一次近似の範囲で確率分布に依存しないICA

最上　嗣生

14 June 2003

I

ICA (independent component analysis) は多くの現実的応用で成功を収めている解析法である。外界に n 個の確率論的に独立な信号源 s_i があるとする。n 次元ベクトル x はこれの線形結合で書かれているときに元の s を復元する線形変換を発見する、これがICA である。(以下では読者がICA について知っているという前提で書かれている。もしICA に不案内なら「独立成分分析」(サイエンス社 2002) などを先に読むことをおすすめする。)

しかし現実の世界では独立な信号源が非線形な混合をおこしてセンサーに届くことは普通のことである。そのためには非線形に拡張したICA を行う必要がある。ICA の確率論的な基礎付けを行えばそれは最尤推定であることが分かる。最尤推定には信号源 s の確率分布が必要であるが、線形のICA の場合は幸運にもこの確率分布が問題にならないという性質があった。しかし非線形化したICA ではこの確率分布の推定が結果に影響をする。従って広汎な非線形応用の世界を開くためには確率分布を問題にしないアルゴリズムを開発するべきである。

本論では、アプリオリに仮定された確率分布 $q(y)$ の真の確率分布 $p(y)$ からの差の一次までの近似を、計算量を増大させることなく計算に取り入れる方法を新たに導入する。一次であっても確率分布関数形は無限次元であるのでこれは無限個の効果を一度に計算することになっている。ただし、復元の変換はやはり線形に限っている。本理論はもともと非線形を意図した物であるが、本論の最後に述べるように線形であってもいくつか難しい問題があったためにまずは線形について充分に理解する必要があるためである。

まず最初に、普通のICA がどのような物であったか見てみよう。信号源 s を変換 $y = f(x)$ で復元したいとする。そのためには信号源の確率分布 $q_i(s_i)$ をつかって相対エントロピー S_R を最小化するように f を選べばよい。相対エントロピーの定義は

$$S = \int d^n y \; p(y) \log p(y) - \int d^n y \; p(y) \log \prod_i q_i(y_i) \tag{1}$$

である。ここでyはn次元ベクトル、$p'(x)$はxの確率分布、$p(y)$はyの確率分布である。ここで $p(y) = \det(\partial x/\partial y)p'(x)$ を使って、Sの第一項は

$$\begin{aligned} &= \int d^n y \ \det \frac{\partial x}{\partial y} p'(x) \log \det \frac{\partial x}{\partial y} p'(x) \\ &= \int d^n y \ p(y) \log \det \frac{\partial x}{\partial y} + \int d^n x \ p'(x) \log p'(x) \end{aligned} \tag{2}$$

と変形でき、この２番目の項は関数 f に依存しないのでSを最小化するfを求める際には以下のS'を最小化すればよいので確率分布$p(x)$を知る必要はない。

$$S' = \int d^n y \ p(y) \log \det \frac{\partial x}{\partial y} - \int d^n y \ p(y) \log \prod_i q_i(y_i) \tag{3}$$

上記のSの積分の中に裸で現れる $p(y)$ はtをサンプルのインデクスとして $p(y) = \sum_t \delta(y - y_t)$ と置き換えることができる。

さらにf(x)を線形関数$y_i = \sum_j V_{ij} x_j$の範囲に限ると

$$S = \sum_t \log \det \left(\frac{\partial x}{\partial y} \right) - \sum_t \sum_i \log q_i(y_i) \tag{4}$$

$$\begin{aligned} \frac{\partial S}{\partial V_{ij}} &= -\sum_t V_{ij}^{-1} - \sum_t \sum_i \frac{1}{q_i} \frac{\partial q_i}{\partial y_i} \\ &= -V_{ji}^{-1} + \int d^n y \ p(y) \varphi_i(y_i) x_j \end{aligned} \tag{5}$$

となって簡単に降下方向を求めることができる。これが ICA (independent component analysis) である。ここで

$$\varphi_i(z) \equiv -\frac{1}{q_i(z)} \frac{dq_i(z)}{dz} \tag{6}$$

とした。

ここで問題は一般にはq_iの関数形が分からないことである。実際の出力のy_i方向の確率分布

$$p_i(y_i) = \int dy_1 \cdots dy_{i-1} dy_{i+1} \cdots dy_n \ p(y) \tag{7}$$

を使って逐次、分布q_iをこれで置き換えていけば良い。しかし線形の場合は理論的には$q_i(y_i)$がそれほど正確に$p_i(y_i)$の近似になっていなくともkurtosisの符号さえあっていれば正しく収束すると言われている。しかし、その場合は相対的に収束が遅くなる。また、もとの信号源が完全に独立な信号の積として書かれていない実世界データの場合に対してもうまくいかなくなることがありうる。

以上の$\log\prod_i p_i(y_i)$を含む第二項の評価の問題を何とか処理できないか考えてみる。この$p_i(y_i)$を$q_i(y_i)$によって近似することによって生じたSの誤差は以下の通りである。(以下、$\prod_i q_i(y_i) \equiv q(y)$と略する。また$q_i$は$q_i(y_i)$を$p_i$は$p_i(y_i)$を表すものとする。)

$$\begin{aligned}\Delta S &= \int dy\ p(y)\log\prod_i q_i(y_i) - \int dy\ p(y)\log\prod_i p_i(y_i) \\ &= \sum_t\{\log\prod_i q_i - \log\prod_i p_i\} \end{aligned} \tag{8}$$

