Supplementary Materials for Submission "Boosting Independent Component Analysis"

1 Explanations for several equations

1.1 Modified maximum likelihood estimation

We don't use the direct logarithm with respect to equation (7) in our manuscript, we used the modified log-likelihood (Section II, Page 2, Column 1, Paragraph 3, Line 48:) from the past research [30][31]. We provide the details below to illustrate the correctness of our equation and comment.
We define the original log-likelihood as $\mathbb{L}(\mathbf{w}_i, f_i)$

$$\begin{aligned}
\mathbb{L}(\mathbf{w}_i, f_i) &= \sum_{j=1}^{N} \ln p_i(y_i^j) \\
&= \sum_{l=1}^{L} q_i^l \ln p_i(y_i^{*l}) \\
&= \sum_{l=1}^{L} q_i^l \left[f_i(y_i^{*l}) - \ln \sum_{l=1}^{L} \Delta_i e^{f_i(y_i^{*l})} \right] \\
&= \sum_{l=1}^{L} q_i^l f_i(y_i^{*l}) \; - \; \ln \sum_{l=1}^{L} \Delta_i e^{f_i(y_i^{*l})} \\
&\geq \sum_{l=1}^{L} q_i^l f_i(y_i^{*l}) \; - \; \sum_{l=1}^{L} \Delta_i e^{f_i(y_i^{*l})} + 1 \\
&= 1 + \sum_{l=1}^{L} q_i^l f_i(y_i^{*l}) - \Delta_i e^{f_i(y_i^{*l})} \\
&= 1 + \mathcal{L}(\mathbf{w}_i, f_i)
\end{aligned} \tag{1}$$

where the $\geq$ is due to the inequality $-\ln v \geq -v + 1$, and the equality is activate when $v = 1$, which means that maximum log-likelihood is obtained when the partition function $\sum_{l=1}^{L} \Delta_i e^{f_i(y_i^{*l})} = 1$. (the proof of equation (8) in our manuscript)

1.2 More details concerning smooth spline

a. the derivation of Line 14 in Algorithm 2 (more details see [30]):
considering equation (12)

$$\max_{\gamma_{i,k}} \quad \mathcal{L}(f_i^{k-1}(y_i) + b(y_i;\gamma_{i,k})) \tag{12}$$

the second-order approximation of equation (12) around $f_i^{k-1}(y_i)$ is

$$\begin{aligned}
&\mathcal{L}(f_i^k(y_i)) \\
&= \mathcal{L}(f_i^{k-1}(y_i) + b(y_i;\gamma_{i,k})) \\
&\approx \mathcal{L}(f_i^{k-1}(y_i)) + \sum_{l=1}^{L} \left(q_i^l - \Delta_i e^{f_i^{k-1}(y_i^{*l})}\right)\left(f_i^k(y_i^{*l}) - f_i^{k-1}(y_i^{*l})\right) \\
&+ \sum_{l=1}^{L} \frac{1}{2}\left(-\Delta_i e^{f_i^{k-1}(y_i^{*l})}\right)\left(f_i^k(y_i^{*l}) - f_i^{k-1}(y_i^{*l})\right)^2
\end{aligned} \tag{2}$$

Maximizing the above equation is equivalent to minimizing the iterative re-weighted least squares below (see Line 10-11, 14 in Algorithm 2):

$$\begin{aligned}
&\min_{\gamma_{i,k}} \sum_{l=1}^{L} \frac{1}{2}\omega_l^k (b_i(y_i^{*l};\gamma_{i,k}) - Y_l^k)^2 \\
&\omega_l^k = \Delta_i e^{f_i^{k-1}(y_i^{*l})} = \omega_l^{k-1} e^{b_i(y_i^{*l};\gamma_{i,k-1})} \\
&Y_l^k = \frac{q_l - \omega_l^k}{\omega_l^k}
\end{aligned} \tag{3}$$

b. connection between the Lagrange multiplier λ and the degree of freedom (more details see Chapter 5.4.1 in [39]):
λ is the hyper-parameter connected with df, and we can tune λ by simply determine the degree of freedom df beforehand (thus, df is the only hyper-parameter we need to choose for single weak learner $b(y_i;\gamma_{i,k})$).
considering the regularized weighted least squares in the Line 14 of Algorithm 2:

$$\sum_{l=1}^{L} \frac{1}{2}\omega_l^k (b_i(y_i^{*l};\gamma_{i,k}) - Y_l^k)^2 + \lambda J(b_i(y_i;\gamma_{i,k})) \tag{4}$$

According to the above equation, linear smoother S_λ is defined as follow,

$$\begin{bmatrix} b_i(y_i^{*1};\gamma_{i,k}) \\ \vdots \\ b_i(y_i^{*L};\gamma_{i,k}) \end{bmatrix} = S_\lambda \begin{bmatrix} Y_1^k \\ \vdots \\ Y_L^k \end{bmatrix} \tag{5}$$

the degree of freedom df is the trace of the linear smoother S_λ:

$$df = \mathrm{trace}(S_\lambda) \tag{6}$$

In summary, we tune df as the replace of λ, and the default value for df is 3 (slighter greater than than lower-bound 2) owing to the principle of boosting: the weak learner's model complexity should be simple enough.

