Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Non-negative matrix factorization algorithms modeling noise distributions within the exponential family

Vincent C. K. Cheung, and Matthew C. Tresch

***Abstract*—We developed non-negative factorization algorithms based on statistical distributions which are members of the exponential family, and using multiplicative update rules. We compared in detail the performance of algorithms derived using two particular exponential family distributions, assuming either constant variance noise (Gaussian) or signal dependent noise (gamma). These algorithms were compared on both simulated data sets and on muscle activation patterns collected from behaving animals. We found that on muscle activation patterns, which are expected to be corrupted by signal dependent noise, the factorizations identified by the algorithm assuming gamma distributed data were more robust than those identified by the algorithm assuming Gaussian distributed data.**

***Keywords*—matrix factorization, blind source separation, multiplicative update rule, signal dependent noise, EMG, muscle synergy**

I. Introduction

In many experimental contexts, investigators are faced with highly complex and high dimensional data. Often, this complexity is only apparent, and the data in fact exist in a simpler, low dimensional subspace of the total possible dimensionality. The problem for the investigator is to identify a representation of the data which captures this low dimensional subspace, thereby characterizing the critical features of the observed data.

Many methods have been proposed to perform this subspace identification, or matrix factorization. Here we describe a general framework for the factorization of data sets consisting of non-negative data. This work extends the framework developed by Lee and Seung [7] [8], to non-negative data sets generated according to any of the probability distributions belonging to the exponential family. This approach allows us to develop factorization algorithms which can be adapted to data sets with different expected statistical properties. We illustrate this approach using a factorization algorithm based on generalized gamma distributions, for which the standard deviation of a data point is proportional to its mean. Such signal dependent noise is expected for many physiological data sets, such as muscle activation patterns [6].

Manuscript received April 8, 2005. Supported by the Chyn Doug Shiah Memorial Fellowship and the Schoemaker Foundation Fellowship to V.C.K.C, and NIH-NINDS grants NS09343 and NS39865 to Emilio Bizzi.

Vincent C. K. Cheung is with the Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Cambridge, MA 02139 USA (corresponding author: 617-253-0771; e-mail: ckcheung@mit.edu).

Matthew. C. Tresch is with the Department of Biomedical Engineering, Northwestern University and Rehabilitation Institute of Chicago, Chicago, IL 60208 USA (e-mail: mtresch@mit.edu).

II. Non-negative algorithms modeling noise within the exponential family

Let D be a non-negative $M\times T$ data matrix comprising T samples of a M-dimensional data vector. We model the data matrix as a linear combination of N basis vectors, $N \le M$, such that

$$D \approx W \bullet C; \quad D,W,C \ge 0; \tag{1}$$

where each column of W ($M\times N$) represents a basis vector, and each row of C ($N\times T$) represents the coefficients for a corresponding basis across all data samples. We further assume that the observed D is generated by corrupting $W\bullet C$ with noise that can be characterized as a distribution in the exponential family. Our problem is to estimate W and C given D.

Assuming that the data samples and the data dimensions are both independently distributed, the likelihood function for D can be expressed as follows for any noise distribution in the exponential family:

$$P(D|W,C,\theta) = \prod_{i=1}^{M}\prod_{j=1}^{T} h(D_{ij})\, y([WC]_{ij}|\theta) \exp\left(\sum_{q=1}^{K} z_q([WC]_{ij}|\theta) \bullet t_q(D_{ij})\right), \tag{2}$$

where h, $y \ge 0$, t_q , z_q , h, y are real-valued functions, and K an integer, which together characterize a particular distribution in the exponential family. We set one of the parameters defining the distribution, e.g., one related to $E(D_{ij})$, to be $W\bullet C$, and let the rest of the parameters in the density function (denoted by θ) to be constants. Thus, for example, for the Gaussian distribution requiring two parameters (μ and σ, respectively representing the mean and standard deviation), μ can be set to $[W\bullet C]_{ij}$, and $\theta = \sigma$.

