A PROJECTION ALGORITHM BASED ON THE PYTHAGORIAN THEOREM AND ITS APPLICATIONS

Shotaro Akaho
AIST
Tsukuba, Ibaraki 305-8568 Japan
e-mail: s.akaho@aist.go.jp

Hideitsu Hino
University of Tsukuba
Tsukuba, Ibaraki 305-8573, Japan
e-mail: hinohide@cs.tsukuba.ac.jp

Neneka Nara, Ken Takano, Noboru Murata
Waseda University
Shinjuku, Tokyo 169-8555, Japan
e-mail: {extraterrestrial@moegi, ken.takano@toki}.waseda.jp,
noboru.murata@eb.waseda.ac.jp

We consider the α-projection from a point p on a dually flat manifold $\mathcal{S}$ to a submanifold $\mathcal{M} \subset \mathcal{S}$, which is a fundamental procedure in statistical inference. Since the α-projection can be found by minimizing an α-divergence[1], gradient descent type algorithms are often used. However, in some applications, the derivative of divergence is not available or numerically unstable. In this poster, we propose a simple and robust algorithm without calculating the derivative of divergence.

The algorithm is based on the Pythagorian theorem for dually flat manifold. Suppose $\{p_i\}_{i=1,\ldots,k} \in \mathcal{S}$ are represented by $-\alpha$-affine coordinate system, they define the $-\alpha$-flat submanifold $\mathcal{M}$ by their affine combinations, $\mathcal{M} = \{\sum_{i=1}^k \theta_i p_i \mid \sum_{i=1}^k \theta_i = 1\}$. Let $q \in \mathcal{M}$ be a candidate of the α-projection of $p \in \mathcal{S}$. When q is actually the α-projection, the Pythagorian theorem holds

$$r_i = D^{(\alpha)}(p, q) + D^{(\alpha)}(q, p_i) - D^{(\alpha)}(p, p_i) = 0. \tag{1}$$

If r_i is more than or less than zero, it means that the α-geodesic connecting p and q does not intersect orthogonally to $\mathcal{M}$.

Based on this fact, the proposed algorithm increases θ_i when $r_i > 0$ while it decreases θ_i when $r_i < 0$. In particular when we can assume all θ_i's are nonnegative, θ_i can be updated by $\theta_i^{(t+1)} = \theta_i^{(t)} f(r_i)$, where $f(r)$ is a positive and monotonically increasing function such that $f(0) = 1$. After the update, θ_i's are normalized so that $\sum_{i=1}^k \theta_i = 1$.

As applications of the proposed algorithm, we consider two problems: nonparametric e-mixture estimation and nonnegative matrix factorization.

The e-mixture is defined as an exponential mixture of k distributions $\{p_i(x)\}$,

$$p(x;\theta) = \exp\left(\sum_{i=1}^{k} \theta_i \log p_i(x) - b(\theta)\right), \quad \sum_{i=1}^{k} \theta_i = 1, \quad \theta_i \geq 0, \tag{2}$$

where $b(\theta)$ is a normalization factor. Compared to an ordinary mixture $\sum \theta_i p_i(x)$, the e-mixture has advantages that it belongs to exponential families and it satisfies the maximum entropy principle. We applied the e-mixture modeling to a transfer learning problem, where we have only a small number of samples for a target task while a lot of samples are given for similar tasks. The problem is to find the m-projection ($\alpha = -1$) of $p(x)$ representing the target data to an e-flat submanifold ($\alpha = 1$) defined by a set of e-mixtures of data distributions $\{p_i(x)\}_{i=1,\dots,k}$ corresponding to the data of similar tasks. We consider the problem in a nonparametric setting, where $p(x)$ and $p_i(x)$'s are empirical distributions. However, since the derivative of divergence is not available in the nonparametric setting, we apply the proposed algorithm to estimate θ_i's by using a characterization of e-mixture[2] and a nonparametric estimation of divergence[3].

Nonnegative matrix factorization (NMF) is a method for dimension reduction, where data matrix X is approximated by a product of low rank matrices W and H, and all components of X, W, H are nonnegative. Letting Π be the column-wise L_1 normalization operator, $\Pi(X) = \Pi(W)\Pi(H)$ holds if $X = WH$. The normalized version of NMF is known as a topic model used in natural language processing. Since the normalized column can be regarded as a probability vector, the NMF is formulated as a fitting problem of an m-flat submanifold[4]. This problem can be solved by alternating e-projections. Exising methods of NMF[5] are numerically unstable when zero components are included in W or H because of the logarithm of zero. To avoid the unstability, we apply the proposed algorithm to estimate the matrices W and H.

Keywords: Pythagorian theorem, α-projection, mixture models, topic models

References

[1] Amari, S. (1985) *Differential-Geometrical Methods in Statistics*, Springer

[2] Murata, N., Fujimoto, Y. (2009) Bregman divergence and density integration, *Journal of Math for Industry*, 1, 97-104

[3] Hino, H., Murata, N. (2013) Information estimators for weighted observations, *Neural Networks*, 1, 260-275

[4] Akaho, S. (2004) The e-PCA and m-PCA: dimension reduction of parameters by information geometry, In *Proc. of IJCNN*, 129-134

[5] Sra, S., Dhillon, I. S. (2005) Generalized nonnegative matrix approximations with Bregman divergences. In *NIPS*, 283-290