

Projet M2 Data Science : ICA et Optimisation Stochastique

Analyse Comparative et Extensions Créatives - 2026

Christophe Ambroise

2026-01-01

Contexte et Objectifs

Motivation

L'Analyse en Composantes Indépendantes (ICA) est traditionnellement résolue par des algorithmes déterministes (FastICA, Infomax avec descente de gradient classique). L'optimisation stochastique permet à l'ICA, notamment de traiter des données massives ou en flux.

Objectifs

Vous devrez :

1. **Implémenter** plusieurs variantes d'algorithmes ICA basés sur le gradient stochastique
2. **Concevoir des expériences** originales pour évaluer ces méthodes
3. **Proposer une contribution créative** : nouvelle variante, application, ou analyse

IMPORTANT : Ce projet nécessite une réflexion personnelle sur les choix algorithmiques et expérimentaux. Les LLMs peuvent vous aider pour le code de base, mais **la démarche scientifique, les choix méthodologiques, et l'analyse critique doivent être les vôtres.**

Liens avec le cours

Ce projet s'appuie sur les chapitres suivants du cours :

- **ICA** : modèle, hypothèses d'indépendance, non-gaussianité
- **Practical online learning** : descente de gradient stochastique, règle de Oja, conditions de Robbins-Monro
- **VAE** (pour la piste ICA+VAE) : ELBO, reparamétrisation, décodeur

Rappel : la log-vraisemblance ICA

Le modèle ICA s'écrit $x = Wz$ avec z un vecteur de sources indépendantes et W la matrice de mélange (generative weights). En posant $V = W^{-1}$ la matrice de démélange (recognition weights), la log-vraisemblance sur la population est :

$$\ell(V) = \log |\det V| + \mathbb{E} \left[\sum_{j=1}^D \log p_j(y_j) \right], \quad y_j = v_j^\top x$$

où p_j est la densité de la j -ème source. La **fonction score** (dérivée de la log-densité) est notée :

$$g_j(y_j) = \frac{d}{dy_j} \log p_j(y_j)$$

Le choix de g (ou de manière équivalente du contraste G avec $g = G'$) est un élément clé de l'algorithme. Par exemple, pour des sources super-gaussiennes : $G(u) = \log \cosh(u)$ et $g(u) = \tanh(u)$.

Partie 1 : Implémentation de Base

Algorithmes à Implémenter

Vous devez implémenter **au minimum 3 algorithmes** parmi :

Algorithmes Classiques (référence)

1. **Infomax avec gradient batch** (Bell & Sejnowski, 1995)
2. **FastICA** (Hyvärinen, 1999) — pour comparaison. Vous pouvez utiliser l'implémentation de `sklearn.decomposition.FastICA` comme baseline ; une implémentation from scratch est un bonus.

Variantes Stochastiques (obligatoires, au moins une)

3. **SGD-ICA** : descente de gradient stochastique classique appliquée au critère Infomax
4. **Adam-ICA** : optimiseur Adam appliqué au critère Infomax

Variante Avancée (optionnelle, bonus)

5. **Natural Gradient SGD** : gradient naturel stochastique (Amari, 1998). La mise à jour prend la forme :

$$V \leftarrow V + \eta (I + g(y)y^T) V$$

où $y = Vx$ et g est la fonction score (dérivée de la log-densité), par exemple $g(y) = \tanh(y)$ pour des sources super-gaussiennes. Cette formulation exploite la géométrie de l'information (matrice de Fisher) pour accélérer la convergence.

Critères d'Évaluation Quantitatifs pour les simulations

Pour mesurer la qualité de la séparation, vous utiliserez l'**indice d'Amari** (Amari et al., 1996). La matrice V (estimée par ICA) est censée être l'inverse de W (matrice de mélange, que vous connaissez si vous simulez), donc idéalement une matrice de permutation-dilatation.

L'indice d'Amari mesure à quel point la matrice de performance $C = VW$ s'écarte de cette structure idéale.

Soit $C = VW$ la matrice de performance (idéalement une matrice de permutation-dilatation):

$$r(C) = \frac{1}{2D(D-1)} \left[\sum_{i=1}^D \left(\sum_{j=1}^D \frac{|c_{ij}|}{\max_k |c_{ik}|} - 1 \right) + \sum_{j=1}^D \left(\sum_{i=1}^D \frac{|c_{ij}|}{\max_k |c_{kj}|} - 1 \right) \right]$$

Cet indice vaut 0 pour une séparation parfaite et tend vers 1 pour une mauvaise séparation. Reportez systématiquement cet indice (moyenné sur plusieurs runs avec barres d'erreur) dans vos expériences.

Critères d'Évaluation du Code

- **Correction mathématique** : implémentation fidèle des algorithmes
- **Modularité du code** : séparation claire entre fonction objectif, gradient, optimiseur
- **Validation** : tests sur données synthétiques avec sources connues
- **Documentation** : code commenté et reproductible

Partie 2 : Contribution Créative

Cette partie est **entièrement libre**. Vous devez proposer et réaliser une extension originale. Quelques pistes (non exhaustives car vous pouvez tout à fait proposer votre propre idée) :

Pistes Algorithmiques

1. **ICA avec algorithme EM** : implémentez une version stochastique de l'EM pour ICA.
2. **ICA et VAE** : utilisez un autoencodeur variationnel pour apprendre un modèle de source non-gaussien. L'idée est de remplacer le choix fixe de la fonction score g par un modèle appris des densités sources p_j . Référence : Khemakhem, I., Kingma, D., Monti, R., & Hyvärinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *AISTATS*.

