Available online at www.sciencedirect.com
SCIENCE DIRECT®
NEUROCOMPUTING
ELSEVIER
Neurocomputing 57 (2004) 469–476
www.elsevier.com/locate/neucom

Letters

An EM algorithm for learning sparse and overcomplete representations

Mingjun Zhong[a,b], Huanwen Tang[a], Hongjun Chen[b,c], Yiyuan Tang[b,d,e,*]

[a] *Institute of Computational Biology and Bioinformatics, Dalian University of Technology, Dalian 116023, People's Republic of China*
[b] *Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, People's Republic of China*
[c] *Department of Foreign Languages, Dalian University of Technology, Dalian 116023, People's Republic of China*
[d] *Laboratory of Visual Information Processing, The Chinese Academy of Sciences, Beijing 100101, People's Republic of China*
[e] *Key Lab for Mental Health, The Chinese Academy of Sciences, Beijing 100101, People's Republic of China*

Abstract

An expectation-maximization (EM) algorithm for learning sparse and overcomplete representations is presented in this paper. We show that the estimation of the conditional moments of the posterior distribution can be accomplished by maximum a posteriori estimation. The approximate conditional moments enable the development of an EM algorithm for learning the overcomplete basis vectors and inferring the most probable basis coefficients.

Keywords: Overcomplete representations; EM algorithm; Maximum a posteriori

1. Introduction

Overcomplete representation [8,10,7,3] is a method of finding a representation of data in which only a few components of the representation are significantly activated at the same time. Important applications of overcomplete representations are in blind source

* Corresponding author. Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, China.
E-mail addresses: sunxl_zhong@yahoo.com (M. Zhong), yy2100@163.net (Y. Tang).

0925-2312/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2003.12.002

separation of more sources than mixtures [6] and sparse coding for natural data [10,7]. In learning overcomplete representations, the observed data vector $x=(x_1,\ldots,x_N)^{\mathrm{T}}$ can be formulated using an overcomplete basis by the following linear generative model:

$$x = As + \varepsilon, \tag{1}$$

where the columns of the matrix $A \in R^{N\times M}$, where $N \leqslant M$ define the overcomplete basis vectors, $s=(s_1,\ldots,s_M)^{\mathrm{T}}$ is the vector of basis coefficients, and the vector $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_N)^{\mathrm{T}}$ is noise which is modelled as Gaussian with zero mean and covariance matrix Σ. The basis coefficients are assumed independent such that $p(s)=\prod_{m=1}^{M} p(s_m)$ where p denotes the probability density function (p.d.f.) which is used throughout this paper, and are also assumed to be as sparse as possible, i.e., only a small number of the available coefficients are required to represent the data. Thus, the distribution of the basis coefficients can be typified by factorable Laplacian distribution such that [3,8]

$$p(s) = \left(\sqrt{2}\right)^{-M} \prod_{m=1}^{M} \exp\left(-\sqrt{2}\,|\,s_m|\right). \tag{2}$$

For simplicity, we assume in this paper that the noise covariance matrix Σ is known.

One approach for learning overcomplete representations derives from Lewicki and Sejnowski's gradient-based method [8] where there is a requirement for the assumption of a low level of noise. Another one derives from Girolami's variational method [3] where there is an additional computational cost for computing the variational parameters in each expectation-maximization (EM) step. Based on the EM algorithm [2], this paper presents a method for inferring the most probable basis coefficients and learning the overcomplete basis vectors. The conditional moments of the intractable posterior distribution are estimated by maximum a posteriori (MAP) estimation. Rather than using an explicit solution for the basis coefficients as in [4] (where coefficients are referred to as independent components in independent component analysis (ICA)), we estimate the basis coefficients by a gradient learning rule. Thus, these approximate conditional moments enable the development of an EM algorithm for learning the overcomplete basis vectors. This proposed EM algorithm generalizes the complete case or regular ICA with additive noise [1,4,5].

