Riemannian Optimization Method on the Flag Manifold for Independent Subspace Analysis

Yasunori Nishimori[1], Shotaro Akaho[1], and Mark D. Plumbley[2]

[1] Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST),
AIST Central2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
y.nishimori@aist.go.jp, s.akaho@aist.go.jp
[2] Department of Electronic Engineering, Queen Mary University of London,
Mile End Road, London E1 4NS, UK
mark.plumbley@elec.qmul.ac.uk

Abstract. Recent authors have investigated the use of manifolds and Lie group methods for independent component analysis (ICA), including the Stiefel and the Grassmann manifolds and the orthogonal group $O(n)$. In this paper we introduce a new class of manifold, the *generalized flag manifold*, which is suitable for independent subspace analysis. The generalized flag manifold is a set of subspaces which are orthogonal to each other, and includes the Stiefel and the Grassmann manifolds as special cases. We describe how the independent subspace analysis problem can be tackled as an optimization on the generalized flag manifold. We propose a Riemannian optimization method on the generalized flag manifold by adapting an existing geodesic formula for the Stiefel manifold, and present a new learning algorithm for independent subspace analysis based on this approach. Experiments confirm the effectiveness of our method.

1 Introduction

Independent component analysis (ICA) assumes a statistical generative model of the form $x = As$ where $s = (s_1, \ldots, s_n)^\top \in \mathbb{R}^n$, the components s_i are generated from statistically independent sources, and $A \in \mathbb{R}^{n \times n}$ is an invertible matrix. Most algorithms for ICA use whitening as a preprocessing step, giving $z = Bx$ such that $E[zz^\top] = I_n$. After the whitening operation, solving the ICA task reduces to an optimization on the orthogonal group [9, 10], i.e. over the set of orthogonal demixing matrices $\{W \in \mathbb{R}^{n \times n} | W^\top W = I_n\}$ where W is a demixing matrix in $y = (y_1, \ldots, y_n)^\top = W^\top z$ which attempts to recover the original signals up to a scaling and permutation ambiguity.

Optimization on a special class of manifolds related to the orthogonal group such as the Stiefel and the Grassmann manifolds frequently appear in the context of neural networks, signal processing, pattern recognition, computer vision, numerics and so on [5]. Principal, minor, and independent component analysis are formalized as optimization on the Stiefel manifold, subspace tracking

and application-driven dimension reduction can be solved by optimization on the Grassmann manifold. Generally, optimization on manifolds raises more intricate problems than optimization on Euclidean space, however, optimization on the Stiefel and the Grassmann manifolds can be tackled in a geometrically natural way based on Riemannian geometry. Because those manifolds are homogeneous, we can explicitly compute various Riemannian quantities relative to a Riemannian metric g, including Riemannian gradient, Hessian, geodesic, parallel translation vector, and curvature. Therefore, by replacing, for instance, an ordinary gradient vector of a cost function by a Riemannian gradient vector, and updating a point along the geodesic in direction to the Riemannian gradient, we get a Riemannian gradient descent method on that manifold:

$$W_{k+1} = W_k - \eta \nabla f(W_k) : \text{Euclidean}$$

$$W_{k+1} = \varphi_M(W_k, -\operatorname{grad}_W f(W_k), \eta) : \text{Riemannian},$$

where $\varphi_M(W, V, t)$ denotes the geodesic equation on a manifold M starting from $W \in M$ in direction $V \in T_W M$ relative to a Riemannian metric g on M. As such, other iterative optimization methods on Euclidean space are directly modified for the Stiefel and the Grassmann manifolds just replacing everything by Riemannian counterparts. Riemannian optimization methods are performed along piecewise geodesic curve on the manifold, so updated points always stays on the manifold and guarantees the stability against the deviation from the manifold. For more information about this approach, see e.g. [5].

In this paper we introduce a new class of manifold: the *generalized flag manifold*. This new manifold naturally arises when we slightly relax the assumption of ICA and consider independent subspace analysis (ISA), allowing dependence between signals within to different subspaces. So far researchers in neural networks, signal processing have mainly concentrated on the Stiefel and Grassmann manifolds for optimization: the generalized flag manifold is a generalization of both the Stiefel and the Grassmann manifolds. We extend the Riemannian optimization method to this new manifold using our previous geodesic formula for the Stiefel manifold [10], and based on it propose a new algorithm for ISA. Finally, we present computer experiments comparing the ordinary gradient method for ISA with the new Riemannian gradient geodesic method based on the flag manifold.

