THE MULTILINEAR ICA DECOMPOSITION WITH APPLICATIONS TO NSS MODELING

Raghu G. Raj and Alan C. Bovik

Center for Perceptual Systems, The University of Texas at Austin, Austin, Texas 78712, USA

ABSTRACT

We refine the classical Independent Component Analysis (ICA) decomposition using a multilinear expansion of the probability density function of the source statistics. In particular, to model the source statistics of natural image textures, we introduce a specific non-linear system that allows us to elegantly capture the statistical dependences between the responses of the Multilinear ICA (MICA) filters. The resulting multilinear probability density is analytically tractable and does not require Monte Carlo simulations to estimate the model parameters. We demonstrate the success of the MICA model on natural textures and discuss applications to non-stationarity detection and natural scene statistics (NSS) modeling.

Index Terms— Multilinear ICA, Non-linear Modeling, Textures, Natural Scene Statistics (NSS).

1. INTRODUCTION

The construction of accurate prior models of source data is essential to many applications, such as low-level vision, for which unsupervised learning methods must be applied due to the inherent lack of labeled training sets. Such prior models give a framework in which to correctly interpret the data thereby serving as the basis for subsequent analysis viewed from different levels of abstraction. The classical statistical tools that exist for this purpose include Principle Component Analysis (PCA), Independent Component Analysis (ICA), Multidimensional Scaling (MDS) etc. [1].

In this paper we develop a novel refinement of the classical ICA decomposition that involves a multilinear expansion of the probability density function of the source. Denote the probability of the source that we are modeling by $P(X)$, where X is a random variable whose realizations have dimensionality d. The goal of ICA is to factor the probability density of the source into a product of distributions: $P(X) = \prod_{i=1}^{d} P(s_i)$, where $s_i = X * \phi_i$ corresponds to a filtered response of the source. The filters $\{\phi_i\}_{i=1}^{d}$ are the ICA filters of the source. Statistical algorithms for computing the ICA filters have been the subject of intense study over the past decade [2], most of which involve the construction of different cost functions (usually variations or special cases of the maximum likelihood based cost function).

Supported by NSF grant 0427372.

Though ICA enables us to compactly represent $P(X)$ and to fruitfully utilize the statistically independent directions of the data in many applications [2], unfortunately the statistics of many real-world sources, such as natural image patches, cannot be factored into such a simple form–as a result the so-called independent components can contain significant dependences between them [3].

In this paper we explicitly capture such statistical dependencies by means of a multilinear representation of $P(X)$: $P(X) = \frac{1}{Z} g(J) \prod_{i=1}^{d} P(s_i)$, where $\{s_i = X * \phi_i\}_{i=1}^{d}$, and such that $g : J = [s_1, \cdots, s_d]^T \rightarrow R$, where $Z \in R$ is a normalizing constant. Of all possible multilinear expansions of this form that could describe the source distribution, we are interested in the one that makes the representation of the source as sparse as possible, i.e., which minimizes the contribution of $g(J)$–in particular, we are interested in obtaining closed form approximations for such a $g(J)$. Such a multilinear form retains all the attractive properties of the ICA decomposition while at the same time lumps the interactions of the filtered responses into the function $g(J)$. Clearly when $g(J)$ is separable with respect to the filter responses (or identity), then this reduces to the classical ICA representation.

The success of the method of course depends upon the verity of our numerical approximation of $g(J)$. Analytical methods of approximating $g(J)$ in terms of Taylor expansions seem formidable. Further complicating the matter is the fact that one has to estimate Z which, in general, requires tedious Monte Carlo simulations.

In Section 2, for the purpose of modeling natural scenes patches, we introduce a non-linear system model that enables us to circumvent the above issues in furnishing a multilinear expansion of $P(X)$. We call the resulting refinement of ICA the *Multilinear ICA (MICA) Model*. We successfully deploy the new method to model natural scene textures in Section 3 and demonstrate advantages relative to classical ICA. Extensions to mixtures of MICA models for modeling NSS are discussed in Section 4 along with applications to non-stationarity detection.

