Modèles augmentés asymptotiquement exacts

Maxime VONO[1], Nicolas DOBIGEON[1], Pierre CHAINAIS[2]

[1]Université de Toulouse, IRIT/INP-ENSEEIHT, Toulouse, France

[2]Université de Lille, Centrale Lille, UMR CNRS 9189 - CRIStAL, Lille, France

maxime.vono@irit.fr, nicolas.dobigeon@enseeiht.fr, pierre.chainais@centralelille.fr

Résumé – L'introduction de variables auxiliaires dans un modèle statistique est communément utilisée afin de simplifier une tâche d'inférence ou augmenter son efficacité. Cependant, l'introduction de ces variables telles que la distribution de probabilité initiale soit préservée relève bien souvent d'un art subtil. Cet article présente un cadre statistique unificateur permettant de lever ces verrous en relâchant l'hypothèse d'augmentation exacte. Ce cadre, appelé *asymptotically exact data augmentation* (AXDA), regroupe certains modèles de mélange, les modèles bayésiens robustes ou encore ceux construits à partir du *splitting* de variables. Afin d'illustrer l'intérêt d'une telle approche, un échantillonneur de Gibbs basé sur un modèle AXDA est présenté.

Abstract – Introducing auxiliary variables to augment an initial statistical model has been widely used to simplify or improve an inference task. However, introducing such variables such that the initial probability distribution is preserved may be a difficult art. To cope with these issues, this paper introduces a unifying statistical framework, called asymptotically exact data augmentation (AXDA), gathering well-established (e.g., robust or mixture-based) but also more recent (e.g., variable splitting-based) models. The benefits of using such an approach are illustrated with a special instance of AXDA-based algorithms, namely the split-and-augmented Gibbs sampler.

1 Introduction

L'introduction de variables auxiliaires au sein d'un modèle statistique initial est une stratégie largement adoptée afin de simplifier et/ou améliorer une tâche d'inférence. L'utilisation de tels modèles augmentés remonte au moins aux travaux sur l'algorithme *expectation-maximization* (EM) [1] et a été démocratisée en statistique [2] et physique statistique [3]. Malheureusement, bien que ces modèles se sont révélés efficaces, ils souffrent d'une hypothèse difficile à satisfaire dans le cas général, à savoir la préservation de la distribution de probabilité initiale [4]. De manière semblable aux méthodes *approximate Bayesian computation* (ABC) [5], ce verrou peut être levé en ayant recours à un modèle approché asymptotiquement exact. Prenant inspiration et étendant les travaux de [6, 7], cet article présente un cadre statistique unificateur, appelé *asymptotically exact data augmentation* (AXDA) permettant de s'affranchir de l'hypothèse d'augmentation exacte tout en gardant ses principaux avantages. AXDA inclut en effet certains modèles de mélange, des modèles bayésiens robustes [8] ou encore ceux construits à partir du *splitting* de variables utilisé en optimisation [6]. Il est à noter qu'AXDA peut être vu comme la contrepartie d'ABC au sein d'un problème d'inférence : les deux méthodes reposent sur une stratégie de *splitting* bien choisie amenant à un modèle approché. Afin de contrôler cette approximation, des garanties non-asymptotiques sont données sous des hypothèses facilement vérifiables en pratique.

Pour ce faire, la partie 2 présente les motivations et les modèles associés à l'approche AXDA. La partie 3 donne les propriétés générales et non-asymptotiques d'un modèle AXDA. La partie 4 présente un algorithme de Gibbs basé sur un modèle AXDA. Ses avantages par rapport à une approche considérant directement la distribution initiale ou à un modèle d'augmentation exact sont discutés. Ce travail s'appuie sur les contributions présentées dans [9] où les preuves des résultats présentés dans ce papier sont détaillées.

2 Modèles augmentés asymptotiquement exacts

Dans cet article, nous supposons que l'inférence de paramètres d'intérêt repose sur la distribution de probabilité ayant pour densité (par rapport à la mesure de Lebesgue)

$$\pi(\mathbf{x}) \propto \exp\left[-f(\mathbf{x})\right], \tag{1}$$

où $f : \mathcal{X} \to (-\infty, +\infty]$ est telle que $\pi \in L^1$. Notons que pour des raisons de simplicité, la même notation sera utilisée pour une distribution de probabilité et sa densité associée. Par souci de généralité et avec un léger abus de notations, π pourra faire référence tout au long de ce papier à une distribution a priori $\pi(\mathbf{x})$, une vraisemblance $\pi(\mathbf{y}|\mathbf{x})$ ou à une distribution a posteriori $\pi(\mathbf{x}|\mathbf{y})$.