2 Independent subspace analysis

Hyvärinen and Hoyer introduced independent subspace analysis (ISA) [8], by relaxing the usual statistically independent condition of each source signal in ICA. The source signal s is decomposed into d_i-tuples ($i = 1, \ldots, r$) where signals within a particular tuple are allowed to be dependent on each other, while signals belonging to different tuples are statistically independent. Since the ISA algorithm uses pre-whitening, we have $W^\top W = I_n$ as for normal ICA. However, because of the statistical dependence of signals within tuples, the manifold of

candidate matrices for ISA is no longer simply the Stiefel manifold, but rather it is the Stiefel manifold with an additional symmetry.

Therefore solving ISA task can be regarded as an optimization on this new manifold, which is known as the *generalized flag manifold*.

3 What is the generalized flag manifold?

Let us introduce the generalized flag manifold. A *flag* in $\mathbb{R}^n$ is an increasing sequence of subspaces $V_1 \subset \cdots \subset V_r \subset \mathbb{R}^n$ of $\mathbb{R}^n$, $1 \leq r \leq n$. Given any sequence $(n_1, \ldots, n_r)$ of nonnegative integers with $n_1 + \cdots + n_r \leq n$, the *generalized flag manifold* $\mathrm{Fl}(n_1, n_2, \ldots, n_r)$ is defined as the set of all flags $(V_1, \ldots, V_r)$ of vector spaces with $V_1 \subset \cdots \subset V_r \subset \mathbb{R}^n$ and $\dim V_i = n_1 + \cdots + n_i, i = 1, 2, \ldots, r$. $\mathrm{Fl}(n_1, n_2, \cdots, n_r)$ is a smooth, connected, compact manifold. We can also consider a modified version of the generalized flag manifold, were we instead consider the set of the vector spaces V

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_r \subset \mathbb{R}^n, \tag{1}$$

where $\dim V_i = d_i$, $d_1 \leq d_2 \leq \cdots \leq d_r, \dim V = \sum_{i=1}^{r} d_i = p \leq n$. With the mapping $V_i \mapsto \bigoplus_{j=1}^{i} V_j$ we can see that the set of all V forms a manifold isomorphic to the original definition so this is also a generalized flag manifold, which we denote by $\mathrm{Fl}(n, d_1, d_2, \cdots, d_r)$. We represent a point on this manifold by a n by p orthogonal matrix W, i.e. $W^\top W = I_p$, which can be decomposed as

$$W = [W_1, W_2, \cdots, W_r],$$

$$W_i = (w_1^i, w_2^i, \cdots, w_{d_i}^i),$$

where $w_k^i \in \mathbb{R}^n$, $k = 1, \cdots, d_i$ for some i, form the orthogonal basis of V_i. Note that we are not so concerned with the individual frame vectors w_k^i themselves, rather with the subspace in $\mathbb{R}^n$ spanned by that set of vectors for some i. In other words any two matrices W_1, W_2 related by

$$W_2 = W_1 \begin{pmatrix} R_1 & & & \\ & R_2 & & \\ & & \ddots & \\ & & & R_r \end{pmatrix} \equiv W_1 \mathrm{diag}(R_1, R_2, \cdots, R_r) \tag{2}$$

where R_i $(1 \leq i \leq r) \in O(d_i)$, i.e. $R_i R_i^T = R_i^T R_i = I_{d_i}$, correspond to the same point on the manifold: we say we *identify* these two matrices. The generalized flag manifold is a generalization of both the Stiefel and the Grassmann manifolds. If all d_i $(1 \leq i \leq r) = 1$, it reduces to the Stiefel manifold, if $r = 1$, it reduces to the Grassmann manifold [4].

4 Geometry of the flag manifold

4.1 Tangent space structure

A tangent space of a manifold is an analogue of a tangent plane of a hypersurface in Euclidean space. For W to represent points in $\mathrm{Fl}(n, d_1, d_2, \cdots, d_r)$ we have

$$W^\top W = I_p \tag{3}$$

$$W\mathrm{diag}(R_1, R_2, \ldots, R_r), \quad R_i \in O(d_i), \quad 1 \le i \le r \quad \text{are identified.} \tag{4}$$

A tangent vector $V = [V_1, V_2, \ldots, V_r]$ of $\mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$ at $W = [W_1, W_2, \ldots, W_r]$ must satisfy the equation obtained by differentiating (3)

$$W^\top V + V^\top W = O, \tag{5}$$

Now let us consider the following curve on the flag manifold passing through W.

