Computing strategies for complex Bayesian models

Marco Banterle

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Université Paris-Dauphine

Computing strategies for complex Bayesian models

École Doctorale de Dauphine — ED 543

Spécialité **Sciences**

COMPOSITION DU JURY :

Christian P. Robert
Université Paris Dauphine
Directeur de thèse

Jean-Michel Marin
Université de Montpellier
Président du jury

Chris Holmes
University of Oxford
Rapporteur

Sophie Donnet
AgroParisTech/INRA
Membre du jury

Robin Ryder
Université Paris Dauphine
Membre du jury

Soutenue le 21/07/2016
par Marco Banterle

Dirigée par **Christian P. Robert**

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

Table des matières

Table des figures

Liste des tableaux

Resumé

La statistique bayésienne est un paradigme important de la statistique moderne. Issue d'une interprétation subjective de la probabilité, qui est considérée comme une mesure du degré de croyance d'un individu sur l'incertitude d'un événement, son vaste cadre inclut l'inférence paramétrique et non paramétrique et la sélection de modèles dans une approche qui reste cohérente avec la théorie de la décision. Dans ce contexte bayésien, l'absence de résultats analytique est souvent liée à l'analyse de modèles complexes. Cependant, au lieu de décourager les praticiens, cette situation a donné lieu à un large éventail de propositions de méthodes de calcul numériques dont le but est de résoudre efficacement ces problèmes. Cette thèse contribue à l'une de ces propositions.

Dans le chapitre 1 nous introduisons formellement les bases de la statistique bayésienne en portant une attention particulière sur les méthodes de calcul. Une fois expliqué le fait que le résultat de l'inférence que nous visons est une distribution de probabilité, dite distribution *a posteriori*, qui met à jour nos croyances *avant* les données empiriques via le théorème de Bayes

$$\pi(\theta|x) = \frac{\pi(\theta)\mathcal{L}(x|\theta)}{\int_\Theta \pi(\theta)\mathcal{L}(x|\theta)d\theta} \tag{1}$$

et compris que les quantités que nous allons vouloir étudier sont des intégrales par rapport à la distribution a posteriori, nous présenterons des résultats basiques sur les méthodes de Monte Carlo, qui reposent sur l'approximation des intégrales intraitables, leur remplaçant par la version empirique basée sur des échantillons aléatoires :

$$\mathcal{I} = \int_\mathcal{X} h(x)\pi(x)dx \approx \hat{\mathcal{I}}_N = \frac{1}{N}\sum_{i=1}^{N} h(x_i) \quad x_i \overset{iid}{\sim} \pi \tag{2}$$

et le détail des résultats sur la convergence de $\hat{\mathcal{I}}_N \to \mathcal{I}$, surtout en ce qui concerne la distribution asymptotique de l'estimateur de Monte Carlo.

Nous introduisont ensuite la méthode de la transformée inverse, ce qui constitue le Théorème fondamental de la simulation, et des transformations des variables aléatoires utiles comme base pour obtenir des échantillons indépendants suivant une distribution quelconque π simulée par un ordinateur.

Quand la complexité des modèles considérés augmente, l'échantillonnage indépendant devient moins pertinent pour l'approximation des intégrales et il est parfois impossible à simuler. Nous allons ensuite présenter en détail une classe d'algorithmes appelés Markov Chain Monte Carlo (MCMC, Méthodes de Monte-Carlo par Chaînes de Markov) qui génèrent une séquence d'échantillons *dépendants*, avec comme distribution limite la distribution π d'intérêt. Nous allons commencer par détailler ici les conditions sous lesquelles une séquence générée par un noyau de Markov va en

effet converger vers la bonne distribution, en introduisant certaines propriétés clés de ces processus, vaguement définies ici :

— irréductibilité, une notion qui assure asymptotiquement une indépendance au le point de départ de la chaîne ;
— apériodicité, une notion qui garantit que la chaîne ne reviendra pas périodiquement au même sous-ensemble de l'espace que nous voulons explorer ;
— récurrence et récurrence de Harris, notions nécessaires pour prouver que tous les ensembles de probabilité positive sous π seront visités asymptotiquement un nombre infini de fois ;
— réversibilité, une notion qui implique que la direction de l'index de temps discret n'a pas d'importance pour la distribution des échantillons ;
— ergodicité, une notion officialisant la convergence vers la distribution désirée et éventuellement son taux de convergence.

En particulier, nous allons décrire la façon dont ces conditions justifient l'utilisation des échantillons produits de cette manière dans un estimateur de Monte Carlo similaire à 2, ainsi que les propriétés de cette approximation. Dans l'introduction, l'algorithme de Metropolis–Hastings, l'un des algorithmes MCMC les plus généraaux et des plus largement appliquées, sera présenté comme un exemple concret de la méthodologie de Monte Carlo par Chaînes de Markov.

Une façon de comparer l'efficacité statistique des différents algorithmes en termes de variance asymptotique de $\hat{\mathcal{I}}_N(h)$, nommé ici $v(h)$, est également donné :

Definition. Si (X_n) et (Y_n) sont des chaînes de Markov avec noyaux de transition K et L, respectivement, et avec distribution stationnaire commune π, alors $K \succeq_E L$ (*lire*, K est à moins aussi efficace que L) si

$$v_X(h) \leq v_Y(h) \qquad \forall h \in L_0^2(\pi)$$

Enfin, quelques résultats sur l'optimalité du noyau de transition définissant chaque chaîne de Markov seront donnés, avec une attention particulière sur la façon dont nous pouvons adapter le mécanisme d'échantillonnage en utilisant des points précédemment échantillonnées dans la chaîne afin d'atteindre cette optimalité.

Tous ces résultats seront utilisés dans les sections suivantes pour prouver la validité et l'efficacité de variantes efficaces de chaîne de Markov classique.

Nous terminerons le chapitre 1 en introduisant une autre classe d'algorithmes appelé Sequential Monte Carlo (SMC, méthodes de Monte-Carlo séquentielles) qui font usage des mêmes noyaux de transition markoviens pour échantillonner une séquence de distributions $\{\pi_t\}$. Nous allons détailler dans quelles situations ces algorithmes peuvent servir nos objectifs et présenter un exemple concret d'un tel système, généralement appelé SMC *tempéré*, pour échantillonner notre distribution d'intérêt π en définissant les éléments de la séquence pour une version 'chauffée' de notre objectif, soit $\pi_t = \pi^{g(t)}$. Ce type de plan d'échantillonnage sera également utilisé dans cette thèse pour effectuer l'inférence dans un modèle statistique spatiale complexe.

Chapitre 2

La première contribution originale se trouve dans le chapitre 2. Lors de l'exécution d'un échantillonneur MCMC, comme Metropolis–Hastings, la complexité de la den-

sité cible peut induire des ralentissements importants dans l'exécution de l'algorithme. Une illustration directe de cette difficulté est la simulation d'une distribution a posteriori impliquant un grand nombre n de données pour lesquels le temps de calcul est au moins de l'ordre $\mathrm{O}(n)$. Plusieurs solutions à ce problème ont était proposée dans la littérature récente (Korattikara, Chen et Welling, 2013 ; Neiswanger, Wang et Xing, 2013 ; Scott et al., 2013 ; Wang et Dunson, 2013), en faisant usage de la décomposition

$$\prod_{i=1}^{n} \ell(\theta|x_i) \tag{3}$$

pour gérer des sous-ensembles de données sur différents processeurs (CPU), différents unités graphiques (GPU), ou même différents ordinateurs. Cependant, il n'y a pas de consensus sur la méthode à appliquer, certaines conduisant à des instabilités en supprimant la plupart des informations à priori et d'autres à des approximations difficiles à évaluer ou même à mettre en œuvre.

Notre approche, denommée Delayed Acceptance, est de retarder l'acceptation dans l'étape de Metropolis-Hastinsg (plutôt que le rejet comme dans Tierney et Mira, 1998) en comparant séquentiellement des parties du rapport d'acceptation à des générations uniformes indépendantes, afin d'arrêter le calcul dès qu'un test est refusé.

Plus formellement, on considère un algorithme générique de Metropolis–Hastings où le taux d'acceptation $\pi(y)q(y,x)/\pi(x)q(x,y)$ est comparé à une variable $\mathcal{U}(0,1)$ pour décider si la chaîne de Markov se déplace de sa valeur courante x à la valeur proposée y (Robert et Casella, 2004). Nous décomposons le rapport en un produit quelconque

$$\pi(y)\, q(y,x)\big/\pi(x)q(x,y) = \prod_{k=1}^{d} \rho_k(x,y)\,, \tag{4}$$

avec pour seule contrainte que les fonctions ρ_k sont toutes positives et qu'elles satisfassent la condition d'équilibre $\rho_k(x,y) = \rho_k(y,x)^{-1}$. Nous acceptons la proposition y avec probabilité

$$\prod_{k=1}^{d} \min\left\{\rho_k(x,y), 1\right\}\,, \tag{5}$$

i.e., en comparant successivement variables uniformes u_k et termes $\rho_k(x,y)$. La motivation de notre approche est que la densité cible initiale $\pi(\cdot)$ demeure stationnaire pour la chaîne de Markov décrite ci-dessus. Une validation empirique de cette procédure peut être visualisée sur la Figure 1.

En pratique, la comparaison séquentielle de ces probabilités avec d uniformes signifie que les comparaisons s'arrêtent au premier rejet, ce qui implique un avantage en temps de calcul si les éléments les plus coûteux dans le produit (4) sont retardées le plus tard possible.

L'inconvénient majeur de ce système est que Delayed Acceptance réduit *efficacement* le coût de calcul uniquement lorsque l'approximation $\tilde{\pi}$ produite dans les premiers termes est 'assez bonne' ou 'assez plate', puisque la probabilité d'acceptation d'un point proposé sera toujours plus petite que dans le Metropolis–Hastings originel. En autres termes, ce Metropolis–Hastings originel domine Delayed Acceptance

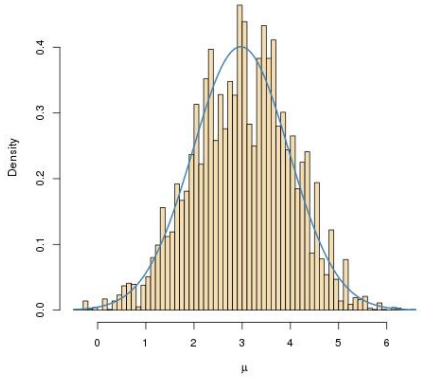


FIGURE 1: Adéquation d'un algorithme de Delayed Acceptance en deux étapes appliqué à un modèle normal-normal, avec une distribution a posteriori $\mu|x \sim N(x/(\{1+\sigma_\mu^{-2}\}, 1/\{1+\sigma_\mu^{-2}\})$ avec $x = 3$ et $\sigma_\mu = 10$, basé sur $T = 10^5$ itérations et deux étapes de Delayed Acceptance, une première acceptation fondée sur le rapport de vraisemblance et une seconde considérant le rapport des distributions a priori. Le taux global d'acceptation est de 12%.

pour l'ordre de Peskun (PESKUN, 1973b). La question la plus pertinente soulevée par CHRISTEN et FOX, 2005 est donc de parvenir à une approximation correcte. Notons que, même si dans la statistique bayésienne une décomposition de la cible est toujours disponible, en séparant les données d'origine en des sous-échantillons ou même simplement en séparant la distribution instrumentale q, la vraisemblance et la probabilité a priori en différents facteurs, ces décompositions peuvent simplement détériorer les propriétés de l'algorithme sans impact sur l'efficacité du calcul.

Plus formellement, nous pouvons définir Delayed Acceptance à partir d'un algorithme de Metropolis–Hastings avec noyau de transition Markovien P tel que

$$P(x, A) := \int_A q(x, y)\alpha(x, y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x, y)\alpha(x, y)\mathrm{d}y\right) \mathbf{1}_A(x),$$

où

$$\alpha(x, y) := 1 \wedge r(x, y), \qquad r(x, y) := \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)},$$

π est la densité cible et $q(x, y)$ représente la densité instrumentale. Ici, $\alpha(x, y)$ est définie comme la probabilité d'acceptation de Metropolis–Hastings et $r(x, y)$ comme le rapport d'acceptation de Metropolis–Hastings. Le noyau de transition markovien du Delayed Acceptance est défini par

$$\tilde{P}(x, A) := \int_A q(x, y)\tilde{\alpha}(x, y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x, y)\tilde{\alpha}(x, y)\mathrm{d}y\right) \mathbf{1}_A(x),$$

où

$$\tilde{\alpha}(x, y) := \prod_{k=1}^{d} \{1 \wedge \rho_k(x, y)\}.$$

Une validation formelle de la méthode suit alors des lemmes suivants :

Lemma 1. *Pour toute chaîne de Markov avec le noyau de transition* Π

$$\Pi(x, A) = \int_A q(x, y)a(x, y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x, y)a(x, y)\mathrm{d}y\right) \mathbf{1}_A(x),$$

*et satisfaisant l'équation de balance ponctuelle (*detailed balance*), la fonction* $a(\cdot)$ *satisfait (pour* π*-a.a.* x, y*)*

$$\frac{a(x, y)}{a(y, x)} = r(x, y).$$

La chaîne de Delayed Acceptance $(\tilde{X}_n)_{n\geq 1}$ est donc associé à la cible visée :

Lemma 2. $(\tilde{X}_n)_{n\geq 1}$ *est une chaîne de Markov réversible et de loi stationnaire* π*.*

Démonstration. De par le Lemme 1 il est suffisant de vérifier que $\tilde{\alpha}(x, y)/\tilde{\alpha}(y, x) = r(x, y)$. On a

$$\begin{aligned}
\frac{\tilde{\alpha}(x, y)}{\tilde{\alpha}(y, x)} &= \frac{\prod_{k=1}^d \{1 \wedge \rho_k(x, y)\}}{\prod_{k=1}^d \{1 \wedge \rho_k(y, x)\}} \\
&= \prod_{k=1}^d \frac{1 \wedge \rho_k(x, y)}{1 \wedge \rho_k(y, x)} \\
&= \prod_{k=1}^d \rho_k(x, y) = r(x, y),
\end{aligned}$$

vu que $\rho_k(y, x) = \rho_k(x, y)^{-1}$ et $(1 \wedge a)/(1 \wedge a^{-1}) = a$ pour $a \in \mathbb{R}_+$. □

Remark 1. Il est immédiat de montrer que

$$\tilde{\alpha}(x, y) = \prod_{k=1}^d \{1 \wedge \rho_k(x, y)\} \leq 1 \wedge \prod_{k=1}^d \rho_k(x, y) = 1 \wedge r(x, y) = \alpha(x, y),$$

vu que $(1 \wedge a)(1 \wedge b) \leq (1 \wedge ab)$ pour $a, b \in \mathbb{R}_+$.

On peut alors comparer un algorithme de Delayed Acceptance avec son un algorithme de Metropolis–Hastings 'parent' en termes d'efficacité, par exemple en comparant leur variance asymptotique. Bien sûr, en restant uniquement concentré sur l'ordre de Peskun nous pouvons montrer que le Metropolis–Hastings original domine toujours Delayed Acceptance vu que $\tilde{\alpha}(x, y) \leq \alpha(x, y)$, mais on peut dériver une relation simple entre les deux grâce au Lemme 34 de Andrieu, Lee et Vihola, 2013.

En veillant à ce que si la probabilité d'acceptation $\alpha(x, y)$ est 1 alors la probabilité d'acceptation $\tilde{\alpha}(x, y)$ est uniformément minorée par une constante positive, on peut montrer que $\tilde{P}$ est relativement proche de P et en particulier que si P admet un trou spectral à droite, la même propriété s'applique pour $\tilde{P}$, avec une implication imédiate quant à son ergodicité géométrique. De plus, et indépendamment de l'existence du trou spectral, les limites quantitatives de la variance asymptotique des estimations MCMC utilisant Delayed Acceptance par rapport à celles qui utilisent Metropolis–Hastings sont disponibles.