2. MULTILINEAR ICA MODEL

We begin by considering the classical ICA model wherein the observation $\tilde{s}$ is modeled as follows: $\tilde{s} = Bz$; where $\tilde{s} = [\tilde{s_1}, \cdots, \tilde{s_d}]^T \in R^d$, $z = [z_1, \cdots, z_d]^T \in R^d$, d is assumed to

be the intrinsic dimensionality of the data, and $B \in R^{d\times d}$ is a full-rank matrix. The goal of ICA is to find a matrix B such that the resulting components of z are independent random variables.

However, as mentioned in the previous section, for many real-world sources such as natural image patches, such an ideal decomposition is not possible and so the components of z will contain residual dependencies. Our aim is to explicitly capture these dependencies. In doing so we must first recognize that z cannot be further decomposed as a combination of independent sources via another full-rank matrix! It is possible, however, that z can be decomposed with respect to an under-complete linear model but this requires knowledge of the subspace dimensionality.

An alternate view which we espouse in this paper is that, given the knowledge of the intrinsic dimensionality d, the residual dependencies can be captured via non-linear combinations of independent sources. The choice of the non-linearity, as well as of the source distribution, must be as simple as possible, and yet must successfully account for the probabilistic structure of the observed natural image patches. To simplify matters further, we first concern ourselves only with modeling *unimodal distributions*; later on we suggest how to extend this to multimodal cases via mixtures of MICA models. Unimodal distributions are applicable to most natural scene textures of which we give some examples in Section3.

Perhaps the simplest possible non-linear system that one can hypothesize for this purpose is a quadratic channel. From our experiments with natural scene textures, we found that the hybrid linear-quadratic model (stimulated by *i.i.d.* Gaussian sources) shown in Fig. 1 does indeed successfully account for the probabilistic structure of natural image patches. We now describe this non-linear system in detail.

The observable source data that we are modeling is $\tilde{s} \in R^d$. $B \in R^{d\times d}$ is a full-rank matrix as mentioned above. B is initially chosen to be the matrix associated with the classical ICA decomposition of $\tilde{s}$ and, as explained in more detail later, can be re-estimated in subsequent iterations. The system F shown in Fig. 1 models the residual interaction between the components of $z \in R^d$. It consists of a core non-linearity φ preceded by a linear system $y = As + \gamma$, where $y = [y_1, \cdots, y_d]^T \in R^d$, $\gamma = [\gamma_1, \cdots, \gamma_d]^T \in R^d$, and $s = [s_1, \cdots, s_d]^T \in R^d$ consists of *i.i.d.* Gaussian sources such that $s_i \tilde{} N(0,1)$. The variance of the Gaussian channels is determined by $\sigma = [\sigma_1, \cdots, \sigma_d]^T \in R^d$ as shown in Fig. 1. Let the Gaussian density of the i^{th} channel be denoted by $q(s_i)$. The parameter $\beta \in R$ is a scalar that is applied to all channels; and finally, let $C = [C_{i,j}] = [C_1^T, \cdots, C_d^T]^T = A^{-1} \in R^{d\times d}$.

The core non-linearity φ consists of complementary linear and quadratic channels. Operators $u_{1,\alpha}$ and $u_{2,\alpha}$ are *complementary limiters* with respect to parameter α such that:

$u_{1,\alpha}(y_i) + u_{2,\alpha}(y_i) = 1$

$u_{1,\alpha}(y_i), u_{2,\alpha}(y_i) \geq 0, \;\; for \;\; all \;\; 1 \leq i \leq d$

A simple choice of these limiters, which we have found to be useful for modeling natural image textures (as shown in Section 3), is the complementary step functions of the following form: $u_{1,\alpha}(y_i) = u(y_i + \alpha) - u(y_i - \alpha)$, $u_{2,\alpha}(y_i) = 1 - u_{1,\alpha}(y_i)$, where $u(x)$ is the unit step function, and $\alpha = 1$.