$$W\mathrm{diag}(R_1(t), R_2(t), \ldots, R_r(t))$$

where $R_i(t) \in O(d_i)$, $R_i(0) = I_{d_i}$, $(1 \le i \le r)$. Since we neglect the effect of each rotation $R_i(t)$, V must be orthogonal to

$$W\mathrm{diag}(R_1'(0), R_2'(0), \ldots, R_r'(0)) = W\mathrm{diag}(X_1, \ldots, X_r),$$

where X_i $(1 \le i \le r)$ is a $d_i \times d_i$ skew symmetric matrix. Thus

$$0 = \mathrm{tr}\left\{\mathrm{diag}(X_1^\top, \ldots, X_r^\top) W^\top V\right\} = -\sum_{i=1}^{n} \mathrm{tr} X_i W_i^\top V_i \text{ for all } X_i$$

so we can show that $W_i^\top V_i$ is symmetric. Now the i-i block of (5) yields

$$W_i^\top V_i + V_i^\top W_i = O.$$

and therefore $W_i^\top V_i = O$, $(1 \le i \le r)$.

So to summarize, a tangent vector $V = [V_1, V_2, \ldots, V_r]$ of $\mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$ at $W = [W_1, W_2, \ldots, W_r]$ is characterized by

$$W^\top V + V^\top W = O \qquad \text{and} \qquad W_i^\top V_i = O, \quad i = 1, \ldots, r. \tag{6}$$

4.2 Natural gradient

In this section we derive the natural gradient of a function over the generalized flag manifold.

We use the following notations.

$$G = I - \frac{1}{2} W W^\top \quad \text{with} \quad G^{-1} = I + W W^T \tag{7}$$

$$X = \nabla_W f = \left(\frac{\partial f}{\partial w_{ij}}\right) = [X_1, \ldots, X_r] \tag{8}$$

$$Y = G^{-1} \nabla_W f = (I + W W^\top) \nabla_W f, \qquad Y_i = G^{-1} X_i \tag{9}$$

Now the natural gradient V of a function f on $\mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$ is equal to the orthogonal projection of X to $T_W \mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$. In other words, $T_W \mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$ is obtained through the minimization of $(V-X)^t G(V-X)$ under the constraints (6). This can be solved by the Lagrange multiplier method. Let us introduce

$$L = \mathrm{tr}[(V-Y)^\top G(V-Y)] - \sum_i \mathrm{tr}(A_i^\top W_i^\top V_i) - \sum_i \sum_{j \neq i} \mathrm{tr}(B_{ij}^\top (W_i^\top V_j + V_i^\top W_j))$$

$$= \mathrm{tr}[(V-Y)^\top G(V-Y)] - \sum_i \mathrm{tr}(A_i^\top W_i^\top V_i) - \sum_i \sum_{j \neq i} \mathrm{tr}(B_{ij}^\top W_i^\top V_j + B_{ji} V_i^\top W_j).$$

Differentiating L with respect to V_i and equating to zero leads to

$$\frac{\partial L}{\partial V_i} = 2G(V_i - Y_i) - W_i A_i - \sum_{j \neq i} W_j (B_{ij} + B_{ji}^\top) = O. \tag{10}$$

$$V_i = Y_i + \frac{G^{-1}}{2} W_i A_i + \frac{G^{-1}}{2} \sum_{j \neq i} W_j (B_{ij} + B_{ji}^\top) \tag{11}$$

Substituting $G^{-1} = I + WW^\top = I + \sum_i W_i W_i^\top$ into (11), we get

$$V_i = Y_i + W_i A_i + \sum_{j \neq i} W_j (B_{ij} + B_{ji}^\top).$$

Therefore, the condition of the tangent vector yields

$$W_i^\top V_i = W_i^\top Y_i + 2A_i = O \Leftrightarrow A_i = -\frac{1}{2} W_i^\top Y_i, \tag{12}$$

$$W_i^\top V_j + V_i^\top W_j = W_i^\top Y_j + (B_{ji} + B_{ij}^\top) + Y_i^\top W_j + B_{ij}^\top + B_{ji} = O. \tag{13}$$

Thus,

$$B_{ij}^\top + B_{ji} = -\frac{1}{2}(W_i^\top Y_j + Y_i^\top W_j),$$

and we get

$$V_i = Y_i - W_i W_i^\top Y_i - \frac{1}{2} \sum_{j \neq i} (W_j W_j^\top Y_i + W_j Y_j^\top W_i) \tag{14}$$

$$= X_i - (W_i {W_i}^\top X_i + \sum_{j \neq i} W_j X_j^\top W_i). \tag{15}$$

It is easy to check that this formula includes the natural gradient formula for the Stiefel and the Grassmann manifolds [4] as special cases.