1.3 Hyper-parameters tuning strategies

Although there are three hyper-parameters (L, M, df) in the proposed method, we can choose to consider the tuning of M (the number of boosting iterations) only in most cases.

- grid value L : the default value of L is 500, which might be proportional to the sample size (it is easy to tune L).

- degree of freedom df: the default value of df is 3, making the weak learner slighter better than the simple line fit (in most cases, we fixed df as default value).

- number of boosting iterations M: M can be regarded as the only hyper-parameter in our method, we can simply choose a large value for M due to the boosting's robust feature concerning M [40].

We provided the sensitivity analysis concerning M in our manuscript (Section III, Subsection C, Page 4, Column 2, Fig. 2), and we actually don't need to pay too much attention on the choice of grid value L (fix it as default value or proportional to the sample size).
In boosting community, people often fix the model complexity (df in the proposed method) of the weak learner, and tune the boosting iterations M. We provide a sensitivity analysis concerning df here.

2 Additional experiments

2.1 Sensitivity analysis experiment concerning df

We designed an images separation experiment with different df, where the three gray-scale images were chosen from the *ICS*[49] package. These images depict a forest road, cat and sheep, and they have been used in mangy ICA researches [49]. We vectorized them to arrive into a $130^2 \times 3$ data matrix and we fixed the mixing matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} 0.8 & 0.2 & 0.3 \\ 0.3 & -0.8 & 0.2 \\ -0.3 & 0.7 & 0.3 \end{bmatrix} \tag{7}$$

df increases from 2 to 100 (the weak learning principle breaks for large df), and the separation performance is measured with Amari metrics and SIR.

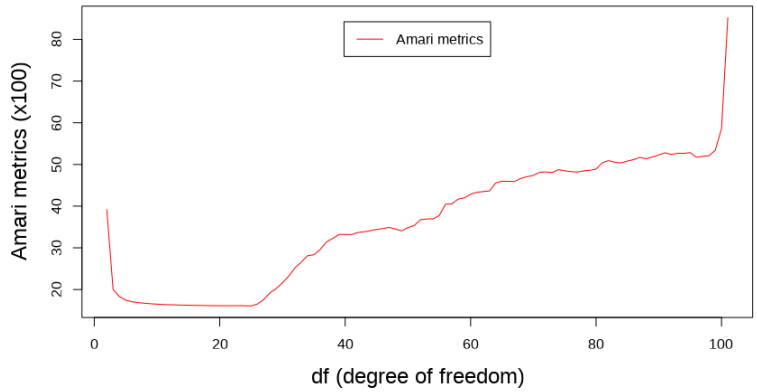

Figure 1: The Amari metrics of natural scene images separation with different degree of freedom. With the increase of df, Amari metrics firstly drops then increases, which indicates extremely large df deteriorates the separation performance.

As can be seen in Figure 1,2, with the increase of df, weak learners become more complex, which breaks the weak learning principle in boosting (leading to a bad separation performance for large M). So, we have no reason to select extremely large value for df, a value slightly greater than 2 is ok.
In conclusion, despite there are three hyper-parameters (L, df, M), the proposed method's hyper-parameter tuning burden is much small.

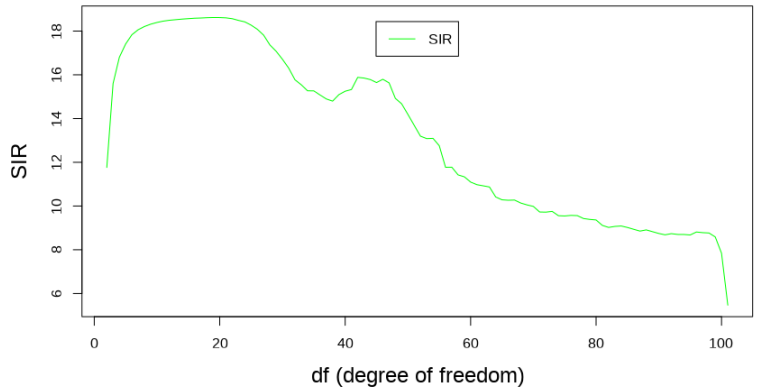

Figure 2: The SIR of natural scene images separation with different degree of freedom. With the increase of *df*, SIR firstly increases then drops, which indicates extremely large *df* deteriorates the separation performance.