Thus we have that $\varphi(y) = yu_{1,\alpha}(y) + \varphi_q(y)u_{2,\alpha}(y)$, where $y = As + \gamma \in R^d$, $\varphi_q(y) = y^2 sgn(y)$ (where all operations on y are applied component-wise). For this choice of $(u_{1,\alpha}, u_{2,\alpha})$, it follows that: $\tilde{\varphi}(\beta z) \equiv \varphi^{-1}(w) = \beta z u_{1,\alpha}(\beta z) + \varphi_q^{-1}(\beta z)u_{2,\alpha}(\beta z)$, where $\varphi_q^{-1}(\beta z) = \{\sqrt{|\beta z_i|}\, sgn(\beta z_i)\}_{i=1}^d$. Since the non-linearity is invertible, system F is also invertible: $s = F^{-1}(z) = C[\tilde{\varphi}(\beta z) - \gamma]$. Thus we have that: $s_k = F_k^{-1}(z) = C_k^T[\tilde{\varphi}(\beta z) - \gamma]$.

Given the above, the distribution of z has the following form: $p(z) = \frac{1}{|J(F)|}\prod_{k=1}^d q(F_k^{-1}(\beta z))$, where $q(s_k)$ is the k^{th} source channel. Expanding the above terms we obtain:

$$p(z) = \frac{K}{|J(F)|} g(J) \prod_{k=1}^{d} p(z_i\beta) \qquad (1)$$

where,

$p(z_i) = exp(-\frac{1}{2}(\sum_{k=1}^d \frac{C_{k,i}^2}{\sigma_k^2})\tilde{\varphi}^2(z_i) - (\sum_{k=1}^d \frac{C_k^T\gamma}{\sigma_k^2} C_{k,i})\tilde{\varphi}(z_i))$

$g(J) = exp(-\sum_{i\neq j} G_{i,j}\tilde{\varphi}(z_i)\tilde{\varphi}(z_j))$, $K = \frac{exp(-\sum_{k=1}^d \frac{(C_k^T\gamma)^2}{2\sigma_k^2})}{(2\pi)^{d/2}\sigma_1 \cdot\cdot \sigma_d}$

and $G_{i,j} = -\sum_{k=1}^d \frac{C_{k,i}C_{k,j}}{\sigma_k^2}$. $J(F)$ is the Jacobian associated with the transformation F. The following lemma (proved in Appendix) yields a closed form expression of the Jacobian.

Lemma 1: The Jacobian $J(F)$ for the above transformation is given by: $J(F) = \frac{1}{|C|\beta^d}\prod_{k=1}^d \psi(\beta z_k)$, where $\psi(\beta z_k) = u_1(\beta z_k) + 2|\varphi_q^{-1}(\beta z_k)|u_2(\beta z_k)$. ♣

Given the structure of the MICA distribution in (1), the qualitative roles of each of the parameters described above are evident. We make special mention of matrix $[G_{i,j}]$ and parameters (β, σ). The *MICA interaction matrix* $[G_{i,j}]$ captures the interaction among the MICA components. In particular, when $[G_{i,j}]_{i\neq j} = 0$, the MICA components are independent. The parameter β determines the amount of nonlinearity in the system which can be qualitatively understood as follows: when training the MICA model (given the data z), β determines the extent to which w is scaled inside the unit interval and consequently determines (i.e. after σ is adjusted as a part of the MICA optimization) the extent to which the linear channel of the system is active.

From the MICA model in (1), closed form expressions of the gradients of the MICA log-likelihood function with respect to all the parameters (i.e. C, σ, γ, β and B) can be derived in a straightforward manner. These gradients can then be used to compute the optimal parameters of the MICA model. As described in Section 1, our goal is to obtain a multilinear expansion of $P(X)$ that corresponds to a sparse representation of the source. We accomplish this by initializing B with the matrix associated with the classical ICA decomposition of source $\tilde{s}$. Thereby we obtain $z = B\tilde{s}$. Thereafter we can invoke the optimization algorithm to compute the param-

Texture	$\sum_{i\neq j} \lvert G_{i,j}^{indep}\rvert$ for Indep. Channels	$\sum_{i\neq j} \lvert G_{i,j}\rvert$ using Actual Data
Sand	0.0269	0.2045
Gravel	27.8389	80.1025

