FLAG MANIFOLDS FOR SUBSPACE ICA PROBLEMS

Yasunori Nishimori, Shotaro Akaho

Neuroscience Research Institute
AIST
1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan

Samer Abdallah, Mark D. Plumbley

Department of Electrical Engineering
Queen Mary, University of London
Mile End Road, London E1 4NS, UK

ABSTRACT

We investigate the use of the Riemannian optimization method over the flag manifold in subspace ICA problems such as independent subspace analysis (ISA) and complex ICA. In the ISA experiment, we use the Riemannian approach over the flag manifold together with an MCMC method to overcome the problem of local minima of the ISA cost function. Experiments demonstrate the effectiveness of both Riemannian methods – simple geodesic gradient descent and hybrid geodesic gradient descent, compared with the ordinary gradient method.

Index Terms— Independent subspace analysis, complex ICA, natural gradient, geodesic, flag manifolds, Riemannian optimization method

1. INTRODUCTION

Several signal processing tasks, such as those related to independent component analysis (ICA), can be approached by optimization over a manifold. Examples of such manifolds include the orthogonal group $O(n)$, corresponding to the set of orthonormal $n \times n$ matrices, the Stiefel manifold $\mathrm{St}(n, p\,; \mathbb{R})$, corresponding to the set of orthonormal $n \times p$ matrices,

$$\left\{W = (w_1, \ldots, w_p) \in \mathbb{R}^{n \times p} | W^\top W = I_p, n \geq p\right\},$$

and the Grassmann manifold $\mathrm{Gr}(n, p\ ; \mathbb{R})$ of unoriented p-planes, corresponding to the subspaces spanned by $n \times p$ full rank matrices. Stiefel manifolds have been used in ICA and PCA in the case where the number of the extracted components is less than the number of the mixed signals [8], while Grassmann manifolds have been utilized for invariant subspace computation and subspace tracking [1].

One-unit ICA extracts one independent component at a time, while ordinary ICA extracts several components simultaneously by optimization over the Stiefel manifold. A single subspace can be represented as a point on the Grassmann manifold, which can be used for subspace analysis. The question then arises: what manifold will be necessary for extracting several subspaces simultaneously? This leads us to the concept of the flag manifold, which is the manifold consisting of orthogonal subspaces. This is closely related to the Stiefel manifold, and include the Grassmannian manifold as a special case.

We extend the Riemannian optimization method to the flag manifold by deriving the formulas for the natural gradient and geodesics on the manifold. We show how the flag manifold method can be applied to subspace ICA problems such as independent subspace analysis [5] and complex ICA.

We show that we can replace an optimization over the complex Stiefel manifold $\mathrm{St}(n, p, \mathbb{C})$ of $n \times p$ complex unitary matrices ($n \geq p$) with an optimization over the real generalized flag manifold of p 2-dimensional subspaces in $\mathbb{R}^{2n}$. Thus, for example, the complex ICA problem of separating p complex independent sources from a sequence of $n > p$ complex observations can also be tackled using a generalized flag manifold.

We also consider the problem of local minima in ISA and propose a hybrid geodesic gradient-MCMC method to tackle the problem. This algorithm takes geodesic gradient descent steps in the flag manifold interleaved with random interchanges of basis vectors between subspaces. These swaps prevent the system from becoming trapped in local minima of the ISA cost function, while the Riemannian optimization method accelerates convergence between swaps.

