma}}$ is given by (10). Surprisingly, (12) is of the same form as (9). The function g is here the derivative of G, and can thus be choosen among the following:

$$g_1(u) = \tanh(u), \quad g_2(u) = u\exp(-u^2/2), \quad g_3(u) = u^3, \tag{13}$$

where g_1 is an approximation of $\varphi^{(-1)}$, which is the gaussian cdf (these relations hold up to some irrelevant constants). These functions cover essentially the nonlinearities ordinarily used in the FastICA algorithm [9, 13]. It can be seen that the addition of $\tilde{\mathbf{\Sigma}}$ in (12) is the key to removing bias. Indeed, using the classical properties of kurtosis, Theorem 1 and the convergence proof of the fixed-point algorithm [13], it can be seen that this modification *removes the asymptotic bias* that noise produces in ordinary ICA algorithms.

As mentioned above, more than one independent components can be estimated using the same orthogonalization schemes as in the noise-free case [15]. It is also simple to derive adaptive one-unit learning rules as in [16].

6. SIMULATIONS

To test our algorithm in (12), we conducted lengthy experiments. The dimension of the data was 4, the independent components had i.i.d. Laplace distributions, and noise covariance was $\sigma^2\mathbf{I}$, where $\sigma = .25$. At each trial, a 4×4 mixing matrix was randomly generated, and normalized so that the total energy of the signals was equal to 1. This corresponds to a signal-to-noise ratio of 4. Badly conditioned mixing matrices (condition number >10) were not accepted, because any estimation procedure for noisy ICA is highly sensitive to the conditioning of the mixing matrix; badly conditioned matrices would produce outliers in the error measure, making the analysis more difficult. Only one independent component was estimated at each trial, and the resulting error was measured as:

$$\text{error} = \min_i |1 - |\mathbf{w}^T\mathbf{b}_i|/\|\mathbf{w}^T\mathbf{B}\|| \tag{14}$$

where $\mathbf{b}_i$ is the i-th row of the mixing matrix after quasi-whitening. Sample size N was varied from 1000 to 64000, and the error for given N was estimated as the median of the errors of 200 trials. At every trial, the FastICA algorithm was run with 50 iterations, which seemed to be always enough for convergence. The results are depicted in Fig. 1. The dotted lines gives log-errors for estimators without bias correction, using the 3 nonlinearities in (13); the horizontal axis shows the logarithm of sample size. In this case, errors do not tend to zero due to asymptotic bias. The solid, dashed, and dash-dotted lines give the errors for the 3 nonlinearities and using the bias correction. Now the errors tend to zero, showing lack of bias. This confirms that our modification of the FastICA algorithm is asymptotically unbiased, i.e. consistent.

7. DISCUSSION

In this paper we introduced a new approach to estimation of the noisy ICA model, using the concept of gaussian moments. The useful property of gaussian moments is that the gaussian moments of underlying random variables can be simply estimated from noisy observations. Thus we derived a FastICA algorithm for noisy ICA that is *computationally simple and very fast*, as well as *statistically consistent and robust against outliers*. Comparing our approach to other methods proposed for noisy ICA, we can conclude:

1. An Expectation-Maximization algorithm for noisy ICA [21] provides a statistically elegant estimator for noisy ICA, and has the benefit of being able to estimate the noise covariance matrix as well. An important disadvantage with that method is, however, that it has computationally exponential complexity (with respect to the dimension of the data), which essentially limits its use to small dimensions. The FastICA algorithm is computationally considerably more efficient; in the noise-free case, it has been succesfully applied for data sets that have more than 100 dimensions [18]. Moreover, it seems likely that the EM method of estimating the noise covariance could be used also in connection with our algorithm as well.