Nous allons maintenant décrire des résultats sur les propriétés de mélange optimal et sur le coût de calcul pour la méthode Delayed Acceptance. Rappelons d'abord

que les performances exploratoires d'un algorithme de Metropolis–Hastings à marche aléatoire sont fortement dépendantes de sa loi de proposition et, comme illustré dans Roberts, Gelman et Gilks, 1997, trouver le paramètre d'échelle optimal conduit à des 'sauts' optimaux dans l'espace des états. Le taux d'acceptation global de la chaîne est relié directement à la distance moyenne de saut et à la variance asymptotique des moyennes ergodiques. Optimiser en ce sens offre aux praticiens une approche effective pour régler l'algorithme de Metropolis-Hastings. L'extension de cette calibration au régime du Delayed Acceptance est donc important, soit pour trouver une échelle raisonnable pour la distribution instrumentale, soit pour éviter les comparaisons avec les algorithmes standard de Metropolis–Hastings. En utilisant les résultats liées à la distribution marginale des chaînes considérées dans la limite de la dimension de la distribution cible (Roberts, Gelman et Gilks, 1997) et Expected Square Jumping Distance (ESJD, distance au carré de saut moyen) (Sherlock et Roberts, 2009) on arrive à cette conclusion :

Lemma 3. *Supposant érifiées les hypothèse H1 - H4 (définies ci-après) sur la loi cible $\pi(x)$, sur la loi de proposition $q(x,y)$ et sur la probabilité d'acceptation factorisée $\tilde{\alpha}(x,y) = \prod_{i=1}^{2}(1 \wedge \rho_i(x,y))$, il vient que*

$$\tilde{\alpha}(x,y) = (1 \wedge \rho_1(x,y))$$

et aussi que lorsque $d \to \infty$

$$\mathbf{Eff}(\delta,\ell) = \frac{h(\ell)}{\delta + \mathbb{E}\left[\tilde{\alpha}\right]} = \frac{2\ell^2\Phi(-\frac{\ell\sqrt{I}}{2})}{\delta + 2\Phi(-\frac{\ell\sqrt{I}}{2})}$$

$$a(\ell) = \mathbb{E}\left[\tilde{\alpha}\right] = 2\Phi(-\frac{\ell\sqrt{I}}{2})$$

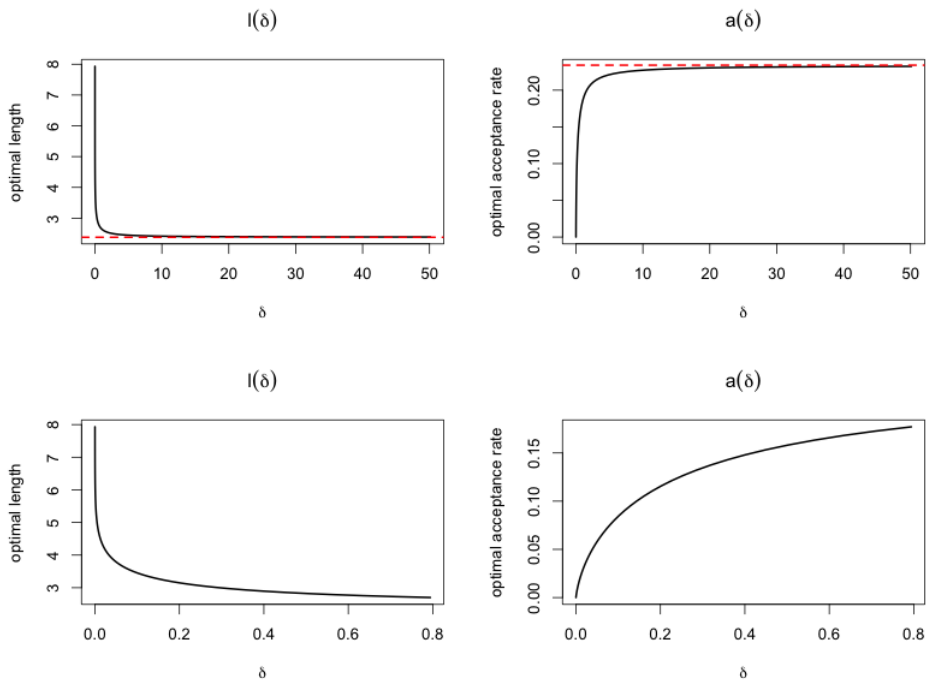
où $I := \mathbb{E}\left[\left(\frac{(\pi(x))'}{\pi(x)}\right)^2\right]$ est défini en Roberts, Gelman *et* Gilks, *1997.*

L'échelle optimale de la loi de proposition $\ell^*(\delta)$ et le taux d'acceptation optimale $a^*(\delta)$ sont obtenus comme des fonctions de δ. En particulier, quand le coût relatif de calcul de $\rho_1(x,y)$ par rapport à celui de $\rho_2(x,y)$ diminue, les mouvements proposés deviennent plus variables, ℓ^* augmente et a^* diminue, puisque chaque rejet coûte peu à l'algorithme en termes de temps, alors que tous les déplacements acceptés résultent presque en un échantillon indépendant. Au contraire, quand δ grandit rapidement, la dynamique de la chaîne se rapproche d'un comportement de Metropolis–Hasting. La Figure 2 permet de visualiser le résultat. Cela implique aussi que le taux d'acceptation optimal $\alpha^*(\delta)$ est indépendant de I.

Un résultat légèrement différent est obtenu pour la procédure de MALA géométrique décrite par Girolami et Calderhead, 2011, qui représente une application idéale pour le Delayed Acceptance. En effet, nous pouvons naturellement diviser son taux d'acceptation dans le produit du rapport a posteriori et le rapport des lois de proposition, celui-ci devant être calculé uniquement lorsque le point proposé est associé à une relativement grande probabilité a posteriori.

Le maillon faible des temps de calcul du G-MALA est en fait le calcul de la dérivée troisième de notre log-cible au point proposé, alors que le calcul de la partie

FIGURE 2: Panneaux supérieurs : comportement de $\ell^*(\delta)$ et $\alpha^*(\delta)$ en fonction du coût relatif δ. Il faut noter que pour $\delta >> 1$ les valeurs optimales convergent vers les valeurs calculées pour l'algorithme standard de Metropolis–Hasting (en pointillés rouges). Panneaux inférieurs : gros plan de la région intéressante pour $0 < \delta < 1$.



a posteriori elle-même a généralement un faible coût relatif. En plus, même avec un mécanisme de proposition efficace non-symétrique (la diffusion discrétisé de Langevin), G-MALA est encore à peu prés une marche aléatoire et nous nous attendons à ce que le rapport de la loi de proposition soit proche de 1, en particulier à l'équilibre et lorsque la taille ε des étapes est faible. Donc, le premier rapport est peu coûteux par rapport au second, alors que la décision prise à la première étape devrait être compatible avec le taux global d'acceptation.

Étant donné que l'échelle optimale pour l'algorithme MALA en fonction de la dimension d de la cible diffère de l'échelle optimale pour une marche aléatoire (voir ROBERTS et ROSENTHAL, 2001), nous avons fixé la variance de la composante normale de la marche aléatoire à $\sigma_d^2 = \frac{\ell^2}{d^{1/3}}$. D'après les résultats ci-dessus, on peut obtenir le taux d'acceptation optimal pour le DA-MALA en maximisant

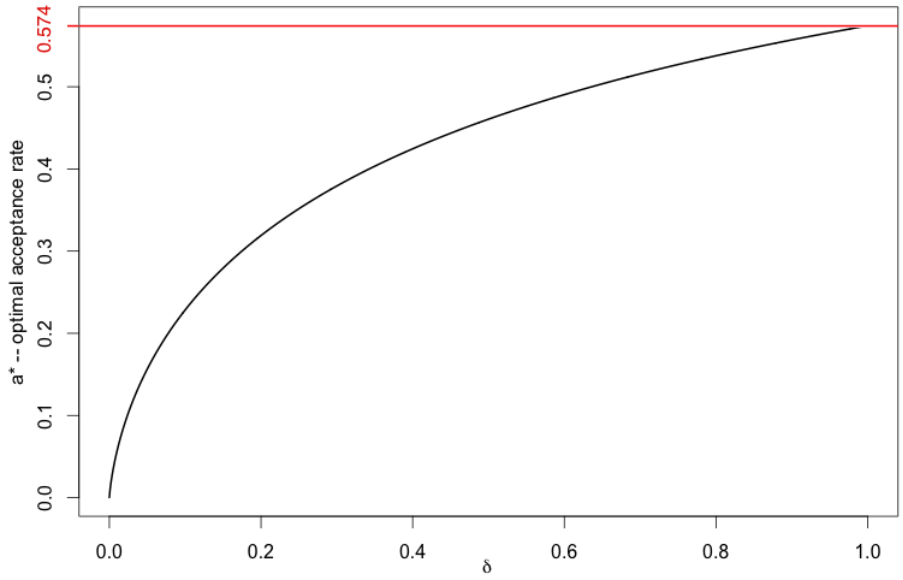
$$\mathbf{Eff}(\delta, \ell) = \frac{h(\ell)}{\delta + \mathbb{E}\left[\tilde{\alpha}\right] - \delta \times \mathbb{E}\left[\tilde{\alpha}\right]} = \frac{2\ell^2 \Phi(-\frac{K\ell^3}{2})}{\delta + 2\Phi(-\frac{K\ell^3}{2}) \times (1-\delta)}$$

ou de manière équivalente

$$\mathbf{Eff}(\delta, a) = -\left(\frac{2}{K}\right)^{\frac{2}{3}} \left[\frac{a\Phi^{-1}\left(\frac{a}{2}\right)^{\frac{2}{3}}}{\delta + a(1-\delta)}\right].$$

Dans cette évaluation, le coût de calcul par itération est fixé à $c = \delta C$ pour le rapport des a posteriori, $C = 1$ pour le rapport des lois de proposition (et donc $c + \mathbb{E}\left[\tilde{\alpha}\right](C - c)$ pour l'ensemble du noyau), $h(\ell)$ est maintenant la vitesse du processus

FIGURE 3: Taux d'acceptation optimal pour l'algorithme DA-MALA en fonction de δ. En rouge, la valeur optimale du taux d'acceptation de MALA standard obtenu par ROBERTS et ROSENTHAL, 2001 reporté pour $\delta = 1$.



de diffusion limite et K est une mesure de 'rugosité' de la distribution cible, fonction de ses dérivées. Le taux d'acceptation optimal a^* est indépendant de K, que nous ne définissons pas plus rigoureusement, nous référant à ROBERTS et ROSENTHAL, 2001. La Figure 3 montre que a^* diminue avec δ, comme c'est le cas pour la marche aléatoire. Il atteint l'optimum connu pour le MALA standard quand $\delta = 1$.

La dernière partie de ce travail illustre notre procédure sur une série de données simulées et réelles.

Chapitre 3

Les modèles spatiaux hiérarchiques pour les processus environnementaux reçoivent de plus en plus d'intérêt vu que la disponibilité des données géostatistiques croît grâce aux systèmes de positionnement global. La majorité de ces modèles font usage du cadre extrêmement flexible des processus gaussiens pour effectuer l'estimation des quantités d'intérêt, mais reposent encore souvent sur l'hypothèse simplificatrice de stationnarité.

Le but du travail derrière ce chapitre est d'introduire une technique générale et intuitive qui modèle explicitement la non-stationnarité, et qui nous permet de tenir compte des effets environnementaux cachés qui correspondent à un phénomène physique ou non, tout en conservant la simplicité des processus gaussiens stationnaires.

Un certain nombre de méthodes existantes traitent déjà de la modélisation de la non-stationnarité et appartiennent généralement à l'une de deux catégories : convolution non triviale des processus localement stationnaires ou 'image-warping', c'est à dire l'emploi de techniques de déformation spatiale. Dans le premier cas, on suppose que lors de l'observation du processus à l'étude *localement* l'effet de la non-stationnarité est négligeable et donc un modèle stationnaire local est utilisé. Ces

processus locaux sont plus tard 'convolués' (dans un sens général) en le processus global (non stationnaire). La partie de la littérature connue comme *process-convolution*, née avec Higdon, 1998 and Higdon, Swall et Kern, 1999 montre en effet que certains processus gaussiens non stationnaires peuvent être représentés par le convolution des noyaux locaux par des mouvements browniens, même si les noyaux sont autorisés à varier spatialement pour modéliser la non-stationnarité ; Xia et Gelfand, 2005 et Paciorek, 2007 ont étendu plus tard l'idée à une plus grande classe de processus. Similaire à la convolution des processus, on peut citer les modelés *low-rank splines* (Ruppert, Wand et Carroll, 2003 ; Lin et al., 2000) et, plus récemment, l'approche par processus prédictifs (Banerjee et al., 2008 ; Finley et al., 2009 ; Eidsvik et al., 2012). Les techniques de *image-warping* comprennent toutes les dérivations du travail de Sampson et Guttorp, 1992, dont l'idée est de déformer la géographie globale de sorte à ce que le processus observé semble stationnaire dans l'espace résultant. Des méthodes de *Multi–Dimensional scaling* sont généralement utilisés pour définir les déformations et des splines gèrent la transformation entre l'espace original et l'espace déformé. Cette idée a ensuite été étendu au cadre bayésien par Damian, Sampson et Guttorp, 2001 et Schmidt et O'Hagan, 2003.

Suite aux résultats de Perrin et Meiring, 2003 et Perrin et Schlather, 2007, qui prouve l'existence d'une représentation de dimension supérieure pour tout champ non stationnaire, Bornn, Shaddick et Zidek, 2012 ont développé une méthode d'optimisation appelé Dimension Expansion (DE) pour modéliser un processus non stationnaire $Y(\mathbf{X})$. Les dimensions étendues (latentes) $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_s)$ sont apprises afin que le processus $Y([\mathbf{X}, \mathbf{Z}])$, où $[\mathbf{X}, \mathbf{Z}]$ représente la concaténation entre les sites observés et les dimensions latentes, semble être stationnaire ; pour être précis les composantes de $\mathbf{Z}$ sont estimées de sorte à ce que le variogramme paramétrique supposé exhibe une distance minimale depuis la version empirique, dans l'espace augmenté.

Sans un cadre inférentiel approprié cette méthode n'a pas moyen d'observer *l'incertitude* de l'espace latent inférée et de correctement *estimer* les paramètres intervenant dans la modélisation de la structure de covariancede ; Tel est l'objectif du travail derrière ce chapitre..

Bien que similaire à l'*image-warping* à première vue, cette méthode diffère fondamentalement vu que l'espace d'origine est conservée et pas nécessairement déformé. On ajoute en fait la flexibilité par l'introduction des dimensions supplémentaires. De plus, cette approche ne prête pas le flan à l'un des inconvénients majeurs de l'*image-warping*, qui est la possibilité de plier l'espace sur lui-même, qui se traduit par une transformation non-injective entre l'espace d'origine et l'espace déformé. Dans un tel cas, deux endroits différents se chevauchent et ils deviennent essentiellement des répétitions du processus, dont la variance est contrôlée uniquement par erreur de mesure indépendante plutôt que d'être, plus logiquement, fortement corrélées.

On considére un processus $\{R(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}$ observé, potentiellement non-stationnaire. Supposons que nous pouvons décomposer R comme

$$R(\mathbf{X}) = Y(\mathbf{X}) + \mu(\mathbf{X}) + \varepsilon(\mathbf{X}) \tag{6}$$

où $\mu(\mathbf{X})$ est une fonction moyenne, qui peut dépendre de certaines covariables, et $\varepsilon(\mathbf{X})$ est un processus d'erreur de mesure indépendante, parfois appelé *nugget*, indépendant de $Y(\mathbf{X})$ qui, à son tour, capture l'association spatiale du processus R

et est l'objectif principal de l'inférence dans ce travail. Modéliser R est habituellement une extension triviale de la modélisation de Y, vu que la fonction moyenne μ est généralement supposée être une fonction déterministe des localisations spatiales $\mathbf{X}$ et, possiblement, de certaines covariables, souvent estimées de manière non-paramétrique (voir Rue, Martino et Chopin, 2009 par exemple). $\varepsilon(\mathbf{X})$ est souvent supposé suivre une distribution gaussienne avec matrice de covariance diagonale.