We estimate W and C through maximization of the log-likelihood function. To accomplish this, we first differentiate log $P(D|W,C)$ with respect to each component of W and C:

$$\frac{\partial}{\partial W_{ia}} \log P(D|W,C,\theta) = [\Phi C^T]_{ia} - [\Psi C^T]_{ia}; \tag{3}$$

$$\frac{\partial}{\partial C_{aj}} \log P(D|W,C,\theta) = [W^T \Phi]_{aj} - [W^T \Psi]_{aj};$$

$$\Phi_{ij} = \sum_{q=1}^{K} t_q(D_{ij}) \bullet z_q'([WC]_{ij}); \qquad \Psi_{ij} = \frac{-y'([WC]_{ij}|\theta)}{y([WC]_{ij}|\theta)}.$$

The log-likelihood function can then be maximized through a 2-step iterative gradient ascent procedure. In the first step, W is updated while keeping C fixed; in the second step, C is updated while keeping W fixed. Denoting the learning rates of the first and second steps by η^W and η^C, respectively, we obtain the following additive update rules:

0-7803-8740-6/05/$20.00 ©2005 IEEE.

$$W_{ia} \leftarrow W_{ia} + \eta^W \bullet \frac{\partial}{\partial W_{ia}} \log P(D \mid W, C, \theta), \quad \eta^W \geq 0; \tag{4}$$
$$C_{aj} \leftarrow C_{aj} + \eta^C \bullet \frac{\partial}{\partial C_{aj}} \log P(D \mid W, C, \theta), \quad \eta^C \geq 0.$$

Since W and C are assumed to be non-negative, both of the above learning rates can be formulated component-wise as functions of W and C. Let $\eta_{ia}^W = W_{ia} / [\Psi C^T]_{ia}$, and $\eta_{aj}^C = C_{aj} / [W^T \Psi]_{aj}$. After some straight-forward rearranging using the expressions in (3), the additive update rules in (3) can be reformulated as multiplicative update rules:

$$W_{ia} \leftarrow W_{ia} \bullet \frac{[\Phi C^T]_{ia}}{[\Psi C^T]_{ia}}; \quad C_{aj} \leftarrow C_{aj} \bullet \frac{[W^T \Phi]_{aj}}{[W^T \Psi]_{aj}}, \tag{5}$$

with the Φ and Ψ matrices as defined in (3).

In this paper, we focus on two non-negative algorithms, derived using (5) from the Gaussian and the generalized gamma distributions, respectively. In the Gaussian case, if we set $\mu = WC$, then $\Phi = D / \sigma^2$, and $\Psi = [WC] / \sigma^2$. Substituting these expressions for Φ and Ψ into (5), we obtain the update rules for the Gaussian algorithm (GAU), which are the same as the non-negative matrix factorization update rules proposed by Lee and Seung [7] [8].

To model signal-dependent noise, we use the generalized gamma distribution having the following form:

$$f(D_{ij} \mid \alpha, \kappa) = \left(\frac{D_{ij} \alpha}{\kappa} \right)^{\alpha} \frac{e^{-D_{ij}(\alpha/\kappa)}}{D_{ij} \Gamma(\alpha)}, \tag{6}$$

where α and κ are parameters for the distribution, and $\Gamma(\alpha)$ is the gamma function (note that (6) can be re-expressed as a standard gamma distribution with $\beta = \kappa/\alpha$). If we set $\kappa = [WC]_{ij}$, and define $\phi = 1/\sqrt{\alpha}$, we obtain the following relationship between the mean and standard deviation of each data point:

$$E(D_{ij}) = [WC]_{ij}; \quad Var(D_{ij}) = [WC]_{ij}^2 / \alpha = \left(\phi \bullet E(D_{ij}) \right)^2. \tag{7}$$

Thus, the generalized gamma distribution defines a signal-dependent noise model in which the standard deviation of the noise is proportional to the data amplitude. Such signal-dependent noise has been observed to underlie variation of control signals in the motor system [6]. Update rules for the generalized gamma algorithm (GGM) can be obtained by deriving expressions for Φ and Ψ:

$$\Phi_{ij} = \frac{\alpha D_{ij}}{[WC]_{ij}^2}; \quad \Psi_{ij} = \frac{\alpha}{[WC]_{ij}}; \quad \alpha, W, C > 0; \tag{8}$$

and the noise proportionality constant ϕ in (7) can be obtained by estimating α using standard maximum likelihood methods.