2.1 Motivations

Si l'inférence basée sur (1) ne peut pas être conduite de manière efficace voire est impossible avec les méthodes existantes, une alternative consiste à introduire des variables auxiliaires

$\mathbf{z} \in \mathcal{Z}$ telles que la densité augmentée associée soit plus simple à manipuler et satisfasse

$$\int_{\mathcal{Z}} p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} = \pi(\mathbf{x}). \tag{2}$$

Cette condition assure que la densité marginale sous p soit la densité initiale π. Comme indiqué en partie 1, satisfaire (2) peut être difficile voire même impossible. De plus, même si un tel schéma augmenté est trouvé, l'inférence basée sur p peut rester très coûteuse pour des problèmes en grande dimension et avec un grand nombre d'observations [10]. Afin de lever ces verrous, nous proposons de relâcher la condition (2) de telle sorte que celle-ci ne soit satisfaite exactement que dans un cas limite.

2.2 Modèle

Ainsi, nous considérons l'introduction de variables auxiliaires $\mathbf{z}$ amenant à travailler avec une densité augmentée p_ρ, où $\rho > 0$, dont la marginale

$$\pi_\rho(\mathbf{x}) = \int_{\mathcal{Z}} p_\rho(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} \tag{3}$$

satisfait l'hypothèse 1 ci-dessous.

Hypothèse 1 *Pour tout* $\mathbf{x} \in \mathcal{X}$, $\lim_{\rho \to 0} \pi_\rho(\mathbf{x}) = \pi(\mathbf{x})$.

Cette hypothèse suppose de faire tendre un paramètre ρ strictement positif vers 0 afin de retrouver π. Nous pouvons également satisfaire une augmentation de modèle asymptotiquement exacte en faisant tendre le nombre d'observations vers l'infini via le théorème de Bernstein-von Mises [9]. Dans la suite, nous nous référerons aux modèles AXDA via la définition 1.

Définition 1 *Un modèle fait partie de la famille des modèles augmentés asymptotiquement exacts (AXDA) par rapport à une densité π s'il satisfait l'hypothèse 1.*

Motivés par l'obtention de méthodes d'inférence simples, rapides et efficaces, ainsi que par les travaux de [6, 7], nous supposons que p_ρ est définie par

$$p_\rho(\mathbf{x}, \mathbf{z}) \propto \exp(-f(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z})), \tag{4}$$

où ϕ_ρ représente une divergence mesurant la différence entre le paramètre d'intérêt $\mathbf{x}$ et la variable auxiliaire $\mathbf{z}$ et telle que (4) soit bien définie. La partie 4 illustrera comment le fait de travailler avec le modèle augmenté p_ρ simplifie l'inférence. Nous présentons ci-dessous deux exemples de densités augmentées p_ρ satisfaisant l'hypothèse 1.

Example 1 : Lissage par noyau gaussien – Soit ϕ_ρ un potentiel associé au noyau gaussien $K_\rho(\mathbf{x} - \mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{x}, \rho^2 \mathbf{I})$. Alors π_ρ correspond au lissage de la densité π par un noyau gaussien de variance ρ^2 :

$$\pi_\rho(\mathbf{x}) = \int_{\mathcal{Z}} \pi(\mathbf{z}) K_\rho(\mathbf{x} - \mathbf{z}) \mathrm{d}\mathbf{z}. \tag{5}$$

De plus, π_ρ satisfait l'hypothèse 1 [6].

Example 2 : Mélange Gamma-Poisson – Soit $\pi(\mathbf{x}) \triangleq \pi(\mathbf{y}|\mathbf{x})$ une densité associée à la distribution de Poisson $\mathcal{P}(\mathbf{y}; \mathbf{x})$ et ϕ_ρ un potentiel associé à une distribution gamma $\mathcal{G}(\mathbf{z}; \rho^{-1}, \mathbf{x})$. Alors π_ρ correspond à la distribution négative binomiale et satisfait l'hypothèse 1.

3 Propriétés

L'inférence basée sur un modèle de type AXDA étant approchée par rapport au modèle initial, il apparait important de pouvoir donner des garanties théoriques sur cette approximation. C'est précisément l'objet de cette partie.