Pistes Applicatives

3. **Inférence causale avec LinGAM** : ICA est à la base de LinGAM (Shimizu et al., 2006) pour l'inférence causale linéaire. Implémentez une version stochastique de LinGAM et testez-la sur des données synthétiques avec des graphes causaux connus.
4. **ICA pour données manquantes** : adaptez SGD-ICA quand certaines observations sont masquées. Une approche possible est d'optimiser la vraisemblance uniquement sur les composantes observées à chaque itération. Référence : Podosinnikova, A., Perry, A., Wein, A. S., Bach, F., d'Aspremont, A., & Sontag, D. (2019). Overcomplete Independent Component Analysis via SDP. *AISTATS*.

Datasets suggérés

Au-delà des données synthétiques, testez vos algorithmes sur des données réelles :

- **Séparation de sources audio** : signaux de type “cocktail party” (disponibles par exemple dans le package `scikit-learn` ou sur [SISEC](#))
- **EEG** : signaux électroencéphalographiques, où l’ICA est utilisée pour séparer les artefacts oculaires et musculaires des signaux cérébraux (datasets MNE-Python)
- **Images** : séparation de textures superposées

Critères d’Évaluation de la Contribution

- **Originalité** : est-ce une idée nouvelle ou une adaptation non triviale ?
- **Rigueur** : la démarche est-elle bien justifiée et testée ?
- **Présentation** : clarté de l’explication (crucial pour la soutenance)

Conseil : Une contribution simple mais bien exécutée vaut mieux qu’une idée ambitieuse mal aboutie.

Livrables et Évaluation

Rapport Écrit (50% de la note finale)

Format : 10 pages maximum (hors code et annexes), PDF. LaTeX ou Quarto recommandé.

Soumission : un dépôt Git (GitHub/GitLab) contenant le rapport (PDF), le code source, et un README expliquant comment reproduire les résultats.

Structure attendue :

1. **Introduction** : contexte, objectifs, positionnement de votre contribution
2. **Méthodes** : description des algorithmes implémentés, choix de la non-linéarité, hyperparamètres
3. **Expériences** : protocole expérimental, datasets, métriques (indice d’Amari), résultats avec barres d’erreur
4. **Discussion** : analyse critique des résultats, limites identifiées, comparaison entre méthodes
5. **Conclusion** : résumé des apports et perspectives

Soutenance Orale (50% de la note finale)

Durée : 7 minutes de présentation + 8 minutes de questions

Ressources et Contraintes

Bibliographie de Départ

ICA classique :

- Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis*. Wiley.
- Bell, A.J., & Sejnowski, T.J. (1995). An information-maximization approach to blind separation. *Neural Computation*, 7(6), 1129-1159.
- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276.

Optimisation stochastique :

- Bottou, L., Curtis, F.E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2), 223-311.
- Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *NIPS*.
- Kingma, D.P., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Gradient naturel :

- Martens, J. (2020). New insights and perspectives on the natural gradient method. *JMLR*, 21(146), 1-76.

ICA et apprentissage profond :

- Khemakhem, I., Kingma, D., Monti, R., & Hyvärinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *AISTATS*.

Inférence causale :

- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 7, 2003-2030.

Bonnes Pratiques Expérimentales

- **Fixez les seeds** : résultats reproductibles
- **Moyennez sur plusieurs runs** : avec barres d'erreur/écart-type
- **Testez sur plusieurs datasets** : la généralité est importante
- **Utilisez des baselines** : comparez toujours à FastICA ou Infomax batch
- **Documentez vos choix** : pourquoi ce learning rate ? Cette initialisation ?
- **Reportez l'indice d'Amari** : c'est la métrique quantitative de référence

Critères de Notation Détaillés

Rapport (50%)

Critère	Points	Description
Clarté rédactionnelle	8	Structure, langue, figures
Rigueur scientifique	12	Protocole, validation, reproductibilité
Qualité de l'implémentation	10	Correction, modularité, tests
Contribution créative	20	Nouveauté, aboutissement

Soutenance (50%)

Critère	Points	Description
Clarté de présentation	15	Slides, pédagogie, timing
Maîtrise technique	20	Compréhension profonde, réponses aux questions
Esprit critique	15	Limites identifiées, choix justifiés

Bonus : Jusqu'à 5 points pour une contribution exceptionnellement originale ou une présentation remarquable.

En guise de conclusion

Ce projet vise à développer votre **autonomie scientifique**. Vous allez :

- Implémenter des algorithmes à partir de papiers de recherche

- Proposer et tester vos propres idées
- Communiquer vos résultats à l'écrit et à l'oral

Le plus important n'est pas d'avoir les “meilleurs” résultats, mais de démontrer une démarche rigoureuse, critique, et créative.

Date limite de rendu du rapport : 5 avril 2026 à minuit

Date des soutenances : 9 avril 2026

Contact : christophe.ambroise@univ-evry.fr