III. Assessing performance of the GAU and GGM algorithms with simulated data sets.

We assessed the ability of the GAU and GGM algorithms described above to identify underlying basis vectors from simulated data sets generated by known bases. Ten different simulated data sets were initially generated. Each contained 1000 data samples, and was generated by linearly combining a set of 5 basis vectors. The components of both the basis vector matrix (W) and the coefficient matrix (C) used for data generation were uniformly distributed in (0, 1). Each of these 10 data sets was then corrupted by two different types of noise – Gaussian noise, and signal-dependent noise with a generalized gamma distribution. In the case of Gaussian noise, the data set, D, was corrupted so that

$$D = g(\overline{D}); \quad \overline{D} \sim N(W \bullet C, \sigma^2),$$

where $g(x) = x$ if $x \geq 0$, and $g(x) = 0$ if $x < 0$. Such thresholding of the data was necessary to ensure that the simulated data set stayed non-negative. Each simulated data set was corrupted by Gaussian noise of 8 different magnitudes, corresponding to the following values of σ: (0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3). In the case of signal-dependent noise, each of the 10 data sets was corrupted so that

$$D \sim Gamma \quad (\alpha, W \bullet C / \alpha).$$

As in the case of Gaussian noise, each of the 10 original simulated data set was corrupted by noise of 8 different magnitudes using the following values of α: (20, 40, 60, 80, 100, 120, 150, 300). Thus, there was a total of 10×2×8 = 160 data sets, 80 of which contained Gaussian noise, and the other 80, signal-dependent noise.

For each type of noise, and for each noise level, we quantified noise magnitude by calculating 1-R^2, where R^2 is the coefficient of determination representing the percentage of variance in the noise-corrupted data set explicable by the original uncorrupted data set.

We then proceeded to extract 5 basis vectors from each of these 160 data sets using both the GAU and GGM algorithms. For each extraction, we also randomly and independently shuffled each row of the data matrix, and extracted basis vectors from this shuffled data set. This allowed us to assess the baseline similarity between the extracted and original basis vectors (see below).

Performance of the two algorithms was assessed by comparing the extracted bases and the original bases generating the simulated data set. Similarity was quantified by calculating the sum of the cosine of the principal angles between the two sets of bases [5]. Since 5 bases were identified in each extraction, the maximum similarity value was 5. To account for the possibility that a fraction of the calculated similarity value is expected by chance, for each data set, each noise magnitude, and each algorithm, the normalized similarity (s^{norm}) was calculated as follows:

$$s^{norm} = (s - s^{baseline}) / (5 - s^{baseline}),$$

where s is the similarity value between the original bases and the bases extracted from unshuffled data, and $s^{baseline}$ is the similarity value between the original bases and those extracted from shuffled data.

Figure 1 summarizes the performance of the two algorithms in the two types of noise-corrupted data sets. Results from both the GAU algorithm (solid line) and the GGM algorithm (dotted line) are presented (mean ± SD, n =10). It is apparent from Fig. 1A that for data sets with Gaussian noise, when the noise magnitude was low, both algorithms performed equally well (with s^{norm} close to 1). But for Gaussian noise magnitudes >20%, the GAU algorithm performed substantially better than the GGM

algorithm. On the other hand, as shown in Fig. 1B, for data sets containing signal-dependent noise, the GGM algorithm performed better than the GAU algorithm for all noise magnitudes tested. Thus, the non-negative algorithm is better able to recover the original generative basis vectors if the actual noise structure of the data set agrees with the noise structure assumed by the algorithm.