3.1 Propriétés générales

Avant de donner des garanties non-asymptotiques, nous caractérisons tout d'abord le comportement asymptotique et la forme de la densité approchée π_ρ.

Théorème 1 *Sous l'hypothèse 1, il vient*

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \xrightarrow[\rho \to 0]{} 0. \tag{6}$$

Proposition 1 *Supposons que π soit log-concave. Soit $\phi_\rho(\mathbf{x}, \mathbf{z}) = \tilde{\phi}_\rho(\mathbf{x} - \mathbf{z})$ tel que $K_\rho \propto \exp(-\tilde{\phi}_\rho)$ soit log-concave, $\mathcal{C}^\infty$ et que pour tout $k \geq 0$, $\partial^k K_\rho$ soit bornée. Supposons également que $\lim_{\rho \to 0} K_\rho(\mathbf{u}) = \delta(\mathbf{u})$ et $\mathbb{E}_{K_\rho}(U) = 0$. Alors, π_ρ a les propriétés suivantes :*

i) l'hypothèse 1 est satisfaite ;

ii) π_ρ est log-concave ;

iii) π_ρ est $\mathcal{C}^\infty$ sur $\mathcal{X}$;

iv) Lorsque π est une loi de $\mathcal{X}$,

$$\mathbb{E}_{\pi_\rho}(X) = \mathbb{E}_\pi(X) \tag{7}$$

$$\mathrm{var}_{\pi_\rho}(X) = \mathrm{var}_\pi(X) + \mathrm{var}_{K_\rho}(X). \tag{8}$$

Les hypothèses de la proposition 1 sont par exemple vérifiées pour le noyau gaussien K_ρ. La propriété *iv)* implique que si π correspond à une distribution a priori autour d'une moyenne μ, π_ρ aura également pour moyenne μ mais une variance plus importante. Cette propriété est directement liée à la robustesse pouvant être apportée par les modèles de type AXDA [9].

3.2 Propriétés non-asymptotiques

Motivés par les propriétés de la proposition 1, par l'équivalence forte entre l'*alternating direction method of multipliers* (ADMM) et l'algorithme de Gibbs dérivé d'AXDA [6], nous faisons l'hypothèse suivante.

Hypothèse 2 *Le potentiel f est L_f-Lipschitz et pour tout $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, $\phi_\rho(\mathbf{x}, \mathbf{z}) = \dfrac{1}{2\rho^2} \|\mathbf{x} - \mathbf{z}\|_2^2$.*

Sous cette hypothèse, la distance en variation totale (TV) entre π_ρ et π peut être contrôlée par le théorème suivant.

Théorème 2 *Sous l'hypothèse 2, il vient pour tout $\rho > 0$,*

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq 1 - \Delta_d(\rho), \tag{9}$$

où $\Delta_d(\rho) = D_{-d}(L_f \rho) / D_{-d}(-L_f \rho)$ et $d = \dim(\mathcal{X})$. La fonction D_{-d} est la fonction spéciale cylindre parabolique définie pour tout $d > 0$ et $z \in \mathbb{R}$ par

$$D_{-d}(z) = \frac{\exp(-z^2/4)}{\Gamma(d)} \int_0^{+\infty} e^{-xz - x^2/2} x^{d-1} \mathrm{d}x. \tag{10}$$

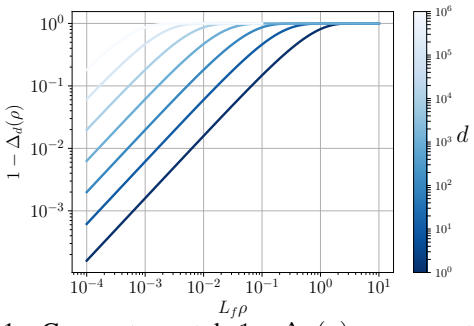

FIGURE 1 – Comportement de $1-\Delta_d(\rho)$ par rapport à $L_f\rho$ en échelle log-log.

Lorsque ρ est suffisamment petit, on montre que la dépendance de cette borne par rapport à ρ est linéaire.