Table 1. Interchannel interaction, which is measured by $\sum_{i\neq j} |G_{i,j}|$, decreases as expected when using truly independently sampled channels

eters $C = A^{-1}$, σ, γ, and β. In the simulations (Section 3) all parameters except for β (which is hand-picked) are computed automatically by the optimization algorithm (although β can also potentially be automated using the closed form gradient descent equation derived from the MICA distribution (1)). At this stage one obtains a multilinear expansion of $P(X)$ as given in (1), the structure of which is fully specified by these computed parameters. One can further refine the estimate of B by invoking a gradient descent algorithm with respect to B (while fixing the parameters $(C, \sigma, \gamma, \beta)$ to the the values obtained in the previous step). Once the matrix B is computed, the above 2-step process of re-estimating (C, σ, γ), followed by re-estimation of B can be performed until the desired level of accuracy is achieved. In the simulations below we find that even just the estimation of (C, σ, γ), *without* subsequent re-estimation of B, can outperform classical ICA based modeling of natural image textures.

3. SIMULATION RESULTS

We applied the parameter estimation algorithm described in the previous section to model natural image textures [4]. The texture images were sampled uniformly at random locations with 3x3 patches (i.e. $d = 9$). An ICA was then performed on the data vectors obtained from each texture using Comon's algorithm [5] to obtain the B matrix described above. Subsequently the parameters (C, σ, γ) of the MICA model were estimated as described in Section 2 (for the hand-picked values of β chosen for each texture as given below). For each texture, we compare the actual data distributions of each channel to the corresponding distribution predicted by the ICA and MICA models. In addition, the average of all the data channels is also compared with those predicted by the ICA and MICA models.

Simulation of the MICA model is accomplished by generating $d = 9$ *i.i.d.* zero mean, unit variance Gaussian channels as shown in Fig. 1 and plotting the histograms outputs of the channels (or filtered outputs thereof) when the optimal parameters computed (for the texture being modeled) are used. The ICA model is simulated by first computing the empirical distribution of each channel. Once the empirical distributions of the ICA-channels are obtained, they are independently sampled and then processed the B matrix. The histograms of the resulting outputs are then compared to the corresponding data distributions and the predictions of the MICA model.

Figs. 2(a) and 2(d) show the histograms of two of the channels for the Sand texture. Figs. 2(b) and 2(e) show the corresponding ICA distributions; and Figs. 2(c) and 2(f) show the corresponding MICA distributions. Fig. 2(g) shows the histogram of the data distribution when all the data channels for the sand texture are averaged together. Fig. 2(h) and Fig. 2(i) show the corresponding distributions obtained from the ICA and MICA models respectively. For the sand texture, $\beta = 1$ was chosen to obtain the MICA results. The distributions have been mean shifted about zero and scaled to fit the same interval. All these results show the MICA distributions to approximate the original data distribution more closely than the corresponding ICA distributions. Similar results are obtained for the Gravel texture in Fig. 3 (for gravel texture $\beta = 0.0033$ was chosen to obtain the MICA results).

We remark that since MICA interaction matrix $[G_{i,j}]_{i\neq j}$ captures the interaction among the multilinear ICA filters (Section 2), $\sum_{i\neq j} |G_{i,j}|$ can give us a measure of the inter-channel interaction. Consequently, if instead of training the MICA model on $\{z_k\}_{k=1}^d$, we instead train it on channels that are truly independently sampled from the empirical ICA distributions corresponding to each channel, we expect the corresponding matrix $[G_{i,j}^{indep}]_{i\neq j}$ to be such that $\sum_{i\neq j} |G_{i,j}| \leq \sum_{i\neq j} |G_{i,j}^{indep}|$. This is indeed what we observe in Table 1, which reinforces the fact that the MICA model captures interchannel interaction which thus enables it to better account for the true data distribution as compared to its ICA counterpart.

4. DISCUSSION

As an application of the concepts discussed in the preceding sections, consider the problem of determining whether a given image segment is non-stationary, i.e., whether it contains different types of textures. The following lemma furnishes a non-stationarity index that measures the degree of non-stationarity of a given image patch.