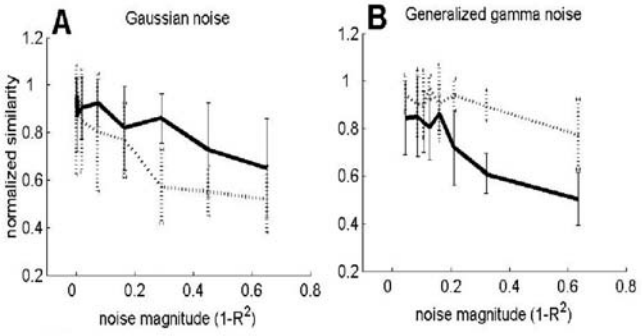

Fig. 1. Assessing performance of the GAU and GGM algorithms with simulated data sets. Basis vectors extracted using the GAU algorithm (solid line) and GGM algorithm (dotted line) were compared with the original basis vectors generating the data sets. Data sets were corrupted by either constant variance Gaussian noise (*A*) or signal-dependent noise (*B*). Mean ± SD plotted (n=10).

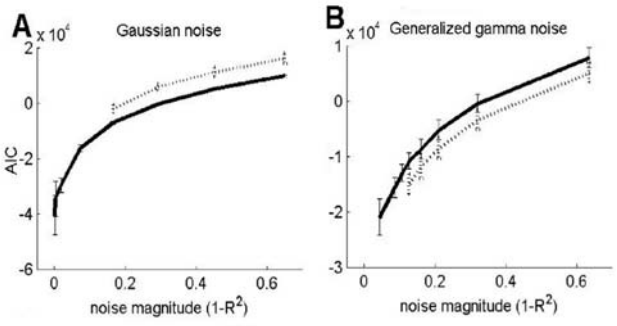

Fig. 2. Testing the performance of the Akaike Information Criterion. The AIC values were computed using results obtained from both the GAU (solid line) and GGM (dotted line) algorithms. The simulated data sets were corrupted by either constant variance Gaussian noise (*A*) or signal-dependent noise (*B*). Note that for data sets with low noise magnitudes, we were unable to obtain the AIC values for the signal-dependent noise model, because the `gamma` function in Matlab cannot handle large input arguments. Mean ± SD plotted (n=10).

IV. Testing the Performance of an Objective Model Selection Criterion – the AIC

As shown above, the closer the data noise structure is to the noise structure assumed by the algorithm, the better the algorithm performs in retrieving the correct basis vectors. However, the noise structure of any given physiological data set is often not known. Whether results of any one algorithm are potentially "better" or "worse" than those of another in capturing underlying data structures has to be determined by some model selection criteria. Here, we tested whether the well-known Akaike Information Criterion (AIC) [1] could reveal the underlying data noise structure in the simulated data sets described above: the lower the AIC, the better the fit of the parameters to the observed data.

For each of the 160 simulated data sets described in the previous section, the AIC was computed for both the GAU and GGM algorithms using their respective extracted bases and coefficients. Figure 2 shows the AIC values obtained from the two algorithms (GAU: solid line; GGM: dotted line; mean ± SD, n =10). It is apparent that for data sets corrupted with constant-variance Gaussian noise, the GAU algorithm consistently yielded results with lower AIC values, indicating that the Gaussian noise model of the GAU algorithm was a better fit to the data than the signal-dependent noise model. On the other hand, for data sets corrupted with signal-dependent noise, the AIC values from the GGM results were lower than those from the GAU results for all noise magnitudes tested. Hence, the underlying noise structure of the data set could be revealed by comparing the AIC values obtained through different algorithms having different assumptions of noise structure.