Proposition 2

$$1-\Delta_d(\rho) \underset{\rho\to 0}{\sim} \frac{2\sqrt{2}\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} L_f\rho. \tag{11}$$

A noter qu'utiliser des formules de type Stirling dans l'équivalent (11) amène à une dépendance par rapport à la dimension du problème d de l'ordre de $\mathcal{O}(L_f d)$. Lorsque $f = \|\cdot\|_p$, $0 < p < 2$, cette dépendance est de l'ordre de $\mathcal{O}(d^{1/2+1/p})$ qui devient sous-linéaire pour $p \in (1,2)$. Le comportement de la borne $\Delta_d(\rho)$ est illustré par la figure 1. En utilisant la preuve du theorème 2, il est également possible de donner un contrôle sur les potentiels et intervalles de crédibilité, voir les propositions 3 et 4 ci-dessous. Le contrôle de ces derniers est très important lorsque l'inférence s'effectue sous le paradigme bayésien.

Proposition 3 *Soit* $f_\rho(\mathbf{x}) = d/2\log(2\pi\rho^2) - \log\int_{\mathcal{Z}} \exp(-f(\mathbf{z}) - (2\rho^2)^{-1}\|\mathbf{z}-\mathbf{x}\|_2^2)\mathrm{d}\mathbf{z}$, *le potentiel associé à* π_ρ. *Alors, sous l'hypothèse 2, pour tout* $\mathbf{x}\in\mathcal{X}$ *et* $\rho > 0$,

$$L_\rho \le f_\rho(\mathbf{x}) - f(\mathbf{x}) \le U_\rho, \tag{12}$$

avec

$$L_\rho = \log M_\rho - \log D_{-d}(-L_f\rho) \tag{13}$$
$$U_\rho = \log M_\rho - \log D_{-d}(L_f\rho) \tag{14}$$
$$M_\rho = \frac{2^{d/2-1}\Gamma(d/2)}{\Gamma(d)\exp\left(L_f^2\rho^2/4\right)}. \tag{15}$$

Une illustration de la proposition 3 est donnée par la figure 2 pour des potentiels f Lipschitz communément utilisés en apprentissage statistique (hinge, Huber, logistic, pinball). Il est à noter que la propriété *iii)* de la proposition 1 est vérifiée et que l'approximation f_ρ partage certaines similarités avec l'enveloppe de Moreau-Yoshida comme la convexité ou encore la différentiabilité.

Proposition 4 *Soit* π *une distribution a posteriori associée à* $\mathbf{x}$. *Soit* $\mathcal{C}_\alpha^\rho$ *un intervalle de crédibilité arbitraire de niveau de crédibilité* $(1-\alpha)$ *sous* π_ρ, *i.e.,* $\mathbb{P}_{\pi_\rho}(\mathbf{x}\in\mathcal{C}_\alpha^\rho) = 1-\alpha$ *avec* $\alpha\in(0,1)$. *Alors, sous l'hypothèse 2 et pour tout* $\rho > 0$,

$$\frac{(1-\alpha)M_\rho}{D_{-d}(-L_f\rho)} \le \int_{\mathcal{C}_\alpha^\rho}\pi(\mathbf{x})\mathrm{d}\mathbf{x} \le \min\left(1, \frac{(1-\alpha)M_\rho}{D_{-d}(L_f\rho)}\right).$$

Lorsque $\rho\to 0$, les bornes ci-dessus tendent bien vers $(1-\alpha)$, le niveau de crédibilité correspondant au modèle initial.

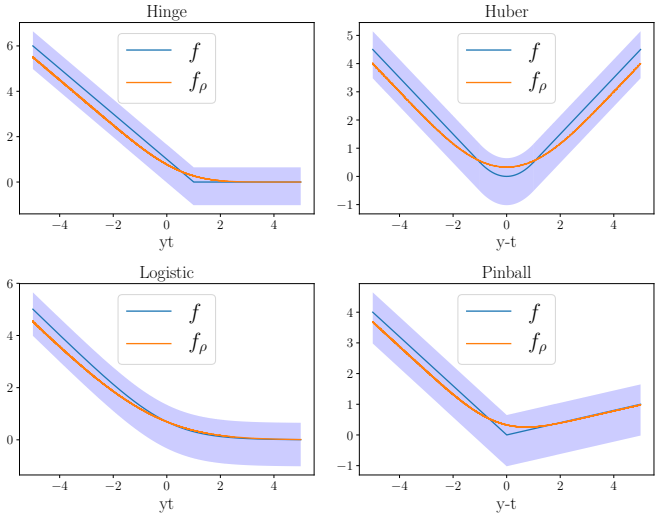

FIGURE 2 – Fonctions de coût Lipschitz f et leurs approximations f_ρ avec $\rho = 1$ estimées par une approximation Monte Carlo. Les contours de la zone bleue voilée correspondent à $f + L_\rho$ et $f + U_\rho$.