2. FLAG MANIFOLD

Manifolds that frequently arise from signal processing tasks are Lie groups, and homogeneous spaces of Lie groups. A homogeneous space M is defined to be a manifold on which a Lie group G acts transitively, and it is expressed as the quotient space of G by its isotropy subgroup H: G/H, where an isotropy subgroup H of $p \in M$ consist of the stabilizers of a point $p \in M$: $H = \{x \in G | x \cdot p = p\}$. It is also worth mentioning that G is a fiber bundle over G/H whose fiber is isomorphic to H. The formula for geodesics over the flag manifold is derived by regarding the Stiefel manifold as a fiber bundle over the flag manifold. Previous use of homogeneous spaces has mainly concentrated on the Stiefel manifold $\mathrm{St}(n, p\,; \mathbb{R})$, which is is diffeomorphic to $O(n)/O(n-p)$ and

This work is partly supported by JSPS Grant-in-Aid for Exploratory Research 16650050, MEXT Grant-in-Aid for Scientific Research on Priority Areas 17022033, and EPSRC Grant GR/S82213/01.

the Grassmann manifold $\mathrm{Gr}(n,p\,;\mathbb{R})$, which is diffeomorphic to $O(n)/O(p)\times O(n-p)$.

Extending the concept of the Grassmann manifold, we introduce a new class of manifold – the flag manifold [6]. Given an ordered sequence $(n_1,\ldots,n_r)$ of nonnegative integers with $n_1+\cdots+n_r\le n$, the flag manifold $\mathrm{Fl}(n_1,n_2,\ldots,n_r)$ is defined to be the set of sequence of vector spaces $V_1\subset\cdots\subset V_r\subset\mathbb{R}^n$ with $\dim V_i = n_1+\cdots+n_i, i=1,2,\ldots,r$. $\mathrm{Fl}(n_1,n_2,\cdots,n_r)$ is a smooth, connected, compact manifold. We need a different expression of the flag manifold to tackle the subspace ICA problems; we consider the set of the orthogonal direct sum of the vector spaces V

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

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

$$W=[W_1,W_2,\ldots,W_r], W_i=(w_1^i,w_2^i,\ldots,w_{d_i}^i),$$

where $w_k^i\in\mathbb{R}^n$, $k=1,\ldots,d_i$ for some i, form the orthogonal basis of V_i. The orthogonal group $O(n)$ acts transitively on the flag manifold $\mathrm{Fl}(n,\mathbf{d})$ by matrix multiplication:

$$O(n)\times\mathrm{Fl}(n,\mathbf{d})\ni(R,W)\mapsto RW\in\mathrm{Fl}(n,\mathbf{d}).$$

It is easily seen that the isotropy subgroup of $O(n)$ at W is:

$$[W,W_\perp]\mathrm{diag}[R_1,R_2,\ldots,R_r,R_{r+1}][W,W_\perp]^\top,$$

where $R_k\in O(d_k),(1\le k\le r), R_{r+1}\in O(n-p)$, and $W_\perp$ is an arbitrary $n\times(n-p)$ matrix satisfying $[W,W_\perp]\in O(n)$. Therefore

$$\mathrm{Fl}(n,\mathbf{d}\,;\mathbb{R})\cong O(n)/O(d_1)\times\cdots\times O(d_r)\times O(n-p). \quad (1)$$

$\mathrm{Fl}(n,\mathbf{d}\,;\mathbb{R})$ is locally isomorphic [1] to $\mathrm{St}(n,p\,;\mathbb{R})$ as a homogeneous space when all d_i $(1\le i\le r)=1$, and it reduces to a Grassmann manifold if $r=1$. It may well be said that the flag manifold is a generalization of the Stiefel and Grassmann manifolds in this sense.

We can obtain the Riemannian optimization method over a manifold by adapting ordinary optimization methods over the Euclidean space to the manifold: first the updated direction is replaced by the Riemannian counterpart, which is geometric in the sense that it does not depend on parametrizations of the manifold; second, the current point is updated to the next along a geodesic on the manifold, thus updated points always stay on the manifold, which guarantees the stability against the deviation from the manifold:

$$W_{k+1}=\varphi_M(W_k,-\mathrm{grad}_W f(W_k),\eta), \quad (2)$$

[1] A homogeneous space G/H is locally isomorphic to G'/H' when the Lie algebras of G,H are locally isomorphic to G',H' respectively.