Supposons maintenant que nous observons $\{Y(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}$, où $\mathcal{S} \subseteq \mathbb{R}^d$ est un processus gaussien univariée avec moyenne nulle et fonction de covariance $\Sigma_{\theta_y}(h)$, h étant la distance entre deux points. Des extensions au cas multivarié pour Y existent et ce qui suit s'applique facilement au prix d'un changement de notations. Nous renvoyons le lecteur à (par exemple) Genton et Kleiber, 2015 et à la discussion relative pour une introduction sur l'opérateur de *cross-covariance*, qui étend la fonction de covariance, fondement d'une telle généralisation.

La structure stationnaire ci-dessus (Cressie, 1993) peut être déraisonnable car elle suppose explicitement qu'il n'y a pas d'association entre les positions dans l'espace, ce qui est souvent irréaliste. Afin de modéliser cette *non stationnarité*, sous la forme d'association spatiale, nous allons agrandir les positions en ajoutant un processus latent $\mathbf{Z} \in \mathcal{Q}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ tel que $Y([\mathbf{X}, \mathbf{Z}])$, maintenant observée sur un sous-ensemble de $\mathbb{R}^{p+d}$, de sorte à être stationnaire.

Pour déduire correctement le processus latent comme une fonction 'lisse' de l'espace d'origine, nous allons élucider une distribution a priori sur $\mathbf{Z}$ tel que $\{\mathbf{Z}_i(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}_{i=1,\dots,p}$ est un processus gaussien univarié avec moyenne nulle et fonction de covariance $\Sigma_{\theta_z}(h)$. Supposer une connaissance préalable sur la corrélation dans un tel processus latent multivarié semble irréaliste et donc nous allons tout simplement modéliser ses lois marginales.

Pour conserver une certaine souplesse, sans aucune hypothèse forte sur la régularité des processus, nous allons supposer ici que les fonctions de covariance ont une forme de fonction de Matérn, pour Y et pour $\mathbf{Z}$. On aura donc $\theta = (\sigma^2, \phi)$ (les deux étant des nombres réels strictement positifs) et

$$\Sigma_{\theta,\nu}(h) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\frac{h}{\phi}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{h}{\phi}\right) \tag{7}$$

où $\Gamma(\cdot)$ est la fonction Gamma et $K_\nu(\cdot)$ est la fonction de Bessel de seconde espèce.

Pour compléter la formulation bayésienne, nous explicitons maintenant une distribution a priori pour σ_y^2, σ_z^2, qui représente la variabilité des processus (par exemple les éléments diagonaux de la matrice de covariance), et ϕ_y, ϕ_z, qui contrôle la régularité de la fonction de covariance ou plutôt le taux de décroissance de la corrélation par rapport à la distance.

Comme ces quantités sont toutes positives, en l'absence d'information a priori, nous allons leur associer une loi a priori $\mathcal{G}amma$ diffuse. N'imposer aucune contrainte sur σ_z^2 peut s'avérer être dangereux, vu que l'effet de $\mathbf{Z}$ en

$$\Sigma_{\sigma_y^2,\phi_y}(d([\mathbf{X}, \mathbf{Z}]))$$

peut potentiellement masquer complètement l'effet de $\mathbf{X}$. En effet, les corrélations ne dépendent que de la distance entre les points. Quel que soit le phénomène latent qu'on mesure implicitement, nous voudrions supposer qu'il se trouve dans un espace

comparable à ce que nous observons dans $\mathbf{X}$; donc nous forçons habituellement la variation du processus latent à être compatible avec la variation de $\mathbf{X}$, cela en ajustant les hyper-paramètres $\alpha_{\sigma_z^2}$ et $\beta_{\sigma_z^2}$.

Pour résumer, le modèle bayésien de Dimension Expansion complet peut être écrit comme :

$$Y|\mathbf{X}, \mathbf{Z}, \sigma_y^2, \phi_y \sim GP(\mathbf{0}, \Sigma_{\sigma_y^2,\phi_y}(d([\mathbf{X}, \mathbf{Z}]))) \tag{8a}$$

$$\mathbf{Z}_i|\mathbf{X}, \sigma_z^2, \phi_z \sim GP(\mathbf{0}, \Sigma_{\sigma_z^2,\phi_z}(d(\mathbf{X}))) \tag{8b}$$

$$\sigma_y^2 \sim \Gamma(\alpha_{\sigma_y^2}, \beta_{\sigma_y^2}),\ \phi_y \sim \Gamma(\alpha_{\phi_y}, \beta_{\phi_y}) \tag{8c}$$

$$\sigma_z^2 \sim \Gamma(\alpha_{\sigma_z^2}, \beta_{\sigma_z^2}),\ \phi_z \sim \Gamma(\alpha_{\phi_z}, \beta_{\phi_z}) \tag{8d}$$

où $d(\cdot)$ est la fonction de la distance euclidienne et $\alpha_{\sigma_y^2}, \beta_{\sigma_y^2}, \alpha_{\sigma_z^2}, \beta_{\sigma_z^2}, \alpha_{\phi_y}, \beta_{\phi_y}, \alpha_{\phi_z}, \beta_{\phi_z}$ sont tous des nombres réels positifs.

Dans le cas où l'espace latent représente des covariables latentes potentielles, p est inconnu et même dans un cadre réaliste, on ne peut connaître cette dimension. En spécifiant une distribution a priori $\pi(p)$, le modèle de base qui correspond à la *stationnarité*, donc à $p = 0$, peut avoir maintenant une probabilité positive. Cette propriété est donc en accord avec pour les principes proposés dans Simpson et al., 2014 et nous sommes sûr de ne pas forcer un modèle non stationnaire sur Y quand, au contraire, l'hypothèse de stationnarité est supportée par les données. D'autre part, cette généralité a pour contrepartie quelques inconvénients calculatoires, vu qu'une forme de *reversible jump* (Green, 1995) doit être inclus dans notre procédure, ce qui souvent provoque une baisse significative du taux d'acceptation.

Une propriété des méthodes de Monte Carlo séquentielles est la possibilité d'obtenir comme un sous-produit de l'algorithme un estimateur de la *vraisemblance marginale* ; si on définit la séquence de distribution cible

$$\pi_j(\cdot) = \pi(p)\pi(\theta_z)\pi(\theta_y)\pi(\mathbf{Z}|\theta_z, \mathbf{X})\ell(Y|\mathbf{X}, \mathbf{Z}, \theta_y)^{t_j},$$

avec t_j allant de ∞ à 1, on peut alors exécuter l'algorithme pour une valeur de p fixe à plusieurs reprises (une fois pour chaque p tel que $\pi(p) > 0$) et enfin, grâce aux facteurs de Bayes associés, décider de la dimension du processus latent le plus probable et ainsi tester formellement la non-stationnarité.

Lorsqu'on est incapable d'exécuter cette procédure coûteuse pour toutes les valeurs possibles de p, on pourrait obtenir une distribution a priori qui évite l'*overfitting* vers la non-stationnarité et satisfaire les principes de Simpson et al., 2014. La solution est de remplacer σ_z^2 par son inverse τ_z, qui représente la précision a priori et expliciter pour ce paramètre une distribution a priori dont les moments ne sont pas défini. Pour une situation similaire Simpson et al., 2014 conseillent une distribution de Gumbel de type 2. Cela permettrait de réduire le processus à une surface plane, donc Y serait effectivement défini sur son espace d'origine.

Même à p fixé on rencontre des difficultés dans l'échantillonnage du modèle ci-dessus, sur $\mathbf{Z}$ en particulier. Tout d'abord la vraisemblance et la distribution a posteriori pour $\mathbf{Z}$ sont invariantes sous certaines transformations *isométriques*, et donc le modèle n'est pas complètement identifiable. D'autre part les éléments de chaque dimension supplémentaire $\mathbf{Z}_i$ sont, par construction, très fortement corrélées et donc l'échantillonnage peut être difficile.

Le premier point est en fait moins grave que prévum parce que nous sommes surtout intéressés à inférer la matrice de covariance ou la fonction $\Sigma_{\sigma_y^2,\phi_y}$ du processus

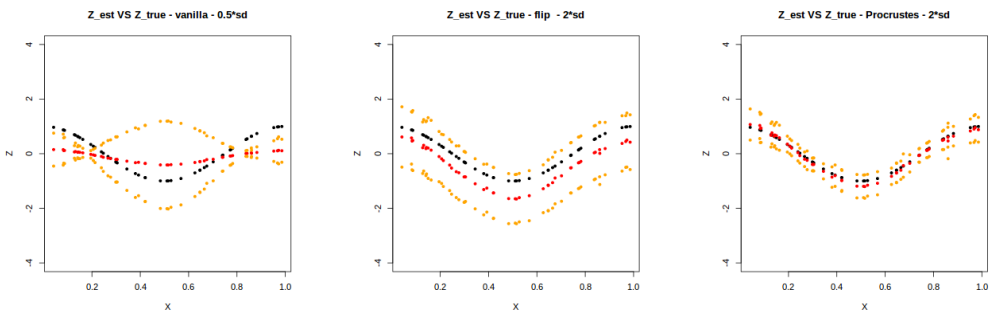


FIGURE 4: Échantillons non traités (à gauche), post-traitement équivalent par rapport à la distribution a posteriori (centre) et post-traitement Procuste (à droite) sur un processus latent unidimensionnel. En noir le processus latent réel, en rouge son espérance a posteriori et en orange ± 0.5 l'écart-type a posteriori (à gauche) ou ± 2 l'écart-type postérieur (au centre et à droite).

observé, plutôt que $\mathbf{Z}$-même. Ces quantités ne sont pas impactées par ce problème d'identifiabilité : on peut par exemple inverser autour de zéro un processus latent unidimensionnel ou tout bien échanger les indices des lois marginales dans un $\mathbf{Z}$ bivarié. Dans les deux cas, les distances entre points ne sont pas modifiés.

Cette résolution pourrait affecter les propriétés de mélange d'une procédure MCMC et une stratégie similaire à celle appliquée dans NOBILE, 1998 peut être mise en œuvre pour atténuer ce problème. Elle consiste à périodiquement proposer un passage à un point de même probabilité a posteriori, afin d'explorer correctement tout l'espace $\mathcal{Q}$. Il faut noter cependant que l'identification des transformations nécessaires pour mettre en place une telle stratégie devient de plus en plus difficile quand p augmente. SMC est intrinsèquement plus robuste face ce problème, que nous ne discuterons pas davantage.

Si le champ latent correspond à un phénomène réelle, comme par exemple l'altitude manquante dans certaines données environnementales, nous pourrions être intéressés par une estimation a posteriori du processus $\mathbf{Z}$. En effet, on a simplement juste besoin d'une stratégie d'échantillonnage de post-traitement afin de choisir une seule *forme* particulière pour les processus latents et en superposant tous les échantillons à celle-ci, en appliquant le transformations invariantes mentionnées ci-dessus.

Les outils d'analyse nommé *Procrustes* (DRYDEN et MARDIA, 1998) peuvent de plus être utilisés pour donner des idées sur l'interprétation du processus latent, au prix d'un sacrifice sur la précision des estimations de l'incertitude. Il existe des transformation *isométrique* telles qu'une routine générale de procrustes ne les rend pas invariantes par rapport à la distribution a posteriori (même si elles sont invariantes par rapport à la vraisemblance).

Un exemple de la manière dont l'incertitude a posteriori est affectée par une méthode procruste peut observée dans la Figure 4.

Pour aborder le second problème une pléthore de solutions sont disponibles dans la littérature, depuis l'*over-relaxation* (NEAL, 1995) jusqu'aux methodes hamiltoniennes de Monte Carlo (NEAL, 2012 ; GIROLAMI et CALDERHEAD, 2011). La plupart d'entre elles doivent être réglées par l'utilisateur. Etant donné que (i) les problèmes environnementaux peuvent varier beaucoup dans leur formulation et (ii) nous cherchons une méthode assez générale, nous avons opté pour une autre

procédure récente qui profite à la fois de l'*over-relaxation* et du *slice sampling* (Neal, 2003). Proposé par ce problème et en particulier pour les processus de Gauss, le *Elliptical Slice Sampler* de Murray, Prescott Adams et MacKay, 2010 s'applique bien dans ce cadre. En dehors des processus gaussiens, les performances de l'*Elliptical Slice Sampler* n'ont pas encore été bien exploréesd (Nishihara, Murray et Adams, 2012), donc HMC peut représenter un concurrent plus attrayant.

Enfin l'échantillonnage de $\sigma_y^2, \phi_y, \sigma_z^2$ et ϕ_z (conditionnées au processus latent) est opéré conjointement par une marche aléatoire avec probabilité instrumentale gaussienne multivariée dans l'espace logarithmique. Ce choix est assez standard dans la littérature et il ne nécessite pas d'autres justifications.

L'algorithme d'échantillonnage est ici présenté dans l'algorithme 0.1, plus de détails peuvent être trouvés dans le chapitre 3 de cette thèse.

Algorithm 0.1 Échantillonneur SMC pour le modèle Dimension Expansion Bayésien

Pour une dimension p, initialiser $t = \infty$, $j = 0$;
for $i \in 1, \ldots, N$ **do**
  $\theta_0^{(i)} \sim \pi(\cdot)$; $Z_0^{(i)} \sim f_0(\cdot)$; $W_0^{(i)} = \frac{1}{N}$.
end for
while $t > 1$ **do**
  $j \leftarrow j + 1$; Calculer t^* tel que $cESS_{t^*} = \rho \times cESS_t$;
  for all i **do**
    – Calculer $W_j^{(i)} \propto W_{j-1}^{(i)} \times \frac{\ell(Y|\theta_{j-1}^{(i)}, Z_{j-1}^{(i)})^{1/t^*}}{\ell(Y|\theta_{j-1}^{(i)}, Z_{j-1}^{(i)})^{1/t}}$;
    – $(\theta_j, Z_j, W_j)^{(i)} \leftarrow (\theta_{j-1}, Z_{j-1}, W_j)^{(i)}$; $\quad t \leftarrow t^*$;
  end for
  Calculer $e = ESS(t)$ et re-échantillonner si $e < (\psi \times N)$;
  for $i \in 1, \ldots, N$ **do**
    – $\theta_j^{(i)} \sim K_\theta^{(j)}(\cdot|\theta_j^{(i)})$;
    – $Z_j^{(i)} \sim K_{Ell}(\cdot|Z_j^{(i)})$;
  end for
end while

où ρ et ψ sont des constantes réelles positives avec valeurs sur $(0, 1)$, qui règlent respectivement la diminution de la température et le seuil pour l'étape de ré-échantillonnage ; π est la distribution a priori pour $\theta = (\sigma_y^2, \phi_y, \sigma_z^2, \phi_z)$, f_0 est la distribution a priori pour $\mathbf{Z}$ et $\ell(Y|\theta, Z)$ est la vraisemblance. K_{Ell} et K_θ sont respectivement les noyaux du elliptical slice sampler et de l'algorithme de Metropolis–Hastings.

Comme d'habitude avec les processus gaussiens, notre approche devient significativement beaucoup coûteuse en temps de calcul si le nombre de positions observées augmente, vu que le calcul de la probabilité a besoin de $\mathcal{O}(n \times s^3)$ opérations. Cela devient problématique même pour un nombre modéré de positions. Diverses approximations par des matrices de bas rang ont été proposées dans la littérature pour surmonter ce problème (Smola et Bartlett, 2001 ; Seeger, Williams et Lawrence, 2003 ; Schwaighofer et Tresp, 2002 ; Quinonero-Candela et Rasmussen, 2005 ; Paciorek, 2007 ; Snelson et Ghahramani, 2005), mais nous allons nous concentrer en particulier sur deux de ces techniques ; nous allons

d'abord présenter le travail de Banerjee et al., 2008 et Datta et al., 2014, qui proposent deux types d'approximations différentes, puis nous expliquerons l'adaptation nécessaire pour qu'ils s'appliquent dans notre cadre de Dimension Expansion.

Dans la Section 3.5 nous montrons comment la stratégie de modélisation décrite peut récupérer différents types de non-stationnarité présentes dans les données, à la fois réelles et simulées. Il est particulier intéressant, en dépit de leur dimension modérée, de considérer l'analyse des données de rayonnement solaire de Hay, 1983. Comme le montrent les figures 5 et 6, la procédure est à même d'estimer un processus latent tel que la matrice de covariance résultante correspond bien à la version empirique. De plus, l'espace élargi estimé où le processus est observé se trouve être très similaire au contour de la région montagneuse autour de Vancouver, ce qui donne une interprétation claire du champ latent.