4 Algorithme de Gibbs basé sur AXDA

Cette partie présente une instance particulière basée sur un modèle AXDA, à savoir le *split-and-augmented Gibbs sampler* (SPA) proposé dans [6]. Cet algorithme a l'avantage de diviser et simplifier le problème d'échantillonnage initial, permettant ainsi de le distribuer et de l'accélérer.

4.1 Split-and-augmented Gibbs sampler

Nous supposons ici que le potentiel de la densité initiale π dans (1) s'écrit $f = \sum_{j=0}^J f^{(j)}$. Cette forme générale couvre un grand nombre de problèmes d'inférence rencontrés sous le paradigme bayésien. Par exemple, π correspond à une loi a posteriori $\pi(\mathbf{x}|\mathbf{y})$ lorsque $f^{(0)}$ est associée à une distribution a priori et $\sum_{j=1}^J f^{(j)}$ à une vraisemblance (e.g., produit de vraisemblances de J blocs d'observations). En introduisant J variables auxiliaires $\mathbf{z}_{1:J}$, le modèle AXDA correspondant cible la distribution de densité

$$p_\rho(\mathbf{x},\mathbf{z}|\mathbf{y}) \propto \exp(-f^{(0)}(\mathbf{x}) - \sum_{j=1}^J f^{(j)}(\mathbf{z}_j;\mathbf{y}) + \phi_\rho^{(j)}(\mathbf{x},\mathbf{z}_j)). \tag{16}$$

Un moyen naturel d'échantillonner selon cette densité jointe est de considérer un algorithme de Gibbs qui cible chaque conditionnelle de manière alternée. Les conditionnelles sous (16)

s'écrivent pour $j \in [\![1, J]\!]$

$$p_\rho(\mathbf{x}|\mathbf{z}_{1:J}, \mathbf{y}) \propto \exp\left(-f^{(0)}(\mathbf{x}) - \sum_{j=1}^{J} \phi_\rho^{(j)}(\mathbf{x}, \mathbf{z}_j)\right) \quad (17)$$

$$p_\rho(\mathbf{z}_j|\mathbf{x}, \mathbf{y}) \propto \exp\left(-f^{(j)}(\mathbf{z}_j; \mathbf{y}) - \phi_\rho^{(j)}(\mathbf{x}, \mathbf{z}_j)\right). \quad (18)$$

Les avantages liés à cet échantillonneur de Gibbs sont multiples. Premièrement, les conditionnelles sont a priori plus simples à échantillonner puisque le potentiel initial composite a maintenant été divisé en J contreparties. Si $\phi_\rho^{(j)}$ est bien choisie (e.g., telle que l'hypothèse 2 est satisfaite), chacune des conditionnelles (18) peut maintenant être échantillonnée directement, avec un schéma d'augmentation exacte ou encore avec des algorithmes de simulation plus sophistiqués en fonction de la difficulté restante après splitting. Bien que non évoqué dans ce papier, un intérêt supplémentaire d'un tel algorithme de Gibbs est de pouvoir enlever un opérateur (e.g., linéaire) gênant pouvant agir sur $\mathbf{x}$ dans un des $f^{(j)}$. Enfin, le fait de diviser le potentiel initial de cette manière permet de distribuer l'inférence sur plusieurs noeuds de calculs en parallèle. Cet aspect est particulièrement important si le jeu de données ne peut pas être stocké sur une seule machine mais est distribué sur J noeuds. Dans ce cas, l'échantillonnage de chaque conditionnelle (18) peut être conduit indépendamment sur chaque noeud sachant l'itérée courante $\mathbf{x}$ [7]. Le noeud central s'occupant de l'échantillonnage de la conditionnelle (17) de $\mathbf{x}$ n'a ainsi pas besoin d'accéder aux données $\mathbf{y}$ permettant le respect de la confidentialité de ces dernières. Notons que comme indiqué dans [6, 7, 9], les étapes de l'algorithme de Gibbs considéré sont intimement liées aux itérations de l'ADMM [11] en optimisation.