where $\varphi_M(W,V,t)$ denotes the equation of a geodesic over manifold M starting from $w\in M$ in the direction of $V\in T_wM$ such that $\varphi_M(W,V,0)=W,\varphi'_M(W,V,0)=V$. We derived in [8] the equation of a geodesic on a flag manifold with respect to the normal metric

$$g_W^{\mathrm{Fl}(n,\mathbf{d};\mathbb{R})}(V_1,V_2)=\mathrm{tr}V_1^\top(I-\tfrac{1}{2}WW^\top)V_2,$$

where $V_1,V_2\in T_W\mathrm{Fl}(n,\mathbf{d}\,;\mathbb{R})$:

$$\varphi_{\mathrm{Fl}(n,\mathbf{d};\mathbb{R})}(W,V,t)=\exp(t(DW^\top-WD^\top))W,$$

where $D=(I-\frac{1}{2}WW^\top)V$. The natural gradient V of a function f on $\mathrm{Fl}(n,\mathbf{d}\,;\mathbb{R})$ at W with respect to $g^{\mathrm{Fl}(n,\mathbf{d};\mathbb{R})}$ is:

$$V_i=X_i-(W_i{W_i}^\top X_i+\textstyle\sum_{j\ne i}W_jX_j^\top W_i),$$

where $V=(V_1,\ldots,V_r)$.

3. COMPLEX ICA

We illustrate how a special class of optimization problems over the complex Stiefel manifold are transformed to optimization problems over the flag manifold, thereby making the Riemannian optimization method over the flag manifold applicable to the complex ICA problem.

Let us consider an optimization problem over the complex Stiefel manifold:

$$F:\mathrm{St}(n,p\,;\mathbb{C})\to\mathbb{R},$$

where $\mathrm{St}(n,p\,;\mathbb{C})=\{W=(w_1,\ldots,w_p)=W_\Re+iW_\Im\in\mathbb{C}^{n\times p}|W^HW=I_p\}$ (H denotes the Hermitian transpose operator). We assume F is invariant under the transformation

$$W=(w_1,\ldots,w_p)\mapsto\left(e^{i\theta_1}w_1,\ldots,e^{i\theta_p}w_p\right), \quad (3)$$

which is satisfied by cost functions of signal processing tasks including complex ICA.

Because the cost function F is real-valued, $\mathrm{St}(n,p\,;\mathbb{C})$ should be regarded as a *real manifold* rather than a complex manifold, for which we embed $\mathrm{St}(n,p\,;\mathbb{C})$ into $\mathbb{R}^{2n\times 2p}$ by the following map:

$$\begin{aligned}\tau:W=\left(w_1^\Re+iw_1^\Im,\ldots,w_p^\Re+iw_p^\Im\right)&\mapsto\tilde{W}\\&=\begin{pmatrix}w_1^\Re & -w_1^\Im & w_2^\Re & -w_2^\Im & \cdots & w_p^\Re & -w_p^\Im\\ w_1^\Im & w_1^\Re & w_2^\Im & w_2^\Re & \cdots & w_p^\Im & w_p^\Re\end{pmatrix}.\end{aligned} \quad (4)$$

It turns out that the embedded manifold $N=\tau(\mathrm{St}(n,p\,;\mathbb{C}))$ coincides with $\mathrm{St}(2n,2p\,;\mathbb{R})\cap T$, where $T=\tau(\mathbb{C}^{n\times p})$ is a subspace in $\mathbb{R}^{2n\times 2p}$. Thus minimizing F over $\mathrm{St}(n,p\,;\mathbb{C})$ is transformed to minimizing f over N, where $f(\tilde{W}):=F(W)$.