Chapitre 4

La détection de dépendances entre variables aléatoires est un problème étudié depuis longtemps en statistique et dans la dernière décennie notamment des mesures non linéaires de dépendance sont apparus comme des outils fondamentaux dans de nombreux domaines appliqués, où assumer une distribution gaussienne pour les données observées est irréaliste. Un grand nombre de méthodes ont été proposées, principalement basées sur l'information mutuelle (voir par exemple Kinney et Atwal, 2014 et ses références), ou les méthodes du noyau (Fukumizu et al., 2007 ; Zhang et al., 2012) ; dans la littérature bayésienne des développements récents sur le sujet peuvent être trouvés par exemple dans Kunihama et Dunson, 2014 et Filippi et Holmes, 2015.

La difficulté avec la plupart de ces méthodes est qu'elles doivent considérer chaque paire de variables aléatoires séparément afin d'inférer la structure entière de la dépendance. La plupart d'entre elles ne disposent pas d'une correction appropriée pour les tests multiples. Kunihama et Dunson, 2014 repose sur une estimation non paramétrique de la densité conjointe pour estimer la dépendance, mais elle manque d'une quantification précise de l'approximation de Monte Carlo et par conséquent d'un certain type de calibrage pour les statistiques de test.

Alternativement, les modèles graphiques (conditionnés aux graphes) fournissent une manière élégante d'exprimer la structure de dépendance complète d'un ensemble de variables aléatoires, ce qui les rend attrayants pour des tâches telles que la réduction de dimension dans un cadre de régression. Cependant, ils reposent habituellement sur des hypothèses de linéarité peu convaincantes (comme le modèle graphique gaussien) ou le besoin de recourir à des approximations dans la procédure d'estimation pour tenir compte des modèles plus réalistes (Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015).

Les modèles de copules ont été introduits exactement pour fournir un outil flexible d'étude des données multidimensionnelles et ils ont été largement étudiées et emplyés notamment du fait de leur aptitude à séparer la modélisation des distributions marginales de l'estimation de la structure de dépendance entre eux. Voir par exemple Joe, 2014 pour une revue récente sur le sujet.

Le modèle de copule gaussien graphique a d'abord été introduit en statistique par Hoff, 2007 ; Liu, Lafferty et Wasserman, 2009 ; Liu et al., 2012, permet-

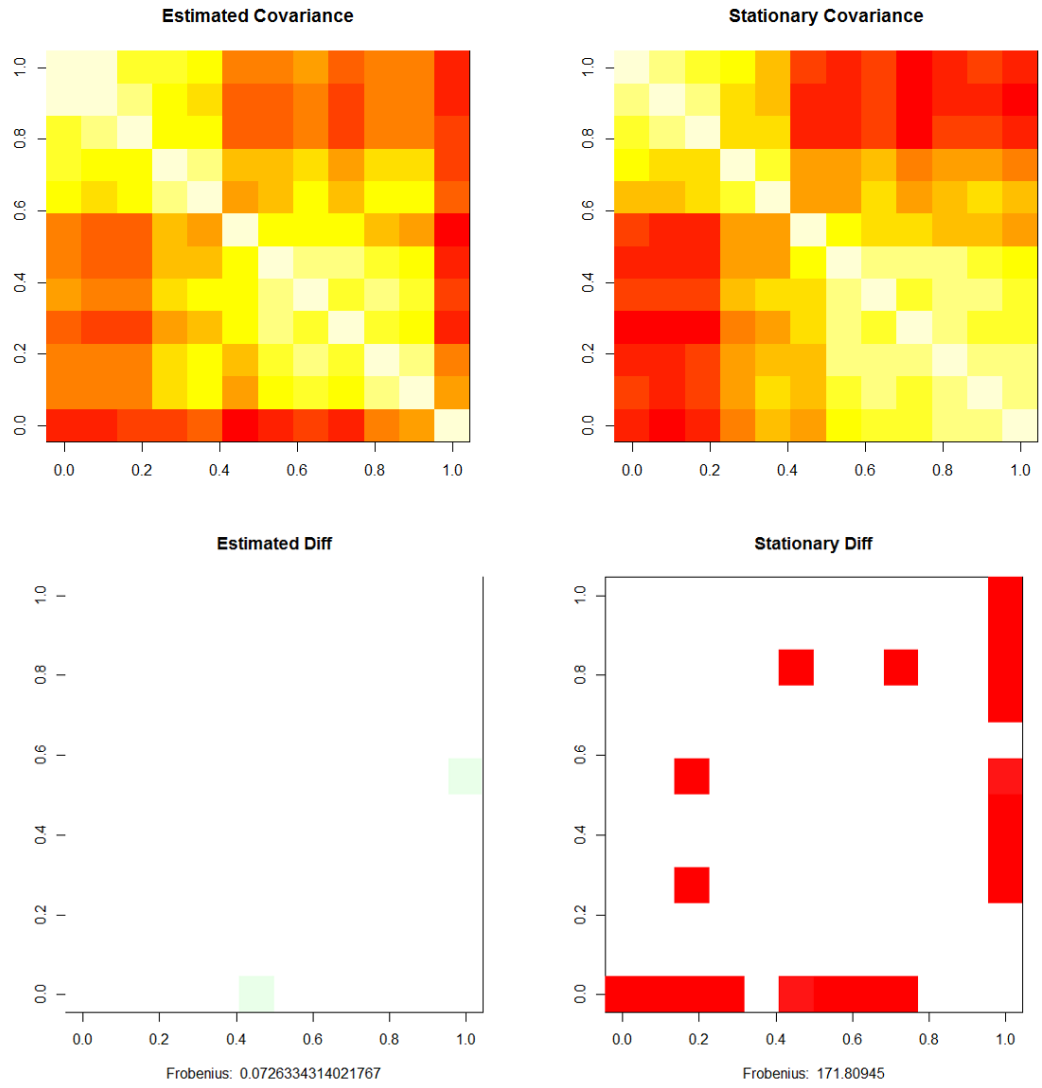


FIGURE 5: Rangée supérieure : Cartes de chaleur pour l'estimation de la matrice de covariance et estimation de la matrice de covariance stationnaire. Rangée inférieure : Cartes de chaleur pour la différence entre matrices de covariance empirique et matrices de covariance estimées, non stationnaire et stationnaire respectivement (incluant la norme de Frobenius), pour des données de rayonnement solaire.

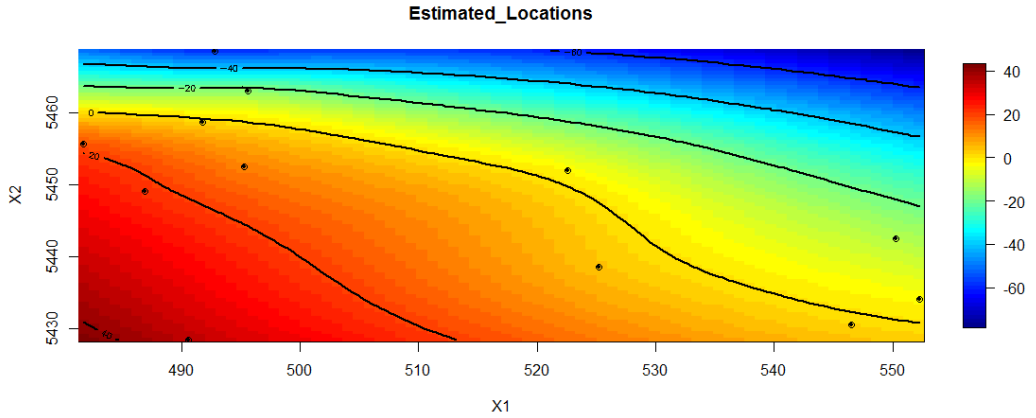


FIGURE 6: Processus latent estimé, données de rayonnement solaire.

tant à la fois une représentation flexible des données multivariées et une inférence précise sur la structure de dépendance grâce au conditionnement à un graphe. Dans une perspective bayésienne il a également été explorée par DOBRA et LENKOSKI, 2011 ; MOHAMMADI et al., 2015 en exploitant la distribution G-Wishart (ROVERATO, 2002) comme loi conjugué pour la matrice de précision Λ de la copule gaussienne.

Jusqu'à récemment, la littérature bayésienne a du se concentrer sur des graphe décomposables, afin de pouvoir calculer la constante de normalisation de la distribution G-Wishart, ou bien introduire des approximations dans la procédure afin de l'estimer. Exploitant la littérature récente sur le G-Wishart (LENKOSKI, 2013 ; UHLER, LENKOSKI et RICHARDS, 2014), nous élaborons une procédure de MCMC exacte pour un modèle de copule gaussien graphique, procédure qui ne partage pas ces limitations. Nous proposons aussi une procédure entièrement bayésienne qui modélise explicitement les lois marginales dans la copule, d'une manière non paramétrique et sans hypothèse forte sur leur forme, grâce à un processus de Dirichlet.

Plus formellement, la densité du modèle de copule gaussien graphique peut être écrit comme

$$f(X|\Theta, R) = |R|^{-\frac{n}{2}} \times \prod_{i=1}^{n} \left(\exp\left\{ -\tfrac{1}{2} \tilde{Z}^{(i)\prime} (R^{-1} - I_d) \tilde{Z}^{(i)} \right\} \prod_{j=1}^{d} f_{v_i}(X_{v_j}^{(i)}, \theta_{v_j}) \right) \tag{9}$$

où

$$\tilde{Z} = \left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\right)$$

et

$$(u_1, \dots, u_d) = (F_{v_1}^{-1}(X_{v_1}), \dots, F_{v_d}^{-1}(X_{v_d})).$$

Nous avons supposé une famille paramétrique pour les marges F_ν, indexé par $\Theta = \{\theta_v : v \in V\}$, mais en général leur estimation peut être délicate pour les modèles les plus réalistes. Notre proposition sera donc de spécifier pour les densités f_v un mélange de distributions h_k, avec k potentiellement infinie, dont les proportions de mélange sont spécifiées par un processus de Dirichlet. Ce modèle non paramétrique est commun dans la littérature sous le nom de mélanges de processus de Dirichlet.

Enfin, pour respecter la structure d'indépendance conditionnelle souhaitée, nous allons d'abord paramétriser la matrice de corrélation R par

$$R_{v_i,v_j}(\Lambda) = \frac{(\Lambda^{-1})_{v_i,v_j}}{\sqrt{(\Lambda^{-1})_{v_j,v_j} \times (\Lambda^{-1})_{v_j,v_j}}}$$

où Λ est une matrice de précision, et on conditionne Λ à un graphe G où l'absence d'arête entre deux nœuds correspond à un zéro pour l'élément associé de Λ.

Le modèle complet est donc

$$\begin{aligned}
Z|\Lambda, G \sim &\quad \mathcal{N}_d(0, \Lambda^{-1}),\\
\tilde{Z}|\Lambda, G \sim &\quad \mathcal{N}_d(0, R(\Lambda)),\\
X_v^{(i)}|\theta_v^{(i)} = &\quad F_v^{-1}\left(\Phi(\tilde{Z}_v^{(i)}); \theta_v^{(i)}\right),\\
\theta_v^{(i)} \sim &\quad P,\\
P \sim &\quad DP(\alpha, P_0),\\
\Lambda|G \sim &\quad W_G(\delta, D),\\
G \sim &\quad \pi_{\mathcal{G}},
\end{aligned} \tag{10}$$

$v \in V$, $i \in \{1, \dots, n\}$ et $\pi_{\mathcal{G}}$ est une distribution a priori sur l'espace des graphe $\mathcal{G}$.

Les avantages de cette stratégie entièrement bayésienne, sachant que le modèle de copule gaussien graphique est complètement général (Liu, Lafferty et Wasserman, 2009), sont liés à la dérivation de l'inférence à partir d'un échantillon de la distribution a posteriori de G. Tout d'abord, l'ensemble de la structure de dépendance est disponible, au lieu de mesures uniquement par paires. De plus, contrairement à la plupart des autres méthodes, l'incertitude peut être évaluée par des intervalles de crédibilité, en plus de la production d'une estimation de la force de la dépendance.

La probabilité que deux variables v_i et v_j soient dépendent peut être estimée :

— par des facteurs de Bayes sur la matrice de corrélation, avec des hypothèses de type $H_0 : |R_{v_i,v_j}| < \varepsilon$ contre des alternatives comme $H_1 : |R_{v_i,v_j}| \geq \varepsilon$ pour laquelle

$$B_{v_i,v_j} = Pr(H_1|X)/Pr(H_0|X) \tag{11}$$

est estimé simplement par le rapport de la proportion d'échantillons qui se situent respectivement dans l'hypothèse alternative et dans l'hypothèse nulle ; le seuil ε nous aide à contrôler le degré de précision nécessaire ;

— marginalement par la probabilité a posteriori d'inclusion de l'arête, soit $Pr(\{v_i, v_j \in V : (v_i, v_j) \in E\})$, calculé simplement comme la proportion d'échantillons du graphe qui contiennent l'arête (v_i, v_j).

Les applications de la méthode ci-dessus comprennent, entre autres, la détermination des réseaux dans les sciences sociales, l'économie et en particulier la biologie, où une multitude de données à *high-throughput* a émergé (données génétiques, transcriptomiques ou protéomiques). Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015, mais aussi Liu, Lafferty et Wasserman, 2009 ; Liu et al., 2012, sont de bonnes ressources où le modèle de copule gaussien graphique s'est montré capable de reproduire avec précision structures graphiques dans le cadre ci-dessus. Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015 en particulier, mis-à-part des légères approximations, partagent le même cadre de modélisation que celui de ce travail.

Nous allons comparer la méthode proposée avec celle de Filippi et Holmes, 2015 sur quelques jeux de données simulées en deux ou plus de dimensions, pour montrer où le modèle de Copule gaussien graphique peut être appliqué afin de détecter la dépendance entre variables.

Un autre domaine d'application, proposé par exemple dans Kunihama et Dunson, 2014, est la réduction de dimension des problèmes de régression. Dans la littérature du *Machine Learning*, ce problème a déjà été abordé avec des estimateurs de type RKHS (Fukumizu et al., 2007 ; Fukumizu, Bach et Jordan, 2009 ; Zhang et al., 2012). Ces estimateurs partagent cependant souvent la même difficulté d'être obligé d'évaluer plusieurs paires de relations. Dans la réduction de la dimension, cela peut se traduire par un algorithme très coûteux en temps de calcul si il est appliqué comme critère dans une procédure du sélection du meilleur sous-ensemble. Nous allons donc tester les performances d'une procédure d'Approximate Bayesian Computation (ABC, Marin et al., 2012, méthode bayésienne approchée) sur un modèle de coalescent proposé dans Joyce et Marjoram, 2008 ; Blum et al., 2013 ; Nunes et Balding, 2010 après avoir réduit la dimension du problème grâce au modèle de copule gaussien graphique.

Chapitre 1

Introduction

In the last decades, Bayesian statistics has seen a rise in popularity due to the large availability of powerful personal computer. Before that, adopting a Bayesian viewpoint often underlaid the use of simplified models so that an analytical solution was possible, limiting greatly the fields of application. This model trade-off between accurate description and explicit answers has since been overcome, thanks particularly to simulation based inference which allows the practitioner to devise samplers for the model under study which enable inference accurate to an arbitrary precision. In this thesis we study the formulation of two of these complex models under a fully Bayesian viewpoint and specifically the development of their corresponding samplers. Finally, especially in the last few years, we have seen the rise of a new trend where the so-called 'Big Data' and 'Big Models' are again challenging the capabilities of modern computers. We will hence study the properties of Delayed Acceptance, a modification of a classic sampler tailored for complex and computationally heavy models. The current chapter details the common framework on which all the followings will rely, introducing Bayesian statistics and simulation-based inference.

1.1 Bayesian statistics

Bayesian inference's eponym is Thomas Bayes ; in 1763 he derived the *inverse probability* distribution, what nowadays would be called *posterior distribution*, of the probability of success —θ, initially coming unobserved from a Uniform distribution— after a sequence of independent Bernoulli experiments (Bayes, 1763).