Lemma 2: Let T be a non-stationary image patch that consists of 2 non-overlapping textures that partition T; one which we call the *center patch* and the other the *surround patch*. Let $\{\phi_i^c\}_{i=1}^d$ be the MICA filters of the center patch. Let I^c and I^s be random variables that correspond, respectively, to *d*-dimensional samples of the center and surround patches. Let $J_c = [I_1^c, \cdots, I_d^c]$ and $J_s = [I_1^s, \cdots, I_d^s]$, where $I_k^c =< I^c, \phi_k^c >$, $I_k^s =< I^s, \phi_k^c >$. Further let μ^c and μ^s denote the joint probability densities associated with J_c and J_s respectively. Then we have that:

$\eta = \frac{\sum_{m\neq n} E_{\mu^s}[<\varphi(I_m^s),\varphi(I_n^s)>]G_{m,n}^s + C_1^s + C_2^s}{\sum_{m\neq n} E_{\mu^c}[<\varphi(I_m^c),\varphi(I_n^c)>]G_{m,n}^c + C_1^c + C_2^c} > 1$, such that $G_{m,n}^c$ and $G_{m,n}^s$ are the MICA interaction matrices corresponding to J_c and J_s respectively, where $C_1^c = D(\mu^c(J_c) \; || g_c(J_c) \prod_{k=1}^d p_c(I_k^c))$ and $C_2^c = E_{\mu^c}[ln(\prod_{k=1}^d \frac{p_c(I_k^c)}{\mu_m^c(I_k^c)})]$ (and where similar expressions hold for C_1^s and C_2^s); given that

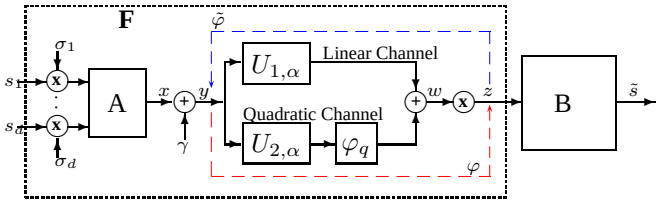

Fig. 1. Non-linear system model that accounts for the multilinear structure of source statistics derived from natural scene data.

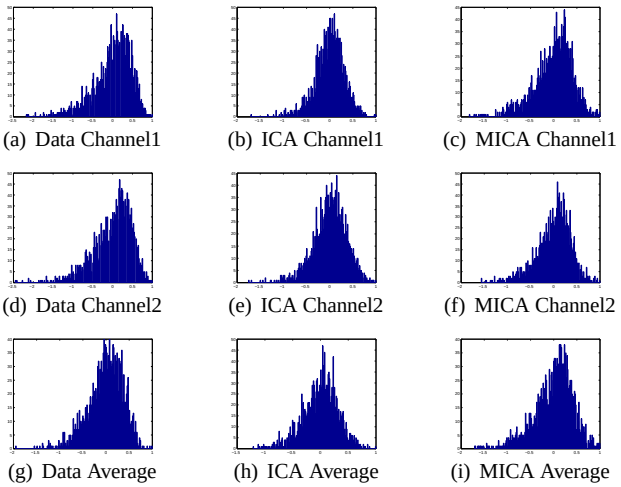

Fig. 2. Sand Texture: Histograms of the Data channels and the corresponding ICA and MICA distributions.

$D(p||q) = \int pln(\frac{p}{q})$ is the Kullback-Leibler divergence, μ_m^c and μ_m^s are the marginals associated with the ICA approximations of J_c and J_s respectively, and $q(J_c) = g_c(J_c).\prod_{k=1}^{d} p_c(I_k^c)$ and $q(J_s) = g_s(J_s).\prod_{k=1}^{d} p_s(I_k^s)$ are the MICA distributions associated with J_c and J_s respectively. ♣

The proof of Lemma2 follows immediately from the fact that: $D(\mu^c(J_c)||\prod_{k=1}^{d}\mu_m^c(I_k^c)) < D(\mu^s(J_s)||\prod_{k=1}^{d}\mu_m^s(I_k^s))$ for non-stationary patches (and where β is assumed to be unity throughout without loss of generality). Furthermore given that different textures have non-identical independent components, it follows that equality holds if and only if the patch being analyzed is stationary.