Furthermore, the assumption of F gives N an additional structure. Since the transformation on $\mathrm{St}(n,p\,;\mathbb{C})$ (3) induces the following transformation on N :

$$\tilde{W}\mapsto\tilde{W}\mathrm{diag}(R(\theta_1),R(\theta_2),\cdots,R(\theta_p)), \quad (5)$$

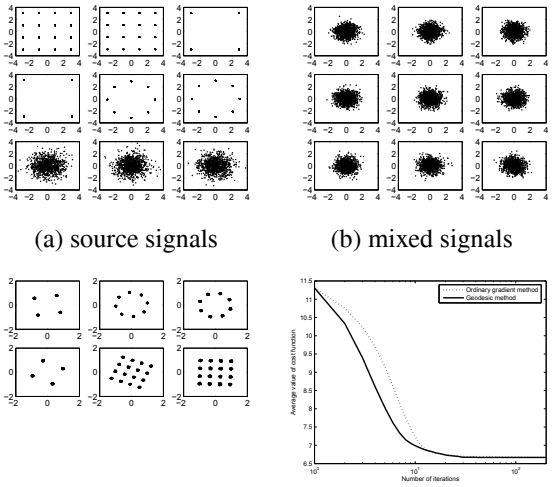
where $R(\theta_i) = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}$, the function f is also invariant under the transformation (5). Therefore, f can be interpreted as a function over a submanifold of the flag manifold[2]: $N' = \mathrm{Fl}(2n, \mathbf{2}\,; \mathbb{R}) \cap T$, where $\mathbf{2} = (2, \ldots, 2)$.

In fact, N' is a *totally geodesic submanifold* of $\mathrm{Fl}(2n, \mathbf{2}\,; \mathbb{R})$, that is, a geodesic on $\mathrm{Fl}(2n, \mathbf{2}\,; \mathbb{R})$ emanating from $\tilde{W} \in N'$ in the direction of $\tilde{V} \in T_{\tilde{W}} N'$ is always contained in N'. This allows us to consider just $\mathrm{Fl}(2n, \mathbf{2}\,; \mathbb{R})$ instead of its submanifold N'. To summarize, minimizing F over $\mathrm{St}(n, p\,; \mathbb{C})$ can be solved by minimizing the function f over the submanifold N' of $\mathrm{Fl}(2n, \mathbf{2}; \mathbb{R})$; to minimize f on N', we have only to apply the Riemannian optimization method for $\mathrm{Fl}(2n, \mathbf{2}\,; \mathbb{R})$ to f.

To explore the behavior of the Riemannian geodesic gradient descent method on the complex Stiefel manifold, we performed a numerical complex ICA experiment. Let us assume we are given 9 signals $x \in \mathbb{C}^9$ (Fig. 1(b)) which are complex-valued instantaneous linear mixture of two independent QAM16 signals, two QAM4 signals, two PSK8 signals, and three complex-valued Gaussian noise signals. We assume we know in advance the number of noise signals. The task of complex ICA under this assumption is to recover only non-noise signals $y = (y_1, \ldots, y_4)^\top$ so that $y = W^\top x$. As a preprocessing stage, we first center the data and then whiten it by SVD. Thus, $n \times p$ demixing matrix W can be regarded as a point on the complex Stiefel manifold $\mathrm{St}(n, p\,; \mathbb{C})$, namely $W^H W = I_p$. As an objective function, we use a kurtosis-like higher-order statistics: $F(W) = \sum_{i=1}^{4} E\left[||y_i(t)||^4\right]$. Then by minimizing $F(W)$ over $\mathrm{St}(n, p\,; \mathbb{C})$ we can solve the task.