Contrarily to frequentist statistics, where probabilities are thought to be limiting relative frequencies, the Bayesian paradigm sees them as a measure of personal belief. Under this approach inference on some unknown parameter θ is conducted by combining not only the data x and their assumed statistical model $(\mathcal{X}, P_\theta(\cdot))$, but incorporating information available on θ prior to the data collection, in the shape of a probability distribution — the *prior distribution* — as well. We can attribute to the prior distribution most of both the arguments that are made in favour and the criticisms against Bayesian statistics. The intuitiveness of the update and the resulting decision-theoretic coherence of the Bayesian framework are an appalling feature to some but the risk of polluting the data with the formal need to incorporate such knowledge, which might not necessarily be present or meaningful, is one of the main controversies in the eyes of skeptics. We will now briefly introduce the principal

concepts necessary for the rest of this thesis ; for an exhaustive and more in-depth discussion on Bayesian statistics along with its associated advantages and criticisms refer for example to ROBERT, 2007.

Formally, define the data as a collection of n values $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, with the sample space $(\mathcal{X}, \sigma_{\mathcal{X}}, dx)$ defined as a measure space, and $\theta \in \Theta$ our parameter of interest [1]. The Bayesian interpretation of probability allows Θ to be formalized as a probability space $(\Theta, \sigma_\Theta, d\theta)$ and hence to endow θ with a probability distribution Π, the *prior* distribution ; the corresponding density with respect to the underlaying measure $d\theta$ will be denoted as π. Now, given the probability density of the data f with respect to dx, define the *likelihood* function as the probability of the data x conditioned on the parameter θ and denote it $\mathcal{L}(x|\theta)$.
Inference is finally conducted via the *posterior distribution* that summarize the 'updated' belief from the prior distribution after observing the data. Its density is obtained through the Bayes theorem as :

$$\pi(\theta|x) = \frac{\pi(\theta)\mathcal{L}(x|\theta)}{\int_\Theta \pi(\theta)\mathcal{L}(x|\theta)d\theta} \tag{1.1}$$

The denominator of (1.1) is usually called *marginal likelihood* or *evidence* and needs to be finite for the posterior distribution to be *proper* or well-defined.

Direct probability statements can now be made from the posterior distribution and, although a full description on how inference is then performed in this context would require an introduction to decision theory, which again can be found in ROBERT, 2007, it is safe to say that usually we are interested in computing functionals of the posterior distribution, most of the times integrals. The posterior expectation

$$\mathbb{E}(\theta|x) = \int_\Theta \theta\pi(\theta|x)d\theta$$

or the *maximum a posterior* (MAP)

$$\arg\max_{\theta\in\Theta} \pi(\theta|x)$$

are often used for parameter estimation and even hypothesis testing and predictions can in fact often be computed as integrals with respect to the posterior distribution.

Example 1.1.1. Consider the case where we are uncertain about the shape of f, the distribution of the data, but we can formulate more than one hypothesis $\mathcal{M}_i$ —say two for simplicity's sake in this example — with $i \in \mu = \{1, 2\}$. As stated above, uncertainty in Bayesian statistics calls to probability distributions and hence we formalise our extended model via a (discrete) prior $p(\mathcal{M}_i) = p_i$ and suppose thus that conditioned on model $\mathcal{M}_i$ we have that $x \overset{iid}{\sim} f_i(\cdot|\theta_i), \quad \theta \sim \pi_i(\theta_i) \in \Theta_i$.

The idea is now to get the marginal posterior distribution for each model, obtained thanks to Bayes theorem, which is

$$Pr(\mu = i|x) = \pi(\mathcal{M}_i|x) = \frac{p_i \int_{\Theta_i} \pi_i(\theta_i)\mathcal{L}_i(x|\theta_i)d\theta_i}{\sum_j p_j \int_{\Theta_j} \pi_j(\theta_j)\mathcal{L}_j(x|\theta_j)d\theta_j}. \tag{1.2}$$

1. Note that if we allow Θ to be an ∞–dimensional space, the same framework holds for non-parametric Bayesian statistics

Inference will proceed by either selecting the maximum, by averaging over all the possible models (mostly used in *prediction* problems) or by making use of an hypothesis testing framework and compute Bayes factors $B_{i,j}$ to quantify the support that the data give to one model with respect to the other :

$$B_{i,j} = \frac{\pi(\mathcal{M}_i|x)}{\pi(\mathcal{M}_j|x))} \Big/ \frac{p_i}{p_j}.$$

Each of this solutions involve the computation of at least the integral at the numerator of (1.2).

Problems arise though if either $\pi(\theta|x)$ is not available in closed form or if the associated integral has no analytical solution. As mentioned at the start this limited Bayesian statistics for a long time, confining its applicability to those cases where calculations were explicitly available, such as *conjugate* priors. The family of conjugate priors is made by distribution for which, after the update though the likelihood, the posterior distribution is again a member of the same family with updated parameters. While this family of distribution is definitely useful and still exists in many applications (like regression models), it is nonetheless restrictive and not fit for general problems.

There is hence a justified need for computational methods to help in these situations. Although deterministic numerical solvers like simple quadrature methods might seem a suitable solution, especially in low dimensional problems, the fact that they are completely unrelated with the probabilistic structure of the problem and their reliance on mathematical tools mostly unfamiliar to statisticians make them less attractive to practitioners then simulation-based methods [2].

The prominent solution, which is also the one adopted in this document, is to rely on Monte Carlo experiments ; we will define and explore them in the next section. We would finally like to mention though that other probabilistic-based solutions are available, Variational methods (Beal, 2003) amongst the most studied especially in the Machine Learning community, differing from Monte Carlo in the type of approximation that the analyst is willing to consider and in their computational complexity.

1.2 Simulation-based inference and Monte Carlo methods

1.2.1 Basic Monte Carlo

Let us formalise our problem first. We want to evaluate a quantity $\mathcal{I}$ defined as an integral of some function $\mathcal{H}$ defined on $\mathcal{X}$:

$$\mathcal{I} = \int_{\mathcal{X}} \mathcal{H}(x)dx$$

2. It is worth noting though that the use of the analytical form of the functions under study and their derivatives, typical of numerical methods, is starting to be more and more incorporated in simulation methods to increase their efficiency ; notable examples are Hamiltonian Monte Carlo methods (see for example Neal, 2012 for an introduction to those techniques).

If there exist a probability measure F with density f [3] and a function h defined on $\mathcal{X}$ so that

$$\mathcal{I} = \int_{\mathcal{X}} h(x)F(dx) = \int_{\mathcal{X}} h(x)f(x)dx = \mathbb{E}_F[h(x)] \tag{1.3}$$

then we can make use of the Law of Large Numbers (LLN), given that $\mathcal{I}$ exists and that we are able to get N *iid* samples from f, to approximate $\mathcal{I}$ by

$$\hat{\mathcal{I}}_N = \frac{1}{N}\sum_{i=1}^{N} h(x_i) \xrightarrow[N\to\infty]{} \mathcal{I} \tag{1.4}$$

where the convergence is at least in probability.

An even stronger result, obtained at the cost of assuming a finite second order moment for h under F is the Central Limit Theorem (CLT) :

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} h(x_i) - \mathcal{I}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma^2) \tag{1.5}$$

where σ^2 is the variance of h under f.

Monte Carlo experiments are exactly simulation studies that rely on samples from F to produce estimates. Following from the CLT, they are hence said to converge toward $\mathcal{I}$ at a rate of order $N^{-\frac{1}{2}}$.

Monte Carlo via Importance Sampling

Similar results can be applied to situations where the samples are obtained from a distribution G via Importance Sampling.

This technique is based on the equivalence

$$\mathcal{I} = \int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} h(x)w(x)g(x)dx = \mathbb{E}_G[h(x)w(x)] \tag{1.6}$$

where $w(x) = \frac{f(x)}{g(x)}$, under the condition that g is absolutely continuous with respect to f (or equivalently that F dominates G).

Importance Sampling (IS), where g is called the *importance function*, is then defined by substituting again the integration with its empirical equivalent :

$$\tilde{\mathcal{I}}_N = \frac{1}{N}\sum_{i=1}^{N} h(x_i)w(x_i) \qquad x_i \overset{iid}{\sim} G \tag{1.7}$$

Note that the convergence of (1.7) is guaranteed under conditions similar to (1.4), but in order to obtain sensible estimates one would want to make sure that the ratio f/g is bounded, in order for the variance of the approximation to be finite.

The use of Importance Sampling is justified not only by the difficulties connected with sampling from f, while g might be straightforward to sample from (see the next Section), but also by the fact that sampling from f might not be necessarily optimal in terms of the variance of the resulting estimator, *i.e.* $Var_G[h(x)w(x)] < Var_F[h(x)]$; see Robert et Casella, 2004 or Ripley, 1987 for example.

3. in all this text we will usually assume the existence of a density with respect to the distributions considered unless otherwise stated

Another advantage of IS that if the ratio f/g is only known up to a normalizing constant, which is often the case in Bayesian statistics where the marginal likelihood is not easily available, a normalised version of the estimator may be used since

$$\frac{1}{N}\sum_{i=1}^{N} w(x_i) \qquad x_i \overset{iid}{\sim} G \tag{1.8}$$

is a convergent estimator of the unknown normalising constant (for non-asymptotic behaviour see Cappé, Moulines et Rydèn, 2005) and hence

$$\frac{\sum_{i=1}^{N} h(x_i)w(x_i)}{\sum_{i=1}^{N} w(x_i)} \qquad x_i \overset{iid}{\sim} G \tag{1.9}$$

converges to the desired quantity. Note that there are cases where this self-normalised estimator is to be preferred in either case under bias-variance trade-off arguments (Casella et Robert, 1998).

All the above require nonetheless the ability to sample *independently* (for now) from either f or g and the next Section will detail some of the available methods. We would like to conclude by saying that even though the focus of this work is on Bayesian statistics and on computing integrals with respect to probability measures, simulation-based inference is not necessarily limited to this domain and optimization methods that rely on simulated samples are available and can be used to compute (for example) *maximum likelihood* estimators (Robert et Casella, 2004).

1.2.2 Independent Random Sampling

In order to make use of the results introduced in the previous Section we need to produce a sequence of random variables distributed according to some distribution F. In all the following we will call the distribution we want to draw samples from as the *target* distribution ; this definition will acquire a more formal and deep meaning in Sections 1.2.3.

The first and the only, but omnipresent, generator that we can find on a computer is the uniform generator. Even though the random variables produced by a machine are only *pseudo-random*, in that they are generated by a deterministic algorithm, we will assume that the uniform number generator available is an implementation of a modern algorithm like the Mersenne twister (Matsumoto et Nishimura, 1998) or any of its derivations, and hence that it can pass a collection of tests like the **Die Harder** battery (Brown, Eddelbuettel et Bauer, 2007). For a more in-depth discussion on the matter the reader is referred to Robert et Casella, 2004 ; Devroye, 1986 ; Matsumoto et Nishimura, 1998 and references therein.

So sampling from any uniform distribution $\mathcal{U}(0,1)$ should be quite straightforward. Assuming that our *target* density f is not a uniform though, we have a wide variety of techniques that can help us producing *independent* samples from almost any arbitrary distribution.

The first basic method is called Generic Inversion and relies on inverting the cumulative distribution F of the target distribution. If in fact we define the generalized

inverse $F^-(u) = \inf_x\{F(x) \geq u\}$ it is easily proven that

$$\text{If } U \sim \mathcal{U}_{[0,1]}, \text{ then } F^-(U) \sim F. \tag{1.10}$$

Even if (1.10) is very general and theoretically requires just one random sample from the uniform *per* random sample from F, it is rarely used in practice as it isn't as efficient as other methods and requires explicit knowledge of the cdf.

Still relying on the structure of the target distribution are the transformation methods, based on transformations of random variables. If the target distribution can be linked to another distribution which is easier to sample from we can often take advantage of this connection to draw samples from f. One notable example is the Box Muller algorithm (Algorithm 1.1) to sample from the normal distribution using again uniform variates :

Algorithm 1.1 Box–Muller Algorithm– Sample $X_1, X_2 \overset{iid}{\sim} \mathcal{N}(0,1)$

1: Sample $U_1, U_2 \overset{iid}{\sim} \mathcal{U}_{[0,1]}$;
2: Compute $X_1 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$;
3: Compute $X_2 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$.

TABLE 1.1: Some examples of General Transformation Method

Target Distribution	Transformation	Auxiliary Samples
$\mathcal{X}^2_{2\nu}$	$2\sum_{j=1}^{\nu} X_j$	$X_j \overset{iid}{\sim} \mathcal{E}xp(1)$
$\mathcal{G}am(a,b)$	$b\sum_{j=1}^{a} X_j$	$X_j \overset{iid}{\sim} \mathcal{E}xp(1)$
$\mathcal{B}eta(a,b)$	$\frac{X}{X+Y}$	$X \sim \mathcal{G}am(a,\theta),\ Y \sim \mathcal{G}am(b,\theta)$
t_ν	X/Y	$X \sim \mathcal{N}(0,1),\ Y \sim \mathcal{X}^2_\nu$
$\mathcal{F}(\nu_1,\nu_2)$	$\frac{X/\nu_1}{Y/\nu_2}$	$X \sim \mathcal{X}^2_{\nu_1},\ Y \sim \mathcal{X}^2_{\nu_2}$

A few examples of transformation methods are provided in Table 1.1, but unfortunately not every random variable can be expressed as a transformation of others. As a consequence we need to introduce now some methods where only a rough knowledge of f's functional form is needed, without needing to exploit any direct probabilistic feature of our target distribution.

The first of these methods is the Accept-Reject algorithm and its basic idea is, somehow similarly to Importance Sampling defined above, to rely on a simpler (to sample from) distribution G in order to obtain this time *unweighted* independent samples from f. Accept-Reject in its most simple form is based on the *Fundamental Theorem of Simulation*, a very simple idea that highlight the fact that $f(x)$ can be written as

$$f(x) = \int_0^{f(x)} du$$

where our target appears now as the marginal density of a joint uniform distribution $(X,U) \sim \mathcal{U}\left((x,u)|0 < u < f(x)\right)$.

The fundamental theorem proceeds then as follow :

Theorem 1.2.1. Sampling from

$$X \sim f$$

is equivalent to sample from

$$(X, U) \sim \mathcal{U}\left((x, u)|0 < u < f(x)\right)$$

and marginalise (discard the samples) with respect to U.

The idea is that if we can sample from the joint distribution of (X, U), which means uniformly sampling in the set $\{(x, u)|0 < u < f(x)\}$, the marginal X samples are already distributed according to our target, even if we used f only through some computations of $f(x)$.

Sampling exact pairs of (X, U) is not always easy nor sometimes feasible at all, so we will instead sample from a bigger set and keep the sampled pairs only if they satisfy the constraint. The distribution is indeed preserved by this procedure (Robert et Casella, 2004). To exemplify the method, think of a continuous one-dimensional target distribution defined on a compact interval $[a, b]$ and suppose that it is bounded by m, i.e. $f(x) \leq m \ \forall x \in [a, b]$. We can thus sample from $(X, U) \sim \mathcal{U}\left((x, u)|a < x < b,\ 0 < u < m\right)$ and further check and keep only the pairs that satisfy the constraint $0 < u < f(x)$; this is guaranteed to get us marginal samples $X \sim f$.

With the help of the aforementioned *instrumental* distribution g ($g \ll f$) this argument can easily generalize to settings different from a 'box' and for targets whose support or maximum are unbounded; consider the set

$$\mathcal{S} = \{(y, u)|0 < u < m(y)\}$$

where the function $m(x) = Mg(x)$ is such that $m(x) \geq f(x)$ for all x in the support of f (notice that $m(x)$ cannot be a probability density as it integrates to M). We can now sample uniformly from (U, Y) by drawing a value $y \sim g$ and then take $u|y \sim \mathcal{U}(0, m(y))$ and finally accept or reject the point iff $u < f(y)$.

The pseudo-code for this general Accept-Reject algorithm is presented in Algorithm 1.2 :

Algorithm 1.2 Accept-Reject Algorithm – Sample from f

1: Sample $x \sim g$ and $u \sim \mathcal{U}_{[0,1]}$;
2: **if** $u \leq f(x)/Mg(x)$ **then**
3: Accept x;
4: **else**
5: Reject and return to 1.
6: **end if**

Of course we would like $m(x)$ to be as close as possible to f in order to waste the least amount of simulation, even if some can be recycled via Importance Sampling; for more details on Accept-Reject and its derivatives see Robert et Casella, 2004 and therein references, like Casella et Robert, 1998.