2. Maximum a posteriori estimation

The estimation of the basis coefficients can be accomplished by MAP estimation. Assuming that we have observed T data samples $x=x(1),\ldots,x(T)$ generated according to model (1), one obtains the log-likelihood (see Appendix A):

$$L(s) = \sum_{t=1}^{T} \left\{-\tfrac{1}{2}\,(x(t)-As(t))^{\mathrm{T}}\Sigma^{-1}(x(t)-As(t)) + \varphi(s(t))\right\} + C, \tag{3}$$

where $\varphi(s(t))=\log\{p(s(t))\}$ is a certain nonlinear function, and C is a constant irrelevant to $s(t)$. Note that this log-likelihood is essentially the log-posterior distribution,

i.e., $L(s) = \log\{p(s\,|\,x, A)\}$. Thus, this log-likelihood is essentially the joint likelihood proposed by Hyvärinen [4] which is used to estimate A and $s(t)$ in noisy ICA, and is also essentially the same as the objective function proposed by Olshausen and Field [10] as an approximation of the likelihood of A.

To infer the most probable basis coefficients, one must maximize the log-likelihood in Eq. (3). Taking the gradient of this log-likelihood with respect to $s(t)$, one obtains the gradient learning rule for the basis coefficients

$$\nabla_s L(s) = A^{\mathrm{T}}\Sigma^{-1}(x - As) + \nabla_s \varphi(s), \tag{4}$$

where ∇_s denotes the gradient with respect to s and the index t has been dropped for simplicity. Thus, the following update equation for the basis coefficients is obtained:

$$s^k = s^{k-1} + \eta \nabla_s L(s^{k-1}), \tag{5}$$

where η is the learning rate and s^{k-1} is obtained in the previous iteration. Because the Laplacian prior is unimodal, given the MAP coefficient estimate $\hat{s}$ inferred by the gradient learning rule in Eq. (4), the approximate conditional moments for the Gaussian posterior distribution, based on the Laplace approximation [8,3], are given as follows (see Appendix B):

$$E\{s\,|\,x(t)\} = \hat{s} = \arg\max_s L(s), \tag{6}$$

$$E\{ss^{\mathrm{T}}\,|\,x(t)\} = H(\hat{s})^{-1} + \hat{s}\hat{s}^{\mathrm{T}}, \tag{7}$$

where the Hessian of the approximate log-posterior computed at the MAP value $\hat{s}$ is denoted as $H(\hat{s}) = -\nabla_s \nabla_s L(\hat{s}) = A^{\mathrm{T}}\Sigma^{-1}A - \nabla_s \nabla_s \varphi(\hat{s})$. Note that $\nabla_{s_m} \varphi(s_m) = -\tanh(\beta s_m)$ and $\nabla_{s_m} \nabla_{s_m} \varphi(s_m) = -\beta\,\mathrm{sech}^2(\beta s_m)$ where β is a large positive constant in the case of a Laplacian prior on s_m $(m = 1, \ldots, M)$ (for details, see [8,3]). Thus, one has $\nabla_s \varphi(s) = (-\tanh(\beta s_1), \ldots, -\tanh(\beta s_M))^{\mathrm{T}}$ and $\nabla_s \nabla_s \varphi(s) = \mathrm{diag}(-\beta\,\mathrm{sech}^2(\beta s_1), \ldots, -\beta\,\mathrm{sech}^2(\beta s_M))$ in which $\mathrm{diag}(\cdot)$ represents a diagonal matrix. It is indicated in the next section that these conditional moments enable the development of an EM algorithm for estimating the parameter A in model (1).