Finally we remark that in the preceding sections we have dealt with the problem of modeling unimodal distributions associated with natural image textures. In general when modeling multimodal distributions obtained from NSS, however, more accurate results could be obtained by mixtures of MICA models. The question of parameter estimation for this case is a subject of future work.

5. APPENDIX

Proof of Lemma 1: We prove the lemma by induction on d. For the base case ($d = 2$), it is easily shown that: $J(F) = \frac{1}{|C|\beta^2}\prod_{k=1}^{2}\psi(\beta z_k)$. Now assume by inductive hypothesis that the lemma is true for $d = 2, \cdots, N$. Now consider the Jacobian when $d = N + 1$:

$$J(F) = \begin{vmatrix} \frac{\partial F_1}{\partial s_1} & \cdot & \cdot & \frac{\partial F_1}{\partial s_{N+1}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial F_{N+1}}{\partial s_1} & \cdot & \cdot & \frac{\partial F_{N+1}}{\partial s_{N+1}} \end{vmatrix} =$$

$$\begin{vmatrix} a_{1,1}\psi(\beta z_1) & \cdot & \cdot & a_{1,N+1}\psi(\beta z_1) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ a_{N+1,1}\psi(\beta z_{N+1}) & \cdot & \cdot & a_{N+1,N+1}\psi(\beta z_{N+1}) \end{vmatrix}$$

(obtained after some manipulation). Now let us expand $J(F)$ with respect to the first row:
$J(F) = \frac{1}{\beta^{N+1}}\sum_{k=1}^{N+1}(-1)^{k+1}(|C^{(1,k)}|a_{1,k}\psi(\beta z_1))$ (where $C^{(1,k)}$ is the minor matrix of $J(F)$ with respect to $(1,k)$, and $|.|$ when applied to matrices denotes the determinant). Now applying the inductive hypothesis we have:
$J(F) = \frac{1}{\beta^{N+1}}\prod_{k=1}^{N+1}\psi(\beta z_k)\sum_{k=1}^{N+1}(-1)^{(k+1)}|A^{(1,k)}|a_{1,k}$ (where $A^{(1,k)}$ is the minor matrix of A with respect to $(1,k)$). Thus $J(F) = \frac{1}{\beta^{N+1}}\prod_{k=1}^{N+1}\psi(\beta z_k)|A| = \frac{1}{\beta^{N+1}|C|}\prod_{k=1}^{N+1}\psi(\beta z_k)$ thereby proving the lemma for all d.

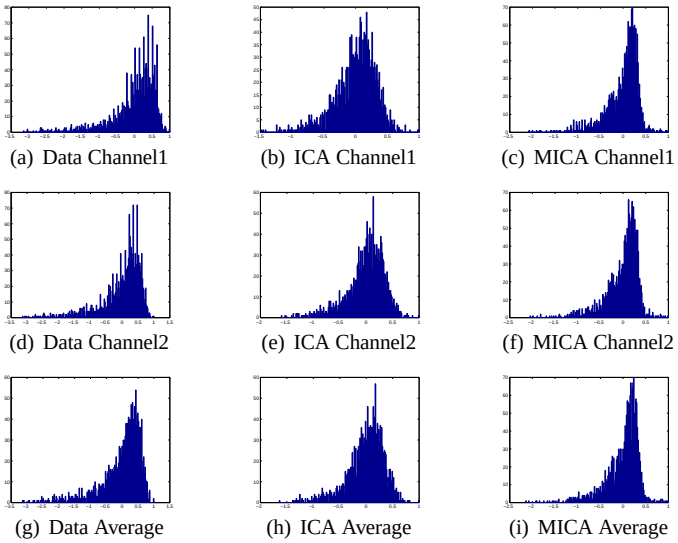

Fig. 3. Gravel Texture: Histograms of the Data channels and the corresponding ICA and MICA distributions.

6. REFERENCES

[1] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification," 2001.

[2] J-F. Cardoso, "Blind source separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[3] A. Hyvarinen, P.O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, July 2001.

[4] "http://sipi.usc.edu/database/database.cgi?volume=textures," .

[5] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.