There are cases though where independent sampling is overall impossible or too expensive in terms of computation; in this situations we would still like to sample *dependent* sequences $(X_n)_{n=1,\dots,N}$ such that their empirical average converges at a decent rate to the integral of interest, similarly to (1.4). It is in fact possible to prove that such sequences exist and particularly popular is the class of Markov Chains which we will briefly introduce in the following Section.

1.2.3 Markov chain Monte Carlo

A Markov Chain is a sequence of random variables (X_n) that satisfies the Markov property, i.e. that the immediate future of the sequence depends only on the present state and not on the past. Formally, we can write that for every measurable set A :

$$Pr(X_{n+1} \in A | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = Pr(X_{n+1} \in A | X_n = x_n).$$

The essential idea behind using Markov Chains to approximate integrals in the form of $\mathbb{E}_F[h(x)]$ is to build a time homogeneous transition kernel $K(x, y)$, that describe the evolution of the sequence, such that it has as an invariant distribution F, our target distribution ; with the addition on some other conditions we will then be able to obtain convergent estimators and in some case characterise their distribution (as in derive a CLT) as we did in Monte Carlo with independent samples.

Markov chain definitions and convergence

In order to accommodate both continuous and discrete targets we will use the measure-theoretic notation $Pr_F(E) = F(E) = \int_E F(dx)$ to indicate probability distributions and will assume that for every random variable a correspondent probability density function or probability mass function exists, denoted $f(x)$.

Given the definition of Markov property above, the Kernel K and an initial distribution $P_0(dx)$ we can describe the evolution of the chain by

$$Pr(X_{n+1} \in A | X_n = x_n) = \int_A K(x_n, dx)$$

$$Pr(X_0 \in A) = \int_A P_0(dx).$$

Note that we will only treat Markov models in *discrete time* and hence we will not describe *Markov processes*.

If we initialise the chain on an arbitrary point coming from P_0 we can thus completely specify the distribution of the chain by its Markov kernel K and the chain rule above ; this translate as well in the definition of the n-transition kernel $K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A)K(x, dy)$, with $K^1(x, A) = K(x, A)$.

We can now formalise the notion of *invariant distribution* introduced above as follows :

Definition 1.2.1. The distribution π on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is *invariant* for K or its associated chain if

$$\pi(A) = \int_{\mathcal{X}} K(x, A)\pi(dx)$$

for every $A \in \mathcal{B}(\mathcal{X})$.

The invariant distribution is often called *stationary* as well, as if $X_0 \sim \pi$ then $X_n \sim \pi \quad \forall n$ and hence the chain is said to be stationary in distribution. Remember that in the Markov chain Monte Carlo methodology the kernels are explicitly constructed such that the target distribution F is the invariant distribution of the

chain and hence all the Markov chains discussed from here on are assumed to satisfy this property.

Let us now introduce the critical conditions for our estimation procedure to work ; the first is the notion of *irreducibility*. A Markov chain is said to be F-irreducible if for any initial value there is a positive probability of entering any set that has a positive probability under F.

Definition 1.2.2. The Markov chain (X_n) is said to be F-irreducible, given a measure F on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, if for every $A \in \mathcal{B}(\mathcal{X})$ such that $F(A) > 0$, there exists n so that $K^n(x, A) > 0$ for all $x \in \mathcal{X}$.

This important property states informally that the Markov chain is not heavily influenced by its starting point and is crucial in practice to assure that we don't have to study in detail how to initialize the procedure, although of course for finite n the initialisation will matter in practice.

Another critical property is *aperiodicity* :

Definition 1.2.3. Given a F-irreducible Markov chain (X_n) defined on a countable space $\mathcal{X}$, its *cycle length* is defined as

$$d(x) = \gcd\{n \in \mathbb{N}^+ | K^n(x, x) > 0\}$$

and its period p is the largest cycle length. A chain is said to be aperiodic if $p = 1$.

The definition for uncountable spaces is slightly more involved and requires some extra definitions that aren't presented here as the main focus of this work lies elsewhere ; see Robert et Casella, 2004 ; Tierney, 1994 for a more in-depth discussion. Intuitively a chain is periodic if some portion of the space $A \in \mathcal{X}$ are only visited regularly with period $d(A)$ while aperiodicity assure us that this is not happening and the chain is not restricted in its movements at any time.

An irreducible an aperiodic chain is then guaranteed to visit every set A where our interest lies and in a non-deterministic fashion, but these properties alone do not provide us a formal assurance that the chain will not just pass through A and not visit it for more than a finite number of times. In this problematic case the chain is said to be *transient*, while we define it *recurrent* otherwise.

In general, if we define η_A the number of visit to the set A, we have that

Definition 1.2.4. An F-irreducible Markov chain is said to be *recurrent* if for every $A \in \mathcal{B}(\mathcal{X})$ with $F(A) > 0$, $\mathbb{E}[\eta_A] = \infty$ for every $x \in A$.

A chain being recurrent is thus very important if we want to asses the asymptotic properties of empirical means on A ; note also that for irreducible chains transience and recurrence are defined for the whole chain and not only for particular sets, with one of the two being inevitably satisfied.

An even stronger property called *Harris recurrence* can be enforced by requiring not only an infinite number of visits to every set A, but also an infinite number of visits for every specific sample path or realisation of the Markov chain.

Definition 1.2.5. A set is said to be *Harris recurrent* if $Pr(\eta_A = \infty) = 1$ for every $x \in A$ and an F-irreducible Markov chain is deemed *Harris recurrent* if for every set $A \in \mathcal{B}(\mathcal{X})$ with $F(A) > 0$, A is Harris recurrent.

Remember now that Markov chains constructed for our purposes are F-invariants and F-irreducibles ; if F is a σ-finite distribution the associated chain is called *positive* and it is possible to prove that this implies recurrence as well.

We can further associate these properties to *reversibility*, that implies the distribution of $X_{n+1}|X_n = x$ is the same of $X_{n+1}|X_{n+2} = x$, i.e. that direction of the discrete time index does not matter, by introducing a stronger condition that the one imposed in Definition 1.2.1 :

Definition 1.2.6. A Markov chain (X_n), with kernel K is said to satisfy *detailed balance condition* if a function f exists such that

$$K(x, y)f(x) = K(y, x)f(y) \qquad \forall (x, y).$$

If f is moreover a probability density we can affirm that
— The density f is the invariant density for the chain ;
— (X_n) is reversible.

The next property we are going to touch on is *ergodicity* ; loosely speaking ergodicity guarantee us that asymptoticly the chain will converge to its unique stationary behaviour regardless of its initialization and is thus at the hearth of convergence theory for Markov chain Monte Carlo methods. Again, a complete understanding of this feature especially under general state spaces require much more formalism that we are prone to include here and the reader is thus referred to Robert et Casella, 2004 ; Tierney, 1994 ; Meyn et Tweedie, 1993. We will just scratch the surface by noting two important results

Definition 1.2.7. If a Markov chain (X_n) is positive Harris recurrent and aperiodic, then

$$\lim_{n\to\infty} \left\| \int K^n(x, \cdot)P_0(dx) - f \right\|_{TV} = 0$$

where $\|\cdot\|_{TV}$ represent the Total Variation measure and f is the stationary distribution for (X_n).

which assure us that (X_n) is *ergodic* and is converging in Total Variation to its limiting distribution independently from its initial distribution P_0. Even more generally we can state that

Definition 1.2.8. If π is the invariant distribution for the chain (X_n), then

$$\left\| \int K^n(x, \cdot)P_0(dx) - f \right\|_{TV}$$

is decreasing as n increases.

As a consequence of these results we can also prove that the convergence holds for expectations of functions under the distribution of the chain initialised via P_0 and under its stationary distribution f

$$\lim_{n\to\infty} \left| \mathbb{E}_{P_0}[h(X_n)] - \mathbb{E}_f[h(X)] \right| = 0$$

for every bounded function h.

This somehow assure us, at least in probabilistic terms, of the fact that is possible to use Markov chains as a substitute for independent sampling in our inferential procedures where the latter would not be possible, allowing us to treat a broader set of problems. Before formally defining the Law of Large Number and Central Limit Theorem counterpart for Markov chain though is worth spending a few more time analysing not only the fact that the chain converges in distribution, but also the *speed* of this convergence. This will also be crucial in the subsequent understanding of the ordering of the chains in terms of performances that will help us in choosing the method to use in practice.

There are two main quantification of rate of convergence for Markov chain Monte Carlo methods, which are as follows :

Definition 1.2.9. An ergodic Markov chain with limiting distribution f is said to be *geometrically ergodic* if there exists a non-negative function M with $[E][|M(x)|] < \infty$ and a positive constant $r > 1$ such that

$$\left\| \int K^n(x, \cdot) P_0(dx) - f \right\|_{TV} \leq M(x) r^{-n}$$

for all x.

Definition 1.2.10. An ergodic Markov chain with limiting distribution f is said to be *uniformly ergodic* if there exists a positive constant $M < \infty$ and a positive constant $r > 1$ such that

$$\left\| \int K^n(x, \cdot) P_0(dx) - f \right\|_{TV} \leq M r^{-n}$$

for all x, or alternatively

$$\lim_{n \to \infty} \sup_{x \in \mathcal{X}} \left\| \int K^n(x, \cdot) P_0(dx) - f \right\|_{TV} = 0.$$

Ergodic theorem and CLT for Markov chains

The last concept we need to address is then the convergence of averages of a single, specific *sample path* of the chain, rather than convergence at a/each fixed point in time. We shall hence introduce at first the so-called *ergodic theorem* that examine asymptotic behaviour of partial sums from a single chain.

Theorem 1.2.2 (Ergodic Theorem, (Robert et Casella, 2004, Theorem 6.63))**.** If (X_n) is a positive Harris recurrent Markov chain with invariant measure f, then for every $h \in L_1(f)$ we have that

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i) \xrightarrow[N \to \infty]{} \int h(x) f(dx).$$

The samples are dependent this time thought and there is thus no natural consequence that lead to a formulation of the Central Limit Theorem ; careful examination of the procedure and of its properties are required in order to approach convergence in distribution of partial sums. We will present here only one version of such a theorem which relies on the assumption of geometric ergodicity described above.

Theorem 1.2.3 (Markov Functional Central Limit Theorem, (Roberts et Rosenthal, 1997 Corollary 2.1 or Robert et Casella, 2004, Theorem 6.67)). If (X_n) is a positive Harris recurrent and irreducible Markov chain, geometrically ergodic with invariant measure f and if the function h satisfies $\mathbb{E}_f[h(x)] = 0$ and $\mathbb{E}_f\left[\left|h(x)\right|^{2+\varepsilon}\right] < \infty$, $\varepsilon > 0$, (or $\mathbb{E}_f[h(x)^2] < \infty$ if additionally the chain is reversible) we have that $\varsigma_h^2 < \infty$ and

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} h(X_i)\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, \varsigma_h^2).$$

It is useful to note that ς_h^2 is not in this case simply the variance of h under the distribution f, as it would be for independent sampling, but instead comprehend all the lag-k auto-covariances of the Markov chain. That is to say

$$\varsigma_h^2 = \mathbb{E}_f[h(X_0)^2] + 2\sum_{k=1}^{\infty} \mathbb{E}_f[h(X_0)h(X_k)] \tag{1.11}$$

Other similar version of this results are available, the most notables relying on the reversibility assumption ; see Robert et Casella, 2004 for a quite comprehensive review.

To conclude we have shown that non-*iid* samples, as is generally the case for the ones produced by Markov chains, can indeed be used in order to approximate integrals of interest and theoretical findings like the Central Limit Theorem for Markov chain provide accurate justifications for Markov chain Monte Carlo methods, even though each sample path might only be distributed according to the target asymptotically. The applicability of these procedure is thus only limited by their implementation and computational costs.

Metropolis–Hastings

Now that we have established the possibility of using sample paths obtained from a Markov chain in order to perform simulation-based inference, we are going to present one general algorithm widely used in practice. Chapter 2 will then introduce and examine a variation of this algorithm developed to meet the increasing computational challenges faced by MCMC methods.

This algorithm namesakes are N. Metropolis, who introduced a first version of the method in Metropolis et al., 1953 particularly for canonical ensemble in physics and Hastings who later generalised it in Hastings, 1970 and spread its use to the statistical community. It is one of the most general Markov chain Monte Carlo methods and probably the most used in practice due to its straightforward implementation and large flexibility. The algorithm is presented in Algorithm 1.3.

Metropolis–Hastings uses a proposal distribution q and an acceptance step with a given probability to form its Markov kernel, which is defined as

$$K(x, y) = a(x, y)q(y|x) + (1 - \rho(x, y))\,\delta_x(y)$$

with $\rho(x, y) = \int a(x, y)q(y|x)dy$ and $\delta_x(\cdot)$ indicating a Dirac mass on x. As shown in Metropolis et al., 1953 this move preserves the stationary distribution f as long as the chain is f-irreducible, that is as long as the proposal density allows the

Algorithm 1.3 Metropolis–Hastings algorithm

Given an initial state $X_0 \sim P_0$, a number of desired iterations N and a proposal distribution q :
for n in 1, ..., N **do**
    Sample $X^* \sim q(\cdot|X_{n-1})$ and $u \sim \mathcal{U}_{[0,1]}$;
    Compute the acceptance probability as $a = \min\{1, r(X^*, X_{n-1})\}$, where

$$r(X^*, X_{n-1}) = \frac{f(X^*)}{f(X_{n-1})} \times \frac{q(X_{n-1}|X^*)}{q(X^*|X_{n-1})};$$

    if $u \leq a$ **then**
        Accept the proposed point and set $X_n = X^*$;
    else
        set $X_n = X_{n-1}$.
    end if
end for

chain to reach any point in the state-space which has positive probability under f. The real potential of the algorithm is thus the fact that, similarly to Accept-Reject, we can sample from almost any proposal distribution q and turn these proposals into valid samples from our target distribution. As is the case for Accept-Reject though, the efficiency of the algorithm and its performances are strictly related to how *good* the proposal mechanism is in relation to our target. Another important shared property is the requirement to know both the target f and the proposal q just up to a normalising constant ; their ratio will in fact simplify when computing the acceptance probability. This is again particularly helpful in Bayesian statistics when the target is the posterior distribution.

The two most widely used proposal mechanism are probably *independent proposals* and *random walk proposals*. In the first case a random point is proposed with distribution $X^* \sim q(\cdot)$ independently from the state the chain is in and we thus need to ensure that the proposal dominates the target distribution and at the same time matches it as closely as possible in order to ensure good behaviour. In the second case instead a point is chosen from a distribution such that $q(X^*|X_{n-1}) = q(X_{n-1}|X^*)$ where X^* is the proposed point and X_{n-1} the current state. In this case the proposal ratio simplifies in the acceptance probability and performances are dictated by the ability of this proposal to quickly explore the space $\mathcal{X}$ without either rejecting too many 'far' proposals or making too many small and highly correlated steps.

It is easily proven that the Metropolis–Hastings chain is aperiodic, irreducible and Harris recurrent under mild condition on the proposal distribution q. We can thus rely on the Ergodic Theorem 1.2.2 to state that under those easily met condition a Metropolis–Hastings chain (X_n) is ergodic.

To better specify the rate of convergence we will have to specify a bit more the proposal distribution q. In the case of independent Metropolis–Hastings it is in fact possible (Mengersen et Tweedie, 1996) to prove that the chain is uniformly ergodic if there exists a constant M such that $f(x) < Mq(x)$ for every $x \in supp(f)$.

Random-walk Metropolis–Hastings does not share uniform ergodicity with the independent version, to little surprise if we consider that the method is *local* in nature, but it is possible to obtain both sufficient and necessary conditions for geo-

metric ergodicity. Mengersen et Tweedie, 1996 shows in fact that for a particular class of *symmetric log-concave targets* and for bounded proposals geometric ergodicity is possible to prove (Theorem 7.15 in Robert et Casella, 2004).

More specific and involved proposal, such as gradient informed proposals (Neal, 2012 ; Girolami et Calderhead, 2011) or extensions to double-intractable target distributions (Andrieu et Roberts, 2009) are constantly being developed and convergence properties for these methods are an active field of research ; see for example Livingstone et al., 2016.