3. An EM algorithm for the parameter estimation

To derive a learning algorithm for estimating the parameter in model (1), i.e., the matrix of basis vectors A, it is required to maximize the probability of the data generated according to model (1). For the T observed data samples $x = x(1), \ldots, x(T)$, one obtains the data likelihood

$$p(x\,|\,A) = \int p(x\,|\,s, A)\, p(s)\, \mathrm{d}s. \tag{8}$$

Rather than using some approximations to this intractable integral as in [8,9], it is desirous to employ the EM framework for estimation and inference for this form of

linear model, as this is a most natural method for maximizing the data likelihood. Given the approximate conditional moments of the posterior distribution, the standard form of M-step for the parameter A emerges (see Appendix C)

$$A^{\text{new}} = \left\{ \sum_{t=1}^{T} x(t) E\{s \mid x(t)\}^{\mathrm{T}} \right\} \left\{ \sum_{t=1}^{T} E\{ss^{\mathrm{T}} \mid x(t)\} \right\}^{-1}. \tag{9}$$

Inserting the approximate conditional moments (6) and(7) into Eq. (9) and noting that $(\Lambda(t)+A^{\mathrm{T}}\Sigma^{-1}A)^{-1} = \Lambda(t)^{-1} - \Lambda(t)^{-1}A^{\mathrm{T}}(\Sigma + A\Lambda(t)^{-1}A^{\mathrm{T}})^{-1}A\Lambda(t)^{-1}$, the following update for A is obtained:

$$A^{\text{new}} = \left\{ \sum_{t=1}^{T} x(t)\hat{s}(t)^{\mathrm{T}} \right\} \left\{ \sum_{t=1}^{T} \left(\Lambda(t)^{-1} - \Lambda(t)^{-1}A^{\mathrm{T}}M(t)A\Lambda(t)^{-1} + \hat{s}(t)\hat{s}(t)^{\mathrm{T}} \right) \right\}^{-1}, \tag{10}$$

where $M(t) = (\Sigma + A\Lambda(t)^{-1}A^{\mathrm{T}})^{-1}$ and $\Lambda(t) = -\nabla_s \nabla_s \varphi(s)$. Note that due to the famous convergence properties of the EM algorithm [2] each EM iteration increases the data likelihood or leaves it unchanged such that $p(x \mid A^{\text{new}}) \geqslant p(x \mid A^{\text{old}})$, where A^{old} is the parameter obtained in the previous iteration.

Since $\hat{s}(t)$ must be inferred in each M-step, a simple alternating variable method, which has already been used in similar estimation tasks [4], should be derived for inferring the most probable coefficients and learning the parameter A. The method is based first on the optimization of the objective function with respect to $s(t)$ for fixed A, then optimization with respect to A for fixed $s(t)$, and so on. The optimization with respect to $s(t)$ for fixed A is accomplished by the gradient learning rule in Eq. (4), and the optimization with respect to A for fixed $s(t)$ is accomplished by the M-step in Eq. (10). Hence, the EM procedure of the following form is obtained:

(i) Take some initial value for A^0. Set $s^0 = (A^0)^+ x$ where $(A^0)^+$ denotes the Moore–Penrose pseudo-inverse of A^0, and let $k = 1$. Normalize each column of A^0 to have unit norm.
(ii) Compute $s^k(t)$ by Eq. (5), using A^{k-1} as the estimate of A.
(iii) Compute A^k by Eq. (10), using A^{k-1} as the estimate of A and substituting $\hat{s}(t)$ by $s^k(t)$. Normalize each column of A^k to have unit norm.
(iv) Set $k = k + 1$, and go back to step (ii) if not converged.

It should be noted that the likelihood of linear model as given by Eq. (1) is a highly nonlinear function of the parameter values, and as such the likelihood will have many local optima. As was pointed out in [3] the convergence of the EM method to a local optimum of the likelihood is guaranteed and so will be dependent on the initial parameter values. Thus, there will be cases where the local optima may yield poor parameter estimates, and reinitialization of the algorithm will be required.