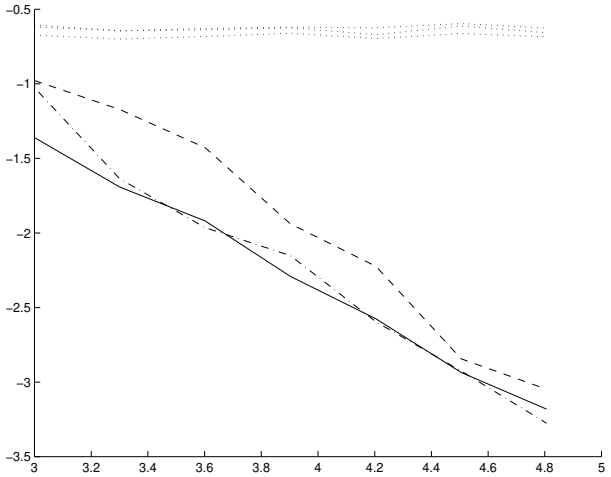

Figure 1: Convergence of the estimators for fixed noise level (SNR=4) and sample size varying from 1000 to 64000. Horizontal axis: log10 of sample size. Vertical axis: log10 of error in (14). Dotted lines: estimators without bias correction, for the three nonlinearities in (13). Other lines: estimators with bias correction (solid: g_3, dashed: g_2, dot-dashed: g_1). Only the estimators with bias correction have errors that tend to zero.

2. Bias reduction techniques for ML estimation of noisy ICA were proposed in [4, 7]. Such methods are computationally simple, but since they are based on Taylor approximations, they usually only reduce the asymptotic bias, whereas our method removes it completely. Moreover, the FastICA algorithm is computationally more efficient than the adaptive gradient methods in [4, 7] in environments where fast tracking is not needed.

3. The main benefit of our method, based on gaussian moments, with respect to cumulant-based methods [20] is that our method is more robust against outliers [10]. Furthermore, our methods seem to have, at least for low noise levels, smaller mean-square errors for most data sets [10].

4. Compared to methods using the joint likelihood of the mixing matrix and the independent components [11, 23], the FastICA algorithm is computationally considerably more efficient, since it reduces the dimension of the search space to a small fraction. Methods based on joint likelihood are useful, however, for solving the additional problem of nonlinear reconstruction of the independent components [11].

8. REFERENCES

[1] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing 8*, pages 757–763. MIT Press, Cambridge, MA, 1996.

[2] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[3] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[4] A. Cichocki, S. C. Douglas, and S.-I. Amari. Robust techniques for independent component analysis with noisy data. *Neurocomputing*, 22:113–129, 1998.

[5] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[6] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[7] S.C. Douglas, A. Cichocki, , and S. Amari. A bias removal technique for blind source separation with noisy measurements. *Electronics Letters*, 34:1379–1380, 1998.

[8] The FastICA MATLAB package. Available at `http://www.cis.hut.fi/projects/ica/fastica/`, 1998.

[9] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, Munich, Germany, 1997.

[10] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pages 388–397, Amelia Island, Florida, 1997.

[11] A. Hyvärinen. Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22:49–67, 1998.

[12] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems 10*, pages 273–279. MIT Press, 1998.

[13] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 1999. To appear.

[14] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 1999. To appear.

[15] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[16] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[17] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[18] J. Karhunen, A. Hyvärinen, R. Vigario, J. Hurri, and E. Oja. Applications of neural blind separation to signal and image processing. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 131–134, Munich, Germany, 1997.

[19] M. Kendall and A. Stuart. *The Advanced Theory of Statistics*. Charles Griffin & Company, 1958.

[20] L. De Lathauwer, B. De Moor, and J. Vandewalle. A technique for higher-order-only blind source separation. In *Proc. ICONIP*, Hong Kong, 1996.

[21] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3617–3620, Munich, Germany, 1997.

[22] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[23] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–340, May 1996.

A. PROOF OF THEOREM

Denote by $p(.)$ the pdf of z. For $k = 0$, we have

$$E\{\varphi_d(z+n)\} = \int \varphi_d(y)[\int \varphi_\sigma(y-t)p(t)dt]dy$$
$$= \int p(t)[\int \varphi_\sigma(y-t)\varphi_d(y)dy]dt = E\{\varphi_c(z)\} \quad (15)$$

which proves the theorem for $k = 0$. For other values of k, introduce a hypothetical location parameter θ. Taking the k-th derivative (resp. integral) under the expectation of the both sides of $E\{\varphi_c(z+\theta)\} = E\{\varphi_d(z+n+\theta)\}$, and setting $\theta = 0$, we obtain the theorem for $k > 0$ (resp. $k < 0$). (The lower integration limit for $k < 0$ must be set to 0 to comply with the definition above).