Ordering of the chains

Even though we only looked at Metropolis–Hastings it is quite clear already that the general formulation of Markov chain Monte Carlo allows for a wide variety of algorithm ; just defining the proposal distribution in the above method gives us almost infinite possible choices. How do we decide then which method suits us the most ? While deciding the *tuning parameters* for a given family of distribution q is a matter of optimisation rather than pure choice (see for example Roberts, Gelman et Gilks, 1997 for a seminal paper in this field), selecting the shape of the proposal distribution or even the type of algorithm to use altogether require us to define an *ordering* between Markov chain Monte Carlo methods.

Let us assume reversibility for all the chains described hereafter. Even though in general *reversibility* is a quite strong assumption, when building Markov chains Monte Carlo algorithms we often times construct the kernel such that the chain satisfies detailed balance condition (Definition 1.2.6) and is thus reversible. In these cases, as long as $\varsigma_h^2 < \infty$ we are able to make use of the Central Limit Theorem to asses performances much more like we would do with basic Monte Carlo Methods. We will hence use this asymptotic variance, named $v(h, K)$ to stress its dependence on the chosen kernel K, to define our criterion for ranking the efficiency of different procedures.

Following for example Tierney, 1998 we introduce two related definitions ; let $L_0^2(F)$ be the subspace of all L^2 functions with respect to F with zero mean, i.e. $L_0^2 = \{h \in L^2(F) | \int \int h(x)F(dx)\}$, then

Definition 1.2.11. If (X_n) and (Y_n) are Markov chains with kernel K and L respectively and stationary distribution F, then we say that $K \succeq_E L$ (read K is at least as efficient as L) if

$$v(h, K) \leq v(h, L) \qquad \forall h \in L_0^2(F).$$

Definition 1.2.12. If (X_n) and (Y_n) are Markov chains with kernel K and L respectively and stationary distribution F, then we say that $K \succeq L$ (read K dominates L off-diagonal) if for F-almost all points $x \in \mathcal{X}$ we have

$$K(x, A \setminus x) \geq L(x, A \setminus x)$$

for all measurable sets A.

This last definition is easy to understand in the case of discrete state-spaces, which is where it originates in Peskun, 1973a, where the transition kernel K is in fact a transition *matrix* and Definition 1.2.12 implies that every off-diagonal element of K is greater than the corresponding element in L. Tierney, 1998 later generalised this notion to general state-spaces and in particular derived the following theorem :

Theorem 1.2.4 (Peskun/Tierney ordering, (Tierney, 1998))**.** If (X_n) and (Y_n) are Markov chains with kernel K and L respectively and stationary distribution F, then

$$K \succeq L \implies K \succeq_E L.$$

Jointly, these properties define the so-called Peskun-Tierney's ordering.

One last important definition to understand the results presented in Chapter 2 is the *(right) spectral gap* of a Markov kernel operator.

Definition 1.2.13. Consider K as an operator on $L^2(F)$ and call its *spectrum* the set $\sigma(K)$ of the λ's such that $\lambda I - K$ is not invertible, where I is the identity operator.

Notice that this is easily applied to discrete state-spaces as the spectrum of an operator on a finite-dimensional space is precisely the set of eigenvalues of the kernel matrix. For general spaces instead the spectrum might include elements other than the eigenvalues, or might not contain them at all. In both cases the spectrum is a non-empty subset of the interval $[-1, 1]$. In Roberts et Rosenthal, 1997 the authors show that for reversible geometrically ergodic chains all the values of λ are bounded away from the extremes of the interval except for the principal $\lambda_0 = 1$ and we will hence not consider the value λ_0, associated with constant functions, when considering spectra of kernel operators.

Define now $\lambda^K_{\max} = \sup_{\lambda \in \sigma(K)} \{\lambda\}$ and $\Lambda^K_{\max} = \sup_{\lambda \in \sigma(K)} \{|\lambda|\}$.
$\Lambda^K_{\max}$ is sometimes called the *spectral radius*.

Definition 1.2.14. The quantity $1 - \Lambda^K_{\max}$ is called *spectral gap* and if the transition kernel K is such that $1 - \Lambda^K_{\max} > 0$ we say that K has a (right) spectral gap.

This quantity is important for us as Theorem 2.1 in Roberts et Rosenthal, 1997 proves that for F-invariant reversible Markov chain with kernel K, having a spectral gap is equivalent to be geometrically ergodic and the spectral gap is further connected to the asymptotic variance ς^2_h. This definition has been moreover used for example in Andrieu, Lee et Vihola, 2013 to inherit properties and asymptotic behaviour for chains derived from simpler Metropolis–Hastings.

Adaptive MCMC

As introduced at the start of the previous Section most of Markov chain Monte Carlo algorithms have *tuning parameter* to adjust in order to control properties and performances of the method. One of the most prominent examples is the choice of the covariance matrix in random-walk Metropolis–Hastings. Remember from Section 1.2.3 that in this kind of procedure the proposal distribution is such that $q(y|x) = q(x|y)$ but no other assumption on q has been made. In order to keep the presentation clean we will assume that $\mathcal{X} \subset \mathbb{R}^d$ and that $q(\cdot|x) \sim \mathcal{N}_d(x, \Sigma_q)$, with Σ_q being essentially the only free parameter to tune.

Roberts, Gelman et Gilks, 1997 showed that, asymptotically in the dimension d, the optimal proposal performs jumps in the state space such that the acceptance rate of the chain, defined as the expectation of the acceptance probability $\alpha = \mathbb{E}_K[a(x, y)]$, is close to $\alpha^* = 0.234$. They moreover show that such an acceptance rate is obtained by tuning Σ_q close to $\frac{(2.38)^2}{d} \times \Sigma$ where Σ is the unknown

covariance of the target distribution. While being only asymptotically true and obtained under some quite restrictive assumptions on both f and q, the result has been generalized multiple times (Roberts et Rosenthal, 2001 ; Breyer et Roberts, 2000 ; Sherlock et Roberts, 2009 ; Neal et Roberts, 2011) and shown to hold with little variation under less restrictive conditions.

While a very potent building block for all the optimisation and adaptation literature since, this still does not offer us an operative criterion to *automatically* tune our chain without resorting to run the procedure multiple times until a suitable parameter is found. We could thus be tempted at first to devise an 'adaptive algorithm' that start the chain on a particular value $\Sigma_q^{(0)}$ and then adapt Σ_q at iteration n ($\Sigma_q^{(n)}$) based either on the *observed acceptance rate* or on the *sample covariance*, both obtained on the past iterations ; say

$$\Sigma_q^{(n)} = g\left(X_1, \ldots, X_{n-1}\right). \tag{1.12}$$

While this idea is in fact successful, we need to put particular care on how this adaptation is done.

It is in fact obvious that this violate the assumption of *homogeneity* of the Markov kernel, which is indeed changing at each iteration now as the proposal is changing based on (1.12), and this easily jeopardise especially stationarity and ergodicity of the chain.

Specifically, let $\{K_\theta\}_{\theta\in\Theta}$ be a family of Markov kernels indexed by the parameter θ, all having the same stationary distribution F, and name $K^{(n)} = K_{\theta^{(n)}}$ the kernel chosen for the *n-th* iteration, with parameter $\theta^{(n)}$, that could potentially be selected based on the whole history of the chain $(X_0, \ldots, X_{n-1})$. Each kernel allowing for F as an invariant distribution is not sufficient to guarantee that the chain $\left\{(X_n, K^{(n)}\right\}$ converges to F, but in recent years many authors have provided sufficient and necessary conditions to preserve the ergodicity of such a procedure (see for example Haario, Saksman et Tamminen, 2001 ; Atchade et Rosenthal, 2005 ; Andrieu et Moulines, 2006).

We will in particular focus on the proposal by Haario, Saksman et Tamminen, 2001, which fits the setting given above, analysed through the condition given in Rosenthal et Roberts, 2007. In the latter the authors prove that both asymptotic convergence of the distribution of the chain and the weak Law of Large Numbers for empirical averages of the produced samples can be verified under two quite intuitive and easy to verify hypothesis, namely :

— **Diminishing adaptation** :

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \left\| K^{(n)}(x,\cdot) - K^{(n-1)}(x,\cdot) \right\|_{TV} = 0 \qquad \text{in probability} ;$$

— **Containment** : the convergence time for each kernel $K^{(n)}(X_n, \cdot)$ started in X_n, for each n, must be bounded in probability.

While *containment* is a rather technical condition verified for nearly any sensible adaptive scheme, *diminishing adaptation* is as much intuitive as it is fundamental. This condition says that the 'amount' of adaptation must decrease as the chain progresses and vanish to 0 in the limit. The sum of all the adaptations made can still diverge toward ∞, but each individual contribution must disappear in time.

An example will help clarifying the above definitions ; consider again the Random–Walk Metropolis–Hastings with Normal proposals with scale parameter $\Sigma_q^{(n)}$. It was

shown in Roberts, Gelman et Gilks, 1997 that the optimal proposal should have $\Sigma_q = \frac{(2.38)^2}{d} \times \Sigma$ but as Σ is of course unknown, let us substitute it with the sample covariance S_n obtained by using all the sample path till iteration n. The proposal at step $n+1$ is thus

$$X^* \sim \mathcal{N}(X_n, \frac{(2.38)^2}{d} S_n).$$

The empirical estimates S_n changes at iteration n by a factor of order $\mathcal{O}(\frac{1}{n})$, which is decreasing in n, and thus satisfy *diminishing adaptation*. Some little adjustment to the diagonal of $\Sigma_q^{(n)}$ is usually made in order to avoid collapse of the proposal covariance, which would violate *containment*, refer to Haario, Saksman et Tamminen, 2001 ; Roberts et Rosenthal, 2009. This simple yet effective algorithm is thus (maybe surprisingly) a valid ergodic Markov chain under very mild conditions.

Andrieu et Moulines, 2006 proves geometric ergodicity for this class of problems under a set of slightly more stringent conditions and there are of course many other types algorithms that fall under the umbrella of 'Adaptive MCMC' see for example for a more extensive review Roberts et Rosenthal, 2009 or Atchade et al., 2009.

1.2.4 Sequential Monte Carlo

Finally, as there will be multiple references to Sequential Monte Carlo throughout the text, we are going to introduce this sampling method as well.

As the name implies Sequential Monte Carlo is devised to sample from a *sequence* of (closely related) distributions. Even though it was developed with dynamic models in mind, in particular in Gordon, Salmond et Smith, 1993 to approximate a sequence of *filtering* distributions for a state-space model, they are widely applied now even to 'static' models.

The first thing that needs to be done is then defining the sequence of distribution under study, called $\{\pi_t\}$; in dynamic models this usually follows spontaneously from the definition of the model, while when only one (static) distribution is our target we are presented with a choice. In Bayesian statistics we could for instance define a smooth path between the prior distribution and the posterior by *tempering* the likelihood (e.g. by defining the sequence as $\{\pi(x) \times (\mathcal{L}(y|x))^{\gamma_t}\}_{t=1}^T$, such that $\gamma_1 = 0$ and $\gamma_T = 1$) ; if the likelihood over n observations factorise into n terms, in the *iid* case for example, one other natural way of expressing the posterior distribution is through the sequence $\{\pi(x|y_{1:t})\}_{t=0}^n$, with the convention that $y_{1:0} = \varnothing$. These ideas were first introduced by Neal, 2001 and Chopin, 2002 respectively ; in this work we are going to focus only on this static case and one case of Sequential Monte Carlo with tempering is presented in Chapter 3.

Now that the target sequence is outlined, the main idea is to sample from the starting distribution at first and sequentially 'mutate' that sample into a sample from the next distribution in the sequence, stopping when the final target distribution is reached. This suggest already that *Importance Sampling* might have a connection with Sequential Monte Carlo and the first sequential method that we will describe is in fact a very simple *Sequential Importance Sampling* (Cappé, Moulines et Rydèn, 2005).

We start by sampling $x_0^{(i)} \sim \pi_0, \quad i = 1, \ldots, P$ i.e. P elements from the *initial* distribution, usually the prior, and we call them *particles*. We then define weights for

all the particles $W_0^{(i)} \propto 1$ since the target for this iteration is the same distribution we sampled from. At the next iteration we re-weight our population of particles according to the next distribution π_1 by defining $W_1^{(i)} \propto W_0^{(i)} \times \frac{\pi_1(x_0^{(i)})}{\pi_0(x_0^{(i)})}$. The weighted particles $(x_0^{(i)}, W_1^{(i)})$ are now a sample distributed as π_1. We iterate the procedure until the target distribution $\pi = \pi_T$ is reached.

It is not hard to see that this procedure rapidly degenerates because of all the weights approaching 0, giving raise to empirical averages with infinite variance that cannot be used in order to perform inference. We thus need to generalise the method adding additional steps that help us in avoiding degeneracy of the weights and at the same time help us maintain diversity in the particle system.

The first thing we might do is mutating the particles at each iteration by inducing a move via Markov kernel K_t which leaves the target distribution at each iteration invariant. This can also be generalised into using a generic kernel that target an auxiliary distribution if we correct accordingly for this in the weights similarly to how we did in Importance Sampling in Section 1.2.1. This allow for patcle diversity as well as making meaninfull the transition to the intermediate distributions and is exactly the algorithm presented in Cappé, Moulines et Rydèn, 2005.

The product of the weights over multiple iterations will still lead a large number of particles to have null weights. We can formalise the degeneracy concept by defining the concept of *Effective Sample Size* (ESS)

$$ESS_t = \frac{\left(\sum_{i=1}^{P} W_t^{(i)}\right)^2}{\sum_{i=1}^{P} (W_t^{(i)})^2}. \tag{1.13}$$

At each iteration ESS_t will be between P if all the particles have equal weight and 0 if for all the particles $W_t^{(i)} = 0$ (degeneracy). We avoid ESS falling to zero by introducing a resampling step.

Douc et Cappé, 2005 presents a wide variety of resampling algorithms for this specific problem and, althought we will not describe them here, we can conceive the resampling step as multinomial sampling from the particle system with probabilities dictated by their weights. What we obtain at the end is an equally weighted sample distributed according to the same distribution as before ; a sample where we will have abandoned some of the less probable particles and duplicated some of the most important. This introduces some additional noise in the procedure and is hence advisable to perform the resampling only when ESS_t falls below a certain threshold.

Finally, in the case of tempering, the temperature schedule $\gamma_{1:T}$ is of course of crucial importance. It is actually the case that we can decide it 'on the fly' at each iteration by examining how distant are two consecutive distribution and specifically how would the current sample fare as an approximation for the next distribution in the sequence. This can be achieved for example through the conditional Effective Sample Size (cESS) defined in Zhou, Johansen et Aston, 2013.

We present a quite general formulation in Algorithm 1.4.

Algorithm 1.4 SMC (tempering) Sampler

Given α, β and P ;
Set $\gamma = 0$, $t = 0$;
for $i \in 1, \dots, P$ **do**
    – $x_0^{(i)} \sim \pi(\cdot)$; $W_0^{(i)} = \frac{1}{P}$.
end for
while $\gamma < 1$ **do**
    $t \leftarrow t + 1$; Compute γ^* such that $cESS_{\gamma^*} = \alpha \times cESS_\gamma$;
    for all i **do**
        – Compute $W_t^{(i)} \propto W_{t-1}^{(i)} \times \frac{\mathcal{L}(y|x_{t-1})^{\gamma^*}}{\mathcal{L}(y|x_{t-1})^{\gamma}}$;
        – $(x_t, W_t)^{(i)} \leftarrow (x_{t-1}, W_t)^{(i)}$; $\gamma \leftarrow \gamma^*$;
    end for
    Compute $e = ESS(t)$;
    if $e < (\beta \times P)$ **then**
        – $\tilde{x}_t^{(i)} \sim \sum_{i=1}^{P} W_t^{(i)} \delta_{x_t^{(i)}}$; $\tilde{W}_t^{(i)} = \frac{1}{P} \quad \forall i$
        – $(x_t, W_t)^{(i)} \leftarrow (\tilde{x}_t, \tilde{W}_t)^{(i)} \quad \forall i$
    end if
    for $i \in 1, \dots, N$ **do**
        – $x_t^{(i)} \sim K_t(\cdot|x_t^{(i)})$;
    end for
end while

Before concluding we would like to mention that, given that the target distribution effectively change at each iteration, sequential methods do not share the same difficulties encountered in Markov chains with respect to adaptation. We can in fact freely tune the current kernel based on the whole set of particles of previous iterations. See for example Fearnhead et Taylor, 2010.