We compared two algorithms. One is the Riemannian optimization method: $\tilde{W}_{k+1} = \varphi_{\mathrm{Fl}(2n,\mathbf{2};\mathbb{R})}(\tilde{W}_k, -\mathrm{grad}_{\tilde{W}_k} f(\tilde{W}_k))$, and another is the standard gradient descent method followed by projection: $W_{s+1} = \mathrm{pro}(W_s - \mu_s \frac{\partial f}{\partial \overline{W}_s})$, where $\frac{\partial f}{\partial \overline{W}_s}$ denotes $\frac{\partial f}{\partial W_\Re} + i\frac{\partial f}{\partial W_\Im}$, and pro means the projection onto $\mathrm{St}(n, p\,; \mathbb{C})$ by complex SVD. Both $\mathrm{grad}_{\tilde{W}_k} f(\tilde{W}_k)$ and $\frac{\partial f}{\partial \overline{W}_s}$ are computed by substituting $\frac{\partial ||y_i||^4}{\partial w_i^\Re} = 2||y_i||^2(y_i^* x + y_i x^*)$ and $\frac{\partial ||y_i||^4}{\partial w_i^\Im} = 2i||y_i||^2(y_i^* x - y_i x^*)$. Recall that we map $\mathrm{St}(n, p; \mathbb{C})$ to $\mathrm{Fl}(2n, \mathbf{2}; \mathbb{R})$ and update the matrices on $\mathrm{Fl}(2n, \mathbf{2}; \mathbb{R})$ using the correspondence between W and $\tilde{W}$ (4). After $\tilde{W}$ converges to $\tilde{W}_\infty$, $\tilde{W}_\infty$ is pulled back to $\mathrm{St}(n, p\,; \mathbb{C})$ to give a demixing matrix W_∞. The learning constant η_k, μ_s was chosen at each iteration based on the Armijo rule [10]. The separation result is shown in Fig. 1(c). The constellations of the source signals were recovered up to phase shifts. Both algorithms were tested for 100 trials. On each trial, a random nonsingular matrix was used to generate the data; a random unitary matrix was chosen as an initial demixing matrix; we iterated for 200 steps. The plots of Fig. 1(d) show

[2]Strictly speaking $\mathrm{Fl}(2n, \mathbf{2}\ ; \mathbb{R})$ should be replaced with $SO(2n)/SO(2) \times \ldots \times SO(2) \times SO(2n - 2p)$ – the universal cover of $\mathrm{Fl}(2n, \mathbf{2}\ ; \mathbb{R})$, yet both are locally isomorphic to each other as a homogeneous space, and we use $\mathrm{Fl}(2n, \mathbf{2}\ ; \mathbb{R})$ by abuse of notation.



(a) source signals (b) mixed signals

(c) recovered signals (d) cost function

Fig. 1. Complex ICA experiment

the average behavior of these two algorithms over 100 trials. We observed that the Riemannian optimization method decreased the cost more rapidly than the standard gradient descent method followed by projection, particularly in the early stages of learning .

4. ISA AND SUBSPACE IDENTIFICATION

In this section we examine the issue of local minima in ISA and the use of Markov chain Monte Carlo methods as a possible solution. These local minima are best observed in artificial problems where the correct solution is known and suboptimal solutions easily detected. Hence, the experiments described here were conducted using data drawn from independent subspaces with spherically symmetric multivariate Student's t distributions. The particular choice of the multivariate t distribution is convenient because it is easy to sample from and analytically tractable. The degrees-of-freedom parameter was set to 3, producing a moderately heavy-tailed distribution.

A maximum-likelihood (ML) estimator for the ISA system was constructed using the multivariate Student's t distribution as the prior within each subspace and the (negative) log-likelihood of the resulting model as the cost function; this was minimized using the geodesic gradient descent method.

The algorithm was tested in multiple runs on a number of 12 dimensional problems composed of, respectively, 2, 3, 4, and 6 dimensional subspaces. For each run, a random sample was drawn from the source distribution, the mixing matrix set to the identity, and the algorithm initialized with a random orthogonal matrix. The correct product of multivariate Student's t distributions was used as the prior; that is, the number and dimensionality of the subspaces was correct in each run.