B. DERIVATION OF FIXED-POINT ALGORITHM FOR NOISY DATA

First note that

$$\varphi_c^{(k)}(x) = \varphi^{(k)}(\frac{x}{c})c^{(-k-1)}, \quad (16)$$

and denote as above

$$d(\mathbf{w}) = \sqrt{c^2 - \mathbf{w}^T\tilde{\mathbf{\Sigma}}\mathbf{w}}. \quad (17)$$

Then the gradient of $\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\mathbf{x})$ with respect to $\mathbf{w}$ can be obtained as

$$\nabla_{\mathbf{w}}\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\mathbf{x}) = \mathbf{x}\varphi_{d(\mathbf{w})}^{(k+1)}(\mathbf{w}^T\mathbf{x})$$
$$+ \tilde{\mathbf{\Sigma}}\mathbf{w}(c^2 - \mathbf{w}^T\tilde{\mathbf{\Sigma}}\mathbf{w})^{-1}(\mathbf{w}^T\mathbf{x})\varphi_{d(\mathbf{w})}^{(k+1)}(\mathbf{w}^T\mathbf{x})$$
$$+ \tilde{\mathbf{\Sigma}}\mathbf{w}(k+1)(c^2 - \mathbf{w}^T\tilde{\mathbf{\Sigma}}\mathbf{w})^{-1}\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\mathbf{x}) \quad (18)$$

To proceed, we need the following lemma

Lemma 1 *For all k, we have*

$$(k+1)\varphi^{(k)}(x) + x\varphi^{(k+1)}(x) = -\varphi^{(k+2)}(x). \quad (19)$$

Proof of lemma: For $k \geq 0$, the lemma follows from the properties of the Tshebyshev-Hermite polynomials [19]. For $k < 0$, take the Fourier transforms of both sides of (19):

$$(k+1)(i\xi)^k\varphi(\xi) + i[(i\xi)^{(k+1)}\varphi]'(\xi) = -(i\xi)^{(k+2)}\varphi(\xi) \quad (20)$$

which can multiplied by $1/(i\xi)$ to give

$$(k+1)(i\xi)^{k-1}\varphi(\xi) + i(i\xi)^k\varphi'(\xi) - (k+1)(i\xi)^{(k+1)}\varphi(\xi)$$
$$= -(i\xi)^{(k+1)}\varphi(\xi) \Leftrightarrow$$
$$k(i\xi)^{k-1}\varphi(\xi) + i[(i\xi)^k\varphi]'(\xi) = -(i\xi)^{(k+1)}\varphi(\xi) \quad (21)$$

which is in fact (20) for $k^* = k - 1$. Thus, by induction, lemma holds for all $k < 0$ as well.

The lemma implies

$$d^{-2}(k+1)d^{-k-1}\varphi^{(k)}(x/d) + d^{-2}xd^{-k-2}\varphi^{(k+1)}(x/d)$$
$$= -d^{-k-3}\varphi^{(k+2)}(x/d)$$
$$\Leftrightarrow$$
$$d^{-2}(k+1)\varphi_d^{(k)}(x) + d^{-2}x\varphi_d^{(k+1)}(x) = -\varphi_d^{(k+2)}(x) \quad (22)$$

This means the gradient in (18) can be expressed as

$$\nabla_{\mathbf{w}}\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\mathbf{x}) = \mathbf{x}\varphi_{d(\mathbf{w})}^{(k+1)}(\mathbf{w}^T\mathbf{x}) - \tilde{\mathbf{\Sigma}}\mathbf{w}\varphi_{d(\mathbf{w})}^{(k+2)}(\mathbf{w}^T\mathbf{x}). \quad (23)$$

The equation giving the fixed-point algorithm was given in [9] as

$$\mathbf{w}^* = E\{\nabla G(\mathbf{w}^T\mathbf{y})\} - \mathbf{w}E\{G''(\mathbf{w}^T\mathbf{y})\}. \quad (24)$$

Choosing $G(u) = \varphi_c^{(k)}(x)$, and using the above derivation, we obtain the gradient part as (23). By Theorem 1, we have

$$E\{G''(\mathbf{w}^T\mathbf{y})\} = E\{\varphi_{d(\mathbf{w})}^{(k+2)}(\mathbf{w}^T\mathbf{x})\}. \quad (25)$$

Thus we obtain (12).