Table 1
SNR values in dBs: Gradient-based method, EM algorithm and variational EM algorithm

Speech sources	Gradient	EM	Variational
1	20	24	26
2	17	21	18
3	21	22	24

4. Simulation

One potential application of learning overcomplete representations is the blind source separation of more sources than mixtures as in [6]. To illustrate the algorithm proposed in this paper, the same three sources of natural speech and mixing matrix as used in [6] are employed in this experiment. The observed data is whitened before the iteration where whitening means that the covariance matrix of x is made equal to unity, i.e., $E\{xx^{\mathrm{T}}\} = I$, which is possible by a simple linear transformation [5]. The proposed EM procedure was randomly initialized and converged in 27 parameter updates with the basis coefficients being updated twice in each iteration. In this simulation, the learning rate η was set to be 0.0005. As expected, the data likelihood monotonically increased during the EM procedure in this simulation. The signal-to-noise ratio (SNR) was computed for each of the inferred sources and compared with the results reported in [6,3]. The results are shown in Table 1. The SNR values are on average improved over those reported in [6]. Fig. 1 shows 10,000 samples of the original sources, the noisy observations and the inferred sources after re-ordering and sign correction obtained by the EM procedure. This simulation serves to demonstrate the ability of the EM algorithm to learn overcomplete representations of natural data that confirm to the standard linear model and infer the most probable coefficients.

5. Conclusions

We have proposed an EM algorithm for learning the basis vectors and inferring the most probable basis coefficients in learning overcomplete representations. The development of an EM algorithm for estimation and inference is made possible by the approximate conditional moments of the posterior distribution derived by modelling the posterior as Gaussian. This EM algorithm can be considered a method for performing standard complete linear ICA with additive noise. This proposed EM algorithm may also be used in learning a sparse and overcomplete dictionary for observed signals and then exploiting this sparse representation for blind source separation as was proposed in [11]. The estimation of the noise covariance matrix using this EM framework will be considered in our future work.

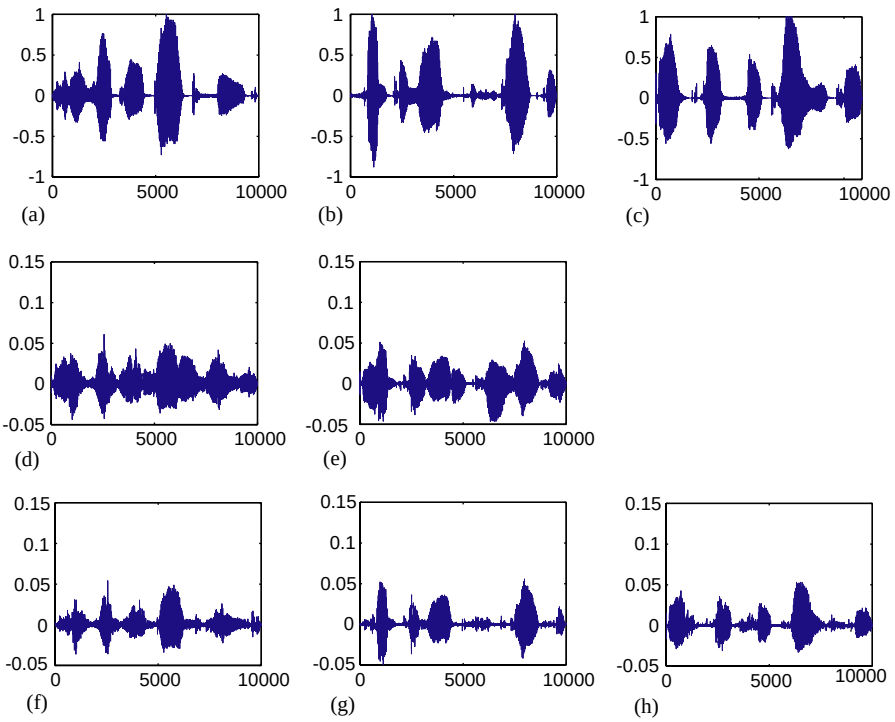

Fig. 1. (a–c)The three original sources. (d–e)The noisy observations. (f–h)The three inferred sources using the EM procedure outlined.

Acknowledgements

We are indebted to Dr. Jinghai Feng for his helpful comments on the manuscript. The authors would like to thank the editor Prof. R. Newcomb for his helpful suggestions. We are also grateful to the anonymous reviewers for their insightful and helpful comments. The work was supported by NSFC (30170321, 90103033), MOST (2001CCA00700) and MOE (KP0302).