V. Applying the Algorithms to EMG Data

We next proceeded to analyze electromyographical (EMG) data sets collected from 13 hindlimb muscles of the frog during jumping in order to identify the basis vectors, or muscle synergies, underlying these high-dimensional data sets. Previous studies [2] [3] [9] have suggested that the frog motor system might simplify control of the many degrees of freedom in the muscle space through linear combination of a small number of muscle synergies. However, none of the above-mentioned studies compared performance of different algorithms assuming different noise structures. It is not known also whether frog EMGs collected during natural behaviors are corrupted by signal-dependent noise similar to the kind suggested by Harris and Wolpert [6]. Here, we address this question by applying both the GAU and GGM algorithms to frog EMG data sets, and by comparing AIC values from both algorithms to see which noise model might be a better fit to these physiological data.

Methods for data collection have been described previously [4]. Data collected from 6 different intact frogs during jumping were analyzed. For each frog, the GAU and GGM algorithms were applied to the EMG data to extract 5 synergies. Such a model order was suggested by previous studies [2] [3]. Also, synergy extraction was repeated 5 times for each frog and each algorithm, each time with different initializing W (the synergy matrix) and C (the coefficient matrix). The AIC value for each extraction was also computed.

Table 1 lists the AIC values for both the GAU and GGM models applied to each of the 6 frogs. For all frogs, the AIC of GGM was smaller than the AIC of GAU. Such a consistent result suggests that the noise underlying frog EMGs might be better described using a signal-dependent noise model whose noise standard deviation is proportional to the mean EMG amplitude. Another implication is that the muscle synergies extracted using GGM might be closer to the actual physiological synergies underlying jumping than the GAU synergies.

Table 1. The Akaike Information Criterion (AIC) values for the GAU and GGM models applied to frog jumping EMG data (13 muscles, 5 synergies; mean, n=5). The MLE estimate for the ϕ parameter in the GGM noise model (eqn. 7) is also listed (mean ± SD, n=5).

Animal	GAU (x10^6)	GGM (x10^6)	MLE est. for ϕ (GGM)
Frog 1	-1.9236	-3.8506	0.5002±0.0117
Frog 2	-0.5617	-1.2010	0.4529±0.0284
Frog 3	-0.5095	-1.3870	0.5275±0.0242
Frog 4	-0.6505	-1.3020	0.4688±0.0112
Frog 5	-0.6310	-1.2783	0.4606±0.0115
Frog 6	-0.3357	-0.7764	0.5162±0.0051

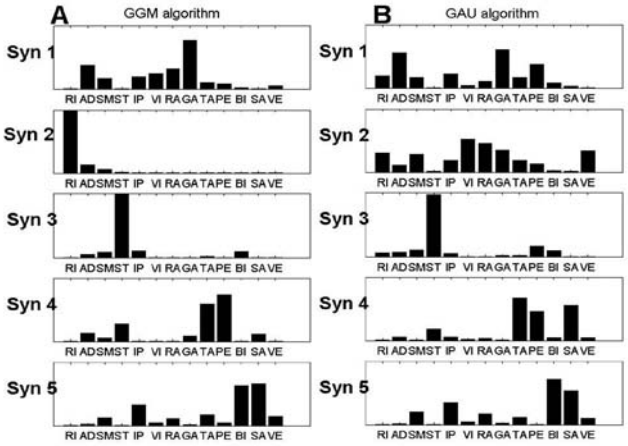

Fig. 3. Muscle synergies (Syn) of frog 2 extracted using the GGM (*A*) and the GAU (*B*) algorithms. Results from the extraction repetition with the lowest AIC were shown. The 13 recorded muscles were rectus internus major (RI), adductor magnus (AD), semimembranosus (SM), semitendinosus (ST), ilio-psoas (IP), vastus internus (VI), rectus femoris anticus (RA), gastrocnemius (GA), tibialis anticus (TA), peroneus (PE), biceps (or ilio-fibularis) (BI), sartorius (SA), and vastus externus (VE). Synergies from the two synergy sets were matched by calculating scalar products.