以下ではこれを評価して、一次の近似においては$p(y)$は積分の中に裸で一次の項としてしか現れないような形にこれを書き直すことができることをしめす。これを使えば$p(y)$の連続分布形を知らずに学習を行うことができる。

関数fは学習で決定されるべきパラメータを持つ。それを抽象的にVと書く。明示的にVのインデックスを書きはしないがVは多次元でかまわない。

$$\begin{aligned}\frac{\partial\Delta S}{\partial V} &= \sum_t\sum_k \frac{\partial}{\partial y_k}\{\log\prod_i q_i(y_i) - \log\prod_i p_i(y_i)\}\times\frac{\partial y_k}{\partial V} \\ &= -\int dy\ p(y)\sum_k\left\{\frac{\partial}{\partial y_k}\sum_i\log\frac{p_i}{q_i}\right\}\times\frac{\partial y_k}{\partial V} \\ &= -\int dy\ p(y)\sum_i\left(\frac{\partial}{\partial y_i}\frac{p_i}{q_i}\right)\frac{q_i}{p_i}\times\frac{\partial y_k}{\partial V} \\ &= \int dy\sum_i\frac{p_i}{q_i}\frac{\partial}{\partial y_i}p(y)\frac{q_i}{p_i}\frac{\partial y_i}{\partial V} \end{aligned} \tag{9}$$

これに全微分の項を加える。

$$= \int dy\sum_i\left(\frac{p_i}{q_i}-1\right)\frac{\partial}{\partial y_i}p(y)\frac{q_i}{p_i}\frac{\partial y_i}{\partial V} \tag{10}$$

すでにこれは差$(p(y)-q(y))$の1次であるから∂y_i微分の中の$p(y)$は$q(y)$にq_i/p_iは1置き換えても1次の近似としては変わらない。

$$\sim \int dy\sum_i\left(\frac{p_i}{q_i}-1\right)\frac{\partial}{\partial y_i}q(y)\frac{\partial y_i}{\partial V} \tag{11}$$

ここでまた全微分を取り除く。

$$= \int dy\sum_i\frac{p_i}{q_i}\frac{\partial}{\partial y_i}q(y)\frac{\partial y_i}{\partial V} \tag{12}$$

さて、以下ではfが線形関数である場合を考える。

$$\frac{\partial\Delta S}{\partial V_{ij}} = \int dy\sum_k\frac{p_k}{q_k}\frac{\partial}{\partial y_k}q(y)(\delta_{ik}x_j)$$

$$
\begin{aligned}
&= \int dy \frac{p_i}{q_i} \frac{\partial}{\partial y_i} q(y) \sum_l V_{jl}^{-1} y_l \\
&= \int dy \frac{p_i}{q_i} q(y) \left\{ \frac{1}{q_i} \frac{\partial q_i}{\partial y_i} \sum_l V_{jl}^{-1} y_l + V_{ji}^{-1} \right\} \\
&= \int dy \frac{p_i}{q_i} q(y) \{ -\varphi_i(y_i) \sum_l V_{jl}^{-1} y_l + V_{ji}^{-1} \} \qquad (13)
\end{aligned}
$$

ここで全ての y_i の $p(y)$ による期待値は 0 であることを仮定した。線形の場合の良いところは $q(y) = \prod_i q_i(y_i)$ と次元ごとに分離しているために、 y_i 以外の y の次元を全て積分してしまうことができることである。したがって

$$
\frac{\partial \Delta S}{\partial V_{ij}} = -\int dy_i \; p_i(y_i) \varphi_i(y_i) (V_{ji}^{-1} y_i) + V_{ji}^{-1} \qquad (14)
$$

であるが、これはオンラインの学習の際に容易に計算することができる。この補正項の第二項は普通の学習則である式 (5) の第一項とキャンセルする。このようにして、少なくとも一次の近似においては出力の各次元の確率分布 p_i を知ることなく ICA を行うことができる。

以上より学習則は

$$
\Delta V_{ij} \propto -\frac{\partial (S + \Delta S)}{\partial V_{ij}} = \int dy_i \; p_i(y_i) \varphi_i(y_i) y_i V_{ji}^{-1} - \int dy \; p(y) \varphi_i(y_i) \sum_l y_l V_{jl}^{-1} \qquad (15)
$$

である。

さて、最後に問題点を指摘しておく。上記 (15) 式のように学習則は求まった。しかしこれを natural gradient の考え方と組み合わせるにはまだ研究が必要である。なぜならば、(15) 式は恒常的に無効な方向を持っている。(15) に右から V を掛けてみれば、行列の対角成分がつねにゼロである。これは次のような事情によるものと考えられる。仮定した確率分布に依存せず現下の確率分布を即時に反映するのであれば、たとえば確率分布の平行移動、確率分布のスケールなどの変換に対応する方向の相互情報量 S は完全にフラットである。そのためにその方向には常に勾配がゼロである。勾配がゼロであること自体は好ましいことであるが、このような平坦方向がある場合は Newton 法を工夫なしに利用することはできない。また natural gradient の考え方を導入するにしてもこのフラット方向をあらかじめ取り除いておかなくてはならない。たとえば上記 (15) 式に $V^T V$ を掛けるというナイーブな natural gradient の導入を行えば V の変化分はフラット方向の成分を持ってしまうが、それはこの理論の本来の目的に反する。

また (15) 式のフラット方向は単純にスケールとも言えないようであるし、その正体をつかみかねている。

またこれは必ずしも問題とは言えないが、(15) 式の学習則は微分されてから近似を受けているので必ず収束する保証がない。