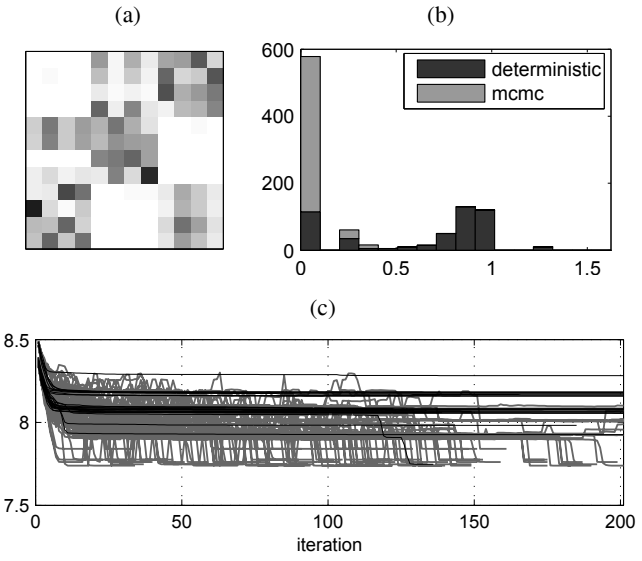


Fig. 2. (a) Example of a sub-optimal basis matrix reached by gradient descent in a problem with 3 subspaces of dimension 4 (white=zero, black=± 1). (b) Distribution of final Amari error over 100 runs of deterministic gradient descent (black) and 500 runs of the MCMC-gradient hybrid method (grey). (c) Evolution of cost function for multiple runs of deterministic (black lines) and MCMC (grey lines) algorithms in a problem with 3×4-dimensional subspaces.

It was observed that in many cases, the system would become stuck at a non-optimal, non-subspace-separating solution, an example of which is shown in fig. 2(a). In these cases, the output components are mixtures of sources from different source subspaces.

One solution to this problem is to allow the system to swap basis vectors at random with a probability related to the resulting change in the likelihood. Specifically, we use a Metropolis-Hastings methodology: a swap is proposed by selecting two columns of the basis matrix at random; this is accepted unconditionally if the likelihood is increased, and with probability $e^{-\beta(t)\Delta L}$ otherwise, where ΔL is the decrease in the likelihood and $\beta(t)$ is a time-varying inverse temperature parameter. Since we are aiming for a ML solution, the temperature is gradually decreased as the algorithm progresses, that is β is increased linearly from 20 to 60 in 200 steps. These MCMC updates to the basis matrix are interspersed with geodesic gradient descent steps, which quickly drive the system to a local minimum in between swaps. The gradient updates are disabled once the change in the likelihood drops below a small threshold (10^{-8} in these experiments), and then re-enabled as soon as a swap occurs. This reduces the amount of computation expended on relatively ineffectual gradient steps, and explains why some of the traces in fig. 2(c) terminate at less than 200 iterations: in these cases, less than 200 gradient steps were required to reach the correct solution.

The results are summarized in fig. 2(b), which shows that the hybrid geodesic gradient-MCMC algorithm is more likely to reach the correct solution to the problem than the purely deterministic gradient method.

5. CONCLUSIONS

We have demonstrated that the flag manifold is useful for tackling subspace ICA problems. The aim of this paper was not to pursue the best learning algorithm for a particular subspace ICA problem, rather the emphasis was to illustrate how the flag manifold naturally arises from the subspace ICA problems, and how we can exploit the geometric structures of the flag manifold to modify existing optimization algorithms to be adapted to these problems. Though we have concentrated on the gradient descent method in this paper, other optimization methods such as the fixed point method could also be formulated over the flag manifold.

6. REFERENCES

[1] P-A. Absil, R. Mahony, and R. Sepulchre, 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. 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.

[4] S. Fiori, Quasi-geodesic neural learning algorithms over the orthogonal group: a tutorial, *Journal of Machine Learning Research*, **6**, pp.743-781, 2005.

[5] 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.

[6] L.J. Mason, and N.M.J Woodhouse, *Integrability, Self-duality, and Twistor Theory*, Oxford University Press, 1996.

[7] 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.

[8] Y. Nishimori and S.Akaho, Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold, *Neurocomputing*, **67** pp.106-135, 2005.

[9] Y. Nishimori, S. Akaho and M.D. Plumbley, Riemannian optimization method on the flag manifold for independent subspace analysis, *Proceedings of ICA2006*, pp.295-1302, 2006.

[10] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, 1997.