4.3 Geodesic of the flag manifold

We use our geodesic formula of the Stiefel manifold relative to the normal homogeneous metric for the Stiefel manifold: $G = I - \frac{1}{2} WW^\top$ obtained in [10].

$$\varphi_{\mathrm{St}(n,p)}(W, -\mathrm{grad}_W f, t) = \exp(-t(\nabla f(W) W^\top - W \nabla f(W)^\top)) W \tag{16}$$

$$\varphi_{\mathrm{St}(n,p)}(W, V, t) = \exp(t(DW^\top - WD^\top)) W, \tag{17}$$

where $D = (I - \frac{1}{2}WW^\top)V$.

We decompose the tangent space of $\mathrm{St}(n,p)$ at W into the vertical space V_W and the horizontal space H_W with respect to G.

$$T_W \,\mathrm{St}(n,p) = H_W \oplus V_W.$$

The vertical space does not depend on the metric; it is determined from the quotient space structure of $T_W \,\mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$:

$V_W = W\mathrm{diag}(X_1^\top, \ldots, X_r^\top)$, where $X_i \in TO(d_i)$ (the set of skew symmetric matrices).

We need to lift a tangent vector V at $T_W \,\mathrm{Fl}(n, d_1, d_2, \ldots, d_r)$ to $\tilde{V} \in H_W$. It turns out that $\tilde{V} = V$, because

$$g_W^{\mathrm{St}(n,p)}(X,V) = \mathrm{tr}\{X^\top (I - \frac{1}{2}WW^\top)V\} \tag{18}$$

$$= -\mathrm{tr}\{\mathrm{diag}(X_1^\top, \ldots, X_r^\top)W^\top (I - \frac{1}{2}WW^\top)V\} \tag{19}$$

$$= -\frac{1}{2}\mathrm{tr}\{\left(\mathrm{diag}(X_1^\top, \ldots, X_r^\top)\right) W^\top V\} = 0, \text{ for all } X \in V_W. \tag{20}$$

Because the projection $\pi : \mathrm{St}(n,p) \to \mathrm{Fl}(n, d_1, \ldots, d_r)$ $(\pi(W) = W)$ is a Riemannian submersion, the following theorem guarantees that the geodesic on the Stiefel manifold starting from W in direction V yields the geodesic on the flag manifold emanating from W with the initial velocity V. Both geodesics are based on the same Riemannian metric G.

Let $p : \tilde{M} \to M$ be a Riemannian submersion, $\tilde{c}(t)$ be a geodesic of $(\tilde{M}, g^{\tilde{M}})$. If the vector $\tilde{c}^{'}(0)$ is horizontal, then $\tilde{c}^{'}(t)$ is horizontal for any t, and the curve $p(\tilde{c}(t))$ is a geodesic of (M, g) of the same length as $\tilde{c}(t)$ [6, 10].

5 Application to independent subspace analysis

In order to validate the effectiveness of the proposed algorithm, we performed independent subspace analysis experiments of natural image data [8]. In this experiment, we attempt to decompose a gray-scale image $I(x,y)$ into linear combination of basis images $a_i(x,y)$ as

$$I(x,y) = \sum_{i=1}^{n} s_i a_i(x,y), \tag{21}$$

where s_i is a coefficient. Let the inverse filter of this model be

$$s_i = \langle w_i, I \rangle = \sum_{x,y} w_i(x,y) I(x,y). \tag{22}$$

The problem is to estimate s_i (or equivalently $w_i(x,y)$) when a set of images are given. For this purpose, we apply the independent subspace criterion: proposed by Hyvärinen et al. [8] Components are partitioned into disjoint subspaces

$S_1, \ldots, S_r$, and s_i and s_j are statistically independent if i and j belong to different subspaces. As a cost function, we take a negative log-likelihood defined by

$$f(\{w_i\}) = -\sum_{k=1}^{K} \log L(I_k; \{w_i\}) = -\sum_{k=1}^{K} \sum_{j=1}^{r} \log p \left(\sum_{i \in S_j} \langle w_i, I_k \rangle^2 \right) \tag{23}$$

where the suffix k denotes the index of samples and p is the exponential distribution $p(x) = \alpha \exp(-\alpha x)$, where the parameter α does not appear in learning rule and hence can be ignored. The cost function is invariant under rotation within the subspace.

We applied the above model to small image patches of size 16×16 pixels. We prepared 10000 image patches at random locations extracted from monochrome photographs of natural images. (The dataset and ISA code is distributed by Hyvärinen http://www.cis.hut.fi/projects/ica/data/images).