Figure 3 shows the jumping muscle synergies extracted using GGM (*A*) and GAU (*B*) for one frog. In this figure, it can be seen that both algorithms yielded similar results for synergies 3-5, all of which were activated during the flexion phase of the jump. However, for synergies 1-2, both of which were activated during jump extension, the GGM and GAU synergies are quite different from each other. A closer examination of these two synergies reveals, for example, that both synergies 1 and 2 from GAU have a strong gastrocnemius (GA) component; however, only synergy 1 from GGM contains GA activation. An interpretation of this observation is that because GA was activated strongly during extension, its EMG was also more variable due to signal-dependent noise. Such variability forced the GAU algorithm to divide the GA extension bursts up between two different synergies. Whether the GAU or GGM synergies correspond better to other physiological measures (such as kinematics) would require further studies and analyses.

VI. Discussion and Conclusion

In this paper, we have derived a non-negative blind source separation algorithm capable of modeling any noise distribution in the exponential family (eqn. 5). It is shown that our update rules can be reduced to the non-negative matrix factorization algorithm proposed by Lee and Seung [7] [8] by assuming constant-variance Gaussian noise. Also, update rules for a signal-dependent noise model were derived (eqn. 6). The ability of both the Gaussian (GAU) and generalized gamma (GGM) algorithms to recover bases from data corrupted with noise was tested using simulated data sets, and it was confirmed that the algorithm performs better if the data noise structure agrees with the noise model of the algorithm (Fig. 1). It was further shown that the AIC could be used as a model selection criterion to decide which algorithm might be better fitted to the data (Fig. 2). Both the GAU and GGM algorithms were then applied to EMG data collected from 6 frogs. Calculation and comparison of AIC values suggested that frog EMG data might be better described using a signal-dependent noise model.

One of the most important features of the algorithm presented here is that it is generalized to any noise distribution in the exponential family, thus allowing modeling of signal-dependent noise. In this paper we have presented one possible formulation of signal-dependent noise using the generalized gamma distribution. But other formulations, such as *Gamma*(*WC*, 1) or *N*(*WC*, *WC*), are in principle possible, provided that the resulting algorithms converge. Such flexibility of our algorithm would be particularly useful in analyses of data sets with unknown noise structures. In such cases, basis vectors can be extracted using multiple versions of our algorithm having different noise model assumptions, and a model selection criterion such as AIC can then be applied to gain insight into the underlying data noise structure. For this reason, we think the algorithms presented here might be useful in analyses of not only EMG signals, but also a wide variety of high-dimensional physiological data (such as neuronal firing rates of multiple neurons).

References

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (Petrov B.N., and Csaki F., Eds). Budapest: Academiai Kiado, pp. 267-281, 1973.

[2] Cheung V.C.K., d'Avella A., Tresch M.C., and Bizzi E. Central and sensory contributions to the activation and organization of muscle synergies during natural motor behaviors. *J. Neurosci.* 25(27): 6419-6434, 2005.

[3] d'Avella A., and Bizzi E. Shared and specific muscle synergies in natural motor behaviors. *Proc. Natl. Acad. Sci. USA* 102(8): 3076-3081, 2005.

[4] d'Avella A., Saltiel P., and Bizzi E. Combinations of muscle synergies in the construction of a natural motor behavior. *Nat. Neurosci.* 6(3): 300-308, 2003.

[5] Golub G.H., and Van Loan C.F. *Matrix computations*. Baltimore, MD: Johns Hopkins UP, 1983.

[6] Harris C.M., and Wolpert D.M. Signal-dependent noise determines motor planning. *Nature* 394: 780-784, 1998.

[7] Lee D.D., and Seung H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788-791, 1999.

[8] Lee D.D., and Seung H.S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems, vol. 13* (Leen T.K., Dietterich T.G., and Tresp V., Eds). Cambridge, MA: MIT Press, pp. 556-562, 2001.

[9] Tresch M.C., Saltiel P., and Bizzi E. The construction of movement by the spinal cord. *Nat. Neurosci.* 2(2): 162-167, 1999.