4.2 Illustration

En plus de pouvoir simplifier et distribuer l'inférence dans le cas d'une distribution compliquée, une approche AXDA peut aussi accélérer la convergence d'algorithmes de l'état de l'art comme les algorithmes de Monte Carlo proximaux [12, 13]. Pour illustrer ce point, nous considérons un problème de déconvolution et de débruitage classique où la distribution a posteriori $\pi(\mathbf{x}|\mathbf{y})$ a pour potentiel $f = f^{(0)} + f^{(1)}$ avec $f^{(0)} = \lambda \mathrm{TV}(\mathbf{x})$ et $f^{(1)}(\mathbf{x}; \mathbf{y}) = (2\sigma^2)^{-1} \|\mathbf{H}\mathbf{x} - \mathbf{y}\|_2^2$ où $\mathbf{y}$ est une image 256^2 floutée et bruitée ; TV est la fonction variation totale, convexe mais non différentiable. La fonction de couplage ϕ_ρ est choisie gaussienne, conforme à l'hypothèse 2. La figure 3 illustre la convergence de l'algorithme de Gibbs du modèle AXDA pour différentes valeurs du paramètre ρ : elle est en général plus rapide que celle de l'algorithme MYULA [13].

5 Conclusion

Ce papier a présenté un cadre unificateur appelé AXDA permettant de lever les verrous associés à une augmentation exacte en proposant une manière quasi-systématique d'augmenter le modèle initial. En plus de bénéficier d'un modèle avec de bonnes propriétés (e.g., robustesse), les algorithmes associés à AXDA ont l'avantage de pouvoir simplifier, rendre plus efficace ou encore distribuer l'inférence [6, 7, 9].

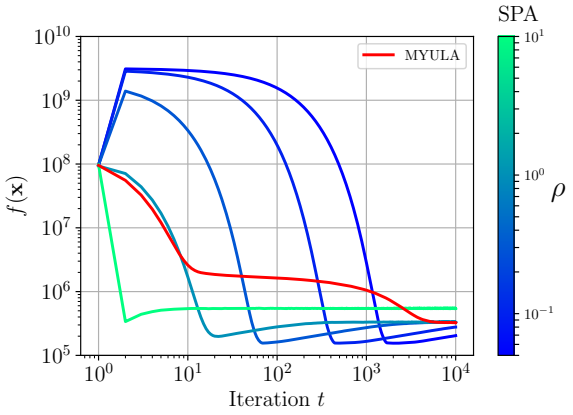

FIGURE 3 – Déconvolution d'image sous prior TV. Convergence des chaînes de Markov AXDA(ρ) et MYULA.

Remerciements

Une partie de ce travail a été soutenue par le projet ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI).

Références

[1] H. O. Hartley, "Maximum likelihood estimation from incomplete data," *Biometrics*, vol. 14, no. 2, pp. 174–194, 1958.
[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[3] R. H. Swendsen and J.-S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 58, pp. 86–88, Jan 1987.
[4] D. A. van Dyk and X.-L. Meng, "The art of data augmentation," *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1–50, 2001.
[5] S. Sisson, Y. Fan, and M. Beaumont (eds), *Handbook of Approximate Bayesian Computation*, ser. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2018.
[6] M. Vono, N. Dobigeon, and P. Chainais, "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *IEEE Trans. Signal Processing*, vol. 67, no. 6, pp. 1648–1661, 2019.
[7] L. J. Rendell, A. M. Johansen, A. Lee, and N. Whiteley, "Global consensus Monte Carlo," 2018. [Online]. Available : https://arxiv.org/abs/1807.09288/
[8] C. Wang and D. M. Blei, "A general method for robust Bayesian modeling," *Bayesian Anal.*, 2018.
[9] M. Vono, N. Dobigeon, and P. Chainais, "Asymptotically exact data augmentation : models, properties and algorithms," submitted. [Online]. Available : https://arxiv.org/abs/1902.05754/
[10] H. M. Choi and J. P. Hobert, "The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic," *Electron. J. Statist.*, vol. 7, pp. 2054–2064, 2013.
[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[12] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Stat. Comput.*, vol. 26, no. 4, pp. 745–760, July 2016.
[13] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo : When Langevin meets Moreau," *SIAM J. Imag. Sci.*, vol. 11, no. 1, pp. 473–506, 2018.