There is of course more to sequential methods then what we shortly introduced here, a good introductory reference which trace sequential methods from their origin to recent days is Doucet et Johansen, 2009 and more specific references can be found inside. In particular a more general and recent formulation of Sequential Monte Carlo methods with an accurate analysis of its theoretical properties based on Feynman–Kac representations was introduced in Del Moral, 2004 ; Del Moral, Doucet et Jasra, 2006.

1.3 Overview

This thesis is further developed in three main chapters ; what follows is a short description of each of them :

1.3.1 Accelerating Metropolis–Hastings algorithms by Delayed Acceptance

In Chapter 2 we study a variation of the classic Metropolis–Hastings algorithm that allows for a reduced computational costs. MCMC algorithms such as Metropolis–Hastings algorithms are in fact know to struggle when paired with complex target distributions, as exemplified by huge datasets, as the cost is always at

least linear in the number of data points. We present a useful generalisation of the Delayed Acceptance approach, devised to reduce the computational costs thanks to a simple and universal divide-and-conquer strategy. The idea behind the generic acceleration is to divide the acceptance step into several parts, aiming at a major reduction in computing time that out-ranks the corresponding reduction in acceptance probability. Each of the components can be sequentially compared with a uniform variate, the first rejection signalling that the proposed value is considered no further. We develop moreover theoretical bounds for the variance of associated estimators with respect to the variance of its 'parent' Metropolis–Hastings and detail some results on optimal scaling and general optimisation of the procedure.

1.3.2 Bayesian Dimension Expansion : modelling nonstationary spatial processes

In Chapter 3 we will propose to set the Dimension Expansion technique, an optimisation procedure devised in Bornn, Shaddick et Zidek, 2012 to model nonstationary environmental processes, in a fully Bayesian framework ; this allows for precise quantification of the connected uncertainty, contrarily to the previous algorithmic proposal. We formally introduce the method directly in a probabilistic setting and derive an efficient sampler to enable inference on it. We aknowledge and discuss its computational pitfalls, describing two possible low-rank or sparse approximations that help the method to scale up to very high dimensions. We apply it finally on both simulated and real data in both its exact formulation when possible and using the approximated designs on larger problems.

1.3.3 Bayesian inference on dependency structure via the Gaussian Copula graphical model

Finally 4 constructs a fully Bayesian MCMC algorithm that performs exact inference on a Gaussian Copula graphical model, a default choice for analysing dependence in multivariate data, with the additional advantage of permitting rigorous density estimation. This work relies on recent results on both sampling and exactly computing desities for G-Wishart variates, used to model precision matrices. The graphical layer in this modelling strategy leads to complete inference about the conditional dependence structure of the data ; it allows for both a flexible modelling procedure and testing for dependence between variables. In contrast with alternative methods, we avoid plug-in estimators and approximations, except for unavoidable model misspecification. Moreover, the resulting testing procedure is moreover not limited to pairs of variables.

References

Andrieu, C. et G.O. Roberts (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In : 37.2, p. 697–725.

Andrieu, Christophe, Anthony Lee et Matti Vihola (2013). "Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers". In : *arXiv preprint arXiv:1312.6432*.

Andrieu, Christophe et Éric Moulines (2006). "On the ergodicity properties of some adaptive MCMC algorithms". In : *Ann. Appl. Probab.* 16.3, p. 1462–1505. issn : 1050-5164.

Atchade, Y. F. et J. S. Rosenthal (2005). "On adaptive markov chain monte carlo algorithm". In : *Bernoulli* 11.5, p. 815–828.

Atchade, Yves et al. (2009). "Adaptive Markov chain Monte Carlo: theory and methods". In : *Bayesian Time Series Models*, p. 33–53.

Bayes, T. (1763). "An Essay Toward Solving a Problem in the Doctrine of Chances". In : *Philosophical Transactions of the Royal Society of London* 53, p. 370–418.

Beal, Matthew James (2003). *Variational algorithms for approximate Bayesian inference*. University of London London.

Bornn, Luke, Gavin Shaddick et James V. Zidek (2012). "Modeling Nonstationary Processes Through Dimension Expansion". In : *Journal of the American Statistical Association* 107.497, p. 281–289. doi : 10.1080/01621459.2011.646919.

Breyer, L.A. et G.O. Roberts (2000). "From metropolis to diffusions: Gibbs states and optimal scaling". In : *Stochastic Processes and their Applications* 90.2, p. 181 –206. issn : 0304-4149. doi : http://dx.doi.org/10.1016/S0304-4149(00)00041-7. url : http://www.sciencedirect.com/science/article/pii/S0304414900000417.

Brown, Robert G, Dirk Eddelbuettel et David Bauer (2007). "dieharder: A Random Number Test Suite". In : *URL http://www. phy. duke. edu/˜ rgb/General/dieharder. php. C program archive dieharder*.

Cappé, O., E. Moulines et T. Rydèn (2005). *Inference in Hidden Markov Models*. New York : Springer-Verlag.

Casella, G. et C.P. Robert (1998). "Post-processing Accept–Reject samples: recycling and rescaling". In : *J. Comput. Graph. Statist.* 7.2, p. 139–157.

Chopin, N. (2002). "A sequential particle filter method for static models". In : *Biometrika* 89, p. 539–552.

Del Moral, P. (2004). *Feynman-Kac formulae*. Probability and its Applications. New York : Springer-Verlag, p. xviii+555.

Del Moral, P., A. Doucet et A. Jasra (2006). "Sequential Monte Carlo samplers". In : 68.3, p. 411–436.

Devroye, Luc (1986). *Nonuniform random variate generation*. New York : Springer-Verlag, p. xvi+843. isbn : 0-387-96305-7.

Douc, Randal et Olivier Cappé (2005). "Comparison of resampling schemes for particle filtering". In : *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*. IEEE, p. 64–69.

Doucet, Arnaud et Adam M Johansen (2009). "A tutorial on particle filtering and smoothing: Fifteen years later". In : *Handbook of Nonlinear Filtering* 12.656-704, p. 3.

Fearnhead, P. et B. M. Taylor (2010). "An Adaptive Sequential Monte Carlo Sampler". In : *ArXiv e-prints*. arXiv : 1005.1193 [stat.CO].

Girolami, M. et B. Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, p. 123–214.

Gordon, N., D. Salmond et A. F. Smith (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In : *IEE Proc. F, Radar Signal Process.* 140, p. 107–113.

Haario, Heikki, Eero Saksman et Johanna Tamminen (2001). "An adaptive metropolis algorithm". In : *Bernoulli* 7.2, p. 223–242. issn : 1350-7265.

Hastings, W.K. (1970). "Monte Carlo sampling using Markov Chains and their applications". In : *Biometrika* 57.1, p. 97–109.

Livingstone, S. et al. (2016). "On the Geometric Ergodicity of Hamiltonian Monte Carlo". In : *ArXiv e-prints*. arXiv : 1601.08057 [stat.CO].

Matsumoto, Makoto et Takuji Nishimura (1998). "Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator". In : *ACM Trans. Model. Comput. Simul.* 8.1, p. 3–30. issn : 1049-3301. doi : 10.1145/272991.272995. url : http://doi.acm.org/10.1145/272991.272995.

Mengersen, K.L. et R.L. Tweedie (1996). "Rates of convergence of the Hastings and Metropolis algorithms". In : 24, p. 101–121.

Metropolis, N. et al. (1953). "Equations of state calculations by fast computing machines". In : *J. Chem. Phys.* 21.6, p. 1087–1092.

Meyn, S. P. et R. L. Tweedie (1993). *Markov chains and stochastic stability*. London : Springer-Verlag London Ltd., p. xvi+ 548. isbn : 3-540-19832-6.

Neal, Peter et Gareth Roberts (2011). "Optimal Scaling of Random Walk Metropolis Algorithms with Non–Gaussian Proposals". English. In : *Methodology and Computing in Applied Probability* 13.3, p. 583–601. issn : 1387-5841. doi : 10.1007/s11009-010-9176-9. url : http://dx.doi.org/10.1007/s11009-010-9176-9.

Neal, R. (2001). "Annealed importance sampling". In : *Statistics and Computing* 11, p. 125–139.

Neal, R.M. (2012). "MCMC using Hamiltonian dynamics". In : *arXiv preprint arXiv:1206.1901*.

Peskun, P. H. (1973a). "Optimum Monte-Carlo sampling using Markov chains". In : *Biometrika* 60, p. 607–612.

Ripley, Brian D. (1987). *Stochastic simulation*. New York : John Wiley & Sons Inc., p. xiv+237. isbn : 0-471-81884-4.

Robert, C.P. (2007). *: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York : Springer.

Robert, C.P. et G. Casella (2004). *Monte Carlo Statistical Methods*. second. Springer-Verlag.

Roberts, G. O., A. Gelman et W. R. Gilks (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". In : *Ann. Appl. Probab.* 7.1, p. 110–120. issn : 1050-5164.

Roberts, Gareth O. et Jeffrey S. Rosenthal (1997). "Geometric ergodicity and hybrid Markov chains". In : *Electron. Comm. Probab.* 2, no. 2, 13–25 (electronic). issn : 1083-589X.

— (2001). "Markov Chains and de-initializing processes". In : *Scandinavian Journal of Statistics* 28, p. 489–504.

Roberts, G.O. et J.S. Rosenthal (2009). "Examples of Adaptive MCMC". In : *J. Comp. Graph. Stat.* 18, p. 349–367.

Rosenthal, J. S. et G. O. Roberts (2007). "Coupling and Ergodicity of adaptive MCMC". In : *Journal of Applied Probablity* 44, p. 458–475.

Sherlock, Chris et Gareth Roberts (2009). "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". In : *Bernoulli* 15.3, p. 774–798. doi : 10.3150/08-BEJ176. url : http://dx.doi.org/10.3150/08-BEJ176.

Tierney, L. (1994). "Markov chains for exploring posterior distributions (with discussion)". In : 22, p. 1701–1786.

Tierney, Luke (1998). "A note on Metropolis-Hastings kernels for general state spaces". In : *Ann. Appl. Probab.* 8.1, p. 1–9. issn : 1050-5164.

Zhou, Y., A. M Johansen et J. A. Aston (2013). "Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach". In : *ArXiv e-prints*. arXiv : 1303.3123 [stat.ME].

Chapitre 2

Delayed Acceptance

Accelerating Metropolis–Hastings algorithms by Delayed Acceptance

This is joint work with Clara Grazian, Anthony Lee and Christian P.Robert.

2.1 Introduction

When running an MCMC sampler such as Metropolis–Hastings algorithms (Robert et Casella, 2004), the complexity of the target density required by the acceptance ratio may lead to severe slow-downs in the execution of the algorithm. A direct illustration of this difficulty is the simulation from a posterior distribution involving a large dataset of n points for which the computing time is at least of order $\mathrm{O}(n)$. Several solutions to this issue have been proposed in the recent literature (Korattikara, Chen et Welling, 2013 ; Neiswanger, Wang et Xing, 2013 ; Scott et al., 2013 ; Wang et Dunson, 2013), taking advantage of the likelihood decomposition

$$\prod_{i=1}^{n} \ell(\theta|x_i) \tag{2.1}$$

to handle subsets of the data on different processors (CPU), graphical units (GPU), or even computers. However, there is no consensus on the method of choice, some leading to instabilities by removing most prior inputs and others to approximations delicate to evaluate or even to implement.

Our approach here is to delay acceptance (rather than rejection as in Tierney et Mira, 1998) by sequentially comparing parts of the MCMC acceptance ratio to independent uniforms, in order to stop earlier the computation of the aforesaid ratio, namely as soon as one term is below the corresponding uniform.

More formally, consider a generic Metropolis–Hastings algorithm where the acceptance ratio $\pi(y)q(y,x)/\pi(x)q(x,y)$ is compared with a $\mathcal{U}(0,1)$ variate to decide whether or not the Markov chain switches from the current value x to the proposed value y (Robert et Casella, 2004). If we now decompose the ratio as an arbitrary

product

$$\pi(y)\, q(y,x) \big/ \pi(x) q(x,y) = \prod_{k=1}^{d} \rho_k(x,y) \tag{2.2}$$

where the only constraint is that the functions ρ_k are all positive and satisfy the balance condition $\rho_k(x,y) = \rho_k(y,x)^{-1}$ and then accept the move with probability

$$\prod_{k=1}^{d} \min\left\{\rho_k(x,y), 1\right\} , \tag{2.3}$$

i.e. by successively comparing uniform variates u_k to the terms $\rho_k(x,y)$, the motivation for our delayed approach is that the same target density $\pi(\cdot)$ is stationary for the resulting Markov chain.

The mathematical validation of this simple if surprising result can be seen as a consequence of Christen et Fox, 2005. This paper reexamines Fox et Nicholls, 1997, where the idea of testing for acceptance using an approximation before computing the exact likelihood was first suggested. In Christen et Fox, 2005, the original proposal density q is used to generate a value y that is tested against an approximate target $\tilde{\pi}$. If accepted, y can be seen as coming from a pseudo-proposal $\tilde{q}$ that simply is formalising the earlier preliminary step and is then tested against the true target π. The validation in Christen et Fox, 2005 follows from standard detailed balance arguments ; we will focus formally on this point in Section 2.2.

In practice, sequentially comparing those probabilities with d uniform variates means that the comparisons stop at the first rejection, implying a gain in computing time if the most costly items in the product (2.2) are saved for the final comparisons.

Examples of the specific two-stage Delayed Acceptance as defined by Christen et Fox, 2005 can be found in Golightly, Henderson et Sherlock, 2014, in the pMCMC context, and in Shestopaloff et Neal, 2013.

The major drawback of the scheme is that Delayed Acceptance *efficiently* reduces the computing cost only when the approximation $\tilde{\pi}$ is "good enough" or "flat enough", since the probability of acceptance of a proposed value will always be smaller than in the original Metropolis–Hastings scheme. In other words, the original Metropolis–Hastings kernel dominates the new one in Peskun's (Peskun, 1973b) sense. The most relevant question raised by Christen et Fox, 2005 is thus how to achieve a proper approximation ; note in fact that while in Bayesian statistics a decomposition of the target is always available, by breaking original data in subsamples and considering the corresponding likelihood parts or even by just separating the prior, proposal and likelihood ratio into different factors, these decompositions may just lead to a deterioration of the algorithm properties without impacting the computational efficiency.

However, even in these simple cases, it is possible to find examples where Delayed Acceptance may be profitable. Consider for instance resorting to a costly non-informative prior distribution (as illustrated in Section 2.5.3 in the case of mixtures) ; here the first acceptance step can be solely based on the ratio of the likelihoods and the second step, which involves the ratio of the priors, does not require to be computed when the first test leads to rejection. Even more often, the converse decomposition applies to complex or just costly likelihood functions, in that the prior ratio may first be used to eliminate values of the parameter that are too unlikely for the prior density. As shown in Figure 2.1, a standard normal-normal example

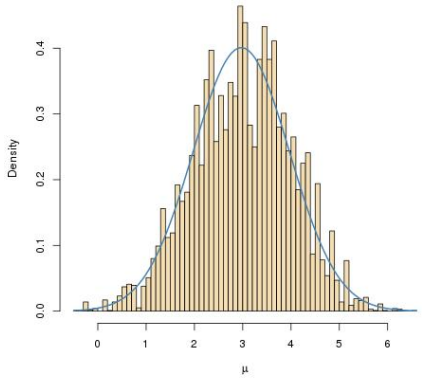


FIGURE 2.1: Fit of a two-step Metropolis–Hastings algorithm applied to a normal-normal posterior distribution $\mu|x \sim N(x/(\{1+\sigma_\mu^{-2}\}, 1/\{1+\sigma_\mu^{-2}\})$ when $x = 3$ and $\sigma_\mu = 10$, based on $T = 10^5$ iterations and a first acceptance step considering the likelihood ratio and a second acceptance step considering the prior ratio, resulting in an overall acceptance rate of 12%