2.2 Nature scene images separation with AWGN

We designed an images separation experiment with AWGN (Additive White Gaussian Noise), where the three gray-scale images were chosen from the *ICS*[49] package. These images depict a forest road, cat and sheep, and they have been used in mangy ICA researches [49]. We vectorized them to arrive into a $130^2 \times 3$ data matrix and we fixed the mixing matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} 0.8 & 0.2 & 0.3 \\ 0.3 & -0.8 & 0.2 \\ -0.3 & 0.7 & 0.3 \end{bmatrix} \tag{8}$$

The observed mixtures with additive noise is

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbf{I}) \tag{9}$$

where ϵ is the AWGN, σ^2 is the variance of noise and $\mathbf{I} \in \mathbb{R}^{m \times m}$. We changed the noise's variance from 0 to 1, and the separation performance is measured with Amari metrics and SIR. Since the proposed method can be roughly regarded as the lowerbound in Figure 3 and the upperbound in Figure 4, $BoostingICA$ actually enjoyed the great separation performance in this experiment (we omitted the comparison with $PICA$ and $FNA2$ in order to keep the Figures as clear as possible).

However, to be honest, we have to state that the proposed method is not particularly designed to tack the noisy cases for the following reasons:

- Our linear model in equation (1) is noiseless, and we actually do not take the observation noise into consideration.
- Since there always exists the whitening preprocessing for linear ICA , the separation performance with noise might deteriorate (taking the noise into the model's design and matrix diagonalization might be appropriate strategies to tack noiseICA).

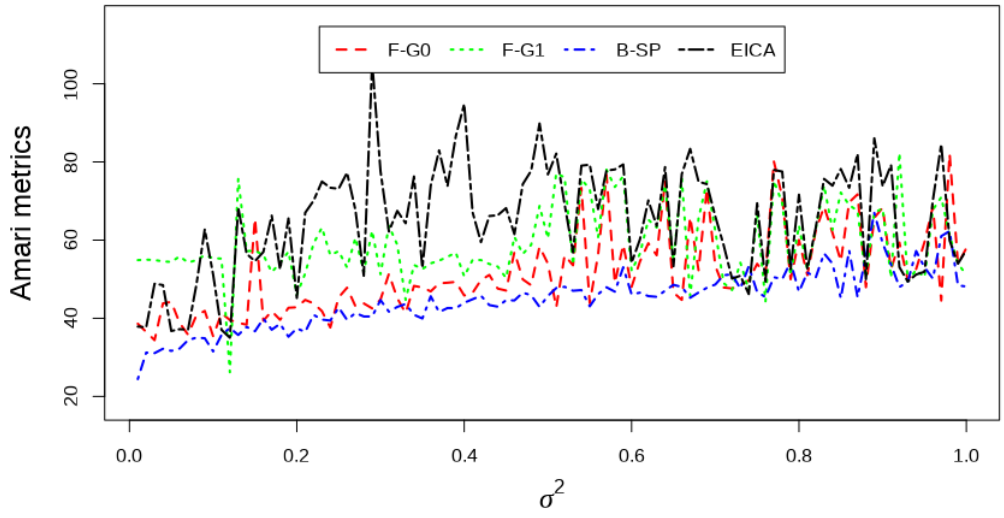

Figure 3: The Amari metrics of noisy natural scene images separation task with different AWGN's variance σ^2. Our method's (in blue)Amari metrics is less than other ICA methods' in most cases.

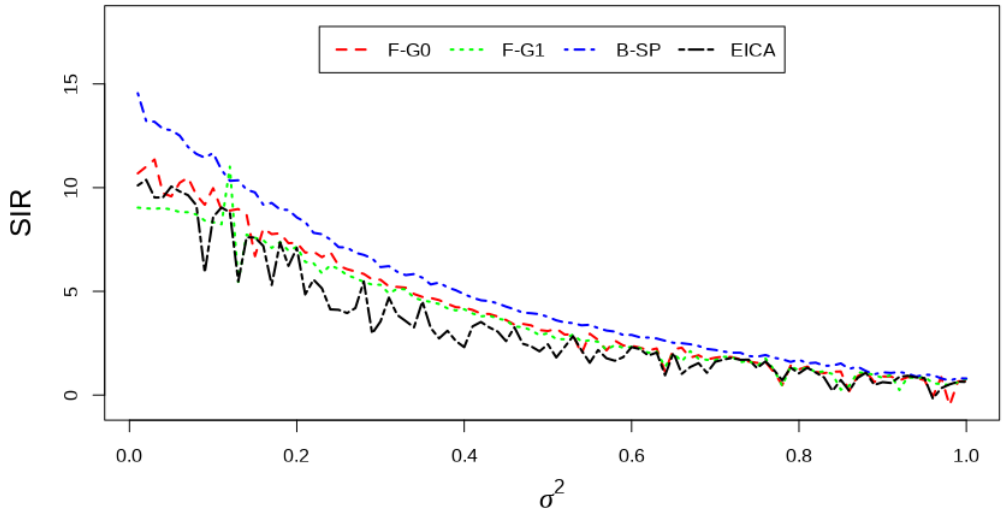

Figure 4: The SIR of noisy natural scene images separation task with different AWGN's variance σ^2. Our method's (in blue) SIR is greater than other ICA methods' SIR.