Appendix A. Derivation of the log-likelihood

In this paper the data samples are assumed to be statistically independent, which has been used in many literatures [8,3,4]. Thus, the posterior of the basis coefficients has the following form:

$$p(s \mid x, A) = \prod_{t=1}^{T} p(s(t) \mid x(t), A). \tag{A.1}$$

For each data sample, Bayes' rule implies

$$p(s(t)\,|\,x(t),A)=\frac{p(x(t)\,|\,s(t),A)\,p(s(t))}{\int\,p(x(t)\,|\,s(t),A)\,p(s(t))\,\mathrm{d}s(t)}=\frac{p(x(t)\,|\,s(t),A)\,p(s(t))}{p(x(t))}. \tag{A.2}$$

Besides, according to model (1), one obtains

$$p(x(t)\,|\,s(t),A)=|\det(2\pi\Sigma)|^{-1/2}\times\exp\left\{-\tfrac{1}{2}(x(t)-As(t))^{\mathrm{T}}\Sigma^{-1}(x(t)-As(t))\right\}. \tag{A.3}$$

Inserting Eq. (A.3) into (A.2) and then Eq. (A.2) into (A.1), the log-likelihood in Eq. (3) is obtained by taking the logarithm of Eq. (A.1).

Appendix B. Derivation of the approximate conditional moments

Based on the Laplace estimation [8], the posterior distribution can be modelled as Gaussian:

$$p(s\,|\,x(t))\approx(2\pi)^{-M/2}\,|\,H(\hat{s})\,|^{1/2}\exp\left\{-\tfrac{1}{2}(s-\hat{s})^{\mathrm{T}}H(\hat{s})(s-\hat{s})\right\}, \tag{B.1}$$

where $\hat{s}$ is the MAP value and $H(\hat{s})$ is the Hessian of the approximate log-posterior computed at the MAP value. Thus, the approximate conditional moments in Eqs. (6) and (7) can be derived by this Gaussian posterior distribution.

Appendix C. Derivation of the M-step

To derive the M-step for the parameter A, it is required to maximize the expected value of the complete data likelihood of the following form, given the observed data x and the current model [2]

$$Q(A^{\mathrm{new}}\,|\,A^{\mathrm{old}})=\sum_{t=1}^{T}\int\,p(s\,|\,x(t),A^{\mathrm{old}})\log\{\,p(x(t),s\,|\,A^{\mathrm{new}})\}\,\mathrm{d}s, \tag{C.1}$$

where A^{old} is the parameter obtained in the previous iteration. Setting the gradient of $Q(A^{\mathrm{new}}\,|\,A^{\mathrm{old}})$ with respect to A^{new} to zero gives the new value of the parameter of the M-step in Eq. (9).

References

[1] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.

[2] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm(with discussion), J. R. Stat. Soc. B 39 (1) (1977) 1–38.

[3] M. Girolami, A variational method for learning sparse and overcomplete representations, Neural Comput. 13 (11) (2001) 2517–2532.

[4] A. Hyvärinen, Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood, Neurocomputing 22 (1–3) (1998) 49–67.
[5] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.
[6] T.W. Lee, M.S. Lewicki, M. Girolami, T.J. Sejnowski, Blind source separation of more sources than mixtures using overcomplete representations, IEEE Signal Process. Lett. 6 (4) (1999) 87–90.
[7] M.S. Lewicki, B.A. Olshausen, Probabilistic framework for the adaptation and comparison of image codes, J. Opt. Soc. Am.: Opt. Image Sci. Vision 16 (7) (1999) 1587–1601.
[8] M.S. Lewicki, T.J. Sejnowski, Learning overcomplete representations, Neural Comput. 12 (2) (2000) 337–365.
[9] R.M. Neal, Bayesian Learning for Neural Networks, in: Lecture Notes in Statistics, Vol. 118, Springer, New York, 1996.
[10] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res. 37 (23) (1997) 3311–3325.
[11] M. Zibulevsky, B.A. Pearlmutter, Blind source separation by sparse decomposition in a signal dictionary, Neural Comput. 13 (4) (2001) 863–882.