In the preprocessing phase, the mean gray-scale value of each image patch was subtracted, and then the dimension of the image was reduced from 256 to 160 by principal component analysis ($n = 160$), and finally the data were whitened. Independent subspace analysis was performed for this dataset, in which each subspace is 4-dimensional ($d_i = 4$), and accordingly the 160 dimensional subspaces were separated into 40 subspaces ($r = 40$).

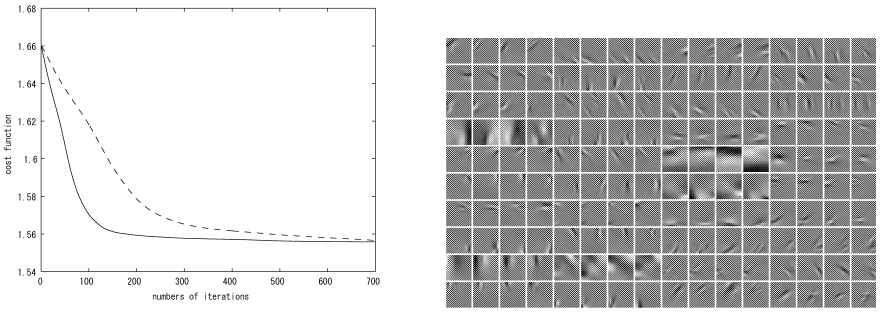
We compared two methods to extract independent subspaces from the dataset. One is the ordinary gradient method used in [8] ($\eta = 1$) and the other is the proposed method based on geodesic of the flag manifold ($\eta = 1.1$). In the ordinary gradient method, the extraction matrix was projected to the orthogonal group by singular value decomposition in each step.

The behavior of the cost function along iterations is shown in Fig. 1(a). In early stages of learning, the cost of the geodesic approach decreases much faster than the ordinary gradient method. The recovered inverse filter $w_i(x, y)$ are shown in Fig. 1(b). The filters are clearly separated into groups. We found no significant difference between the points of convergence of the two methods, and neither method appeared to get 'stuck' in a local optimum.

6 Conclusion

We have introduced a new manifold, the generalized flag manifold, for solving the independent subspace analysis (ISA) problem, and have developed a Riemannian gradient descent geodesic method on the generalized flag manifold. Computer experiments confirm that our algorithm gives good performance compared with the ordinary gradient descent method.

While we have concentrated on the gradient descent method in this paper, conjugated gradient and the Newton methods could also be used for searching over manifold using geodesics. Also, while we used orthogonal matrices to represent points on the flag manifold, the algorithm could be described using non-orthogonal matrices, as Absil et al have done for the Grassmann manifold [1].



(a) (b)

Fig. 1. Results, showing (a) learning curves; (b) recovered inverse filters

7 Acknowledgements

This work is supported in part by JSPS Grant-in-Aid for Exploratory Research 16650050, MEXT Grant-in-Aid for Scientific Research on Priority Areas 17022033, and EPSRC grants GR/S75802/01, GR/S82213/01, EP/D000246/1 and EP/C005554/1.

References

1. P-A. Absil, R. Mahony, R. Sepulchre, and P. Van Dooren, Riemannian geometry of Grassmann manifolds with a view on algorithmic computation, *Acta Applicandae Mathematicae*, **80**(2), pp.199-220, 2004.
2. S. Amari, Natural gradient works efficiently in Learning, *Neural Computation*, **10**, pp.251-276, 1998.
3. A. Besse, Einstein Manifolds, Springer-Verlag, 2002.
4. A. Edelman, T.A. Arias, and S.T. Smith, The Geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications*, **20**(2), pp.303-353, 1998.
5. S. Fiori, Quasi-Geodesic Neural Learning Algorithms over the Orthogonal Group: A Tutorial, *Journal of Machine Learning Research*, **6**, pp.743-781, 2005.
6. S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian Geometry*, Springer, 1990.
7. U. Helmke, J. B. Moore, *Optimization and dynamical systems*, Springer, 1994.
8. A. Hyvärinen and P.O. Hoyer, Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, **12**(7), pp.1705-1720, 2000.
9. Y. Nishimori, Learning Algorithm for Independent Component Analysis by Geodesic Flows on Orthogonal Group, *Proceedings of International Joint Conference on Neural Networks (IJCNN1999)*, **2**, pp.1625-1647, 1999.
10. Y. Nishimori, Learning Algorithms Utilizing Quasi-Geodesic Flows on the Stiefel Manifold, *Neurocomputing*, **67** pp.106-135, 2005.
11. M. D. Plumbley, Algorithms for non-negative independent component analysis. *IEEE Transactions on Neural Networks*, **14**(3), pp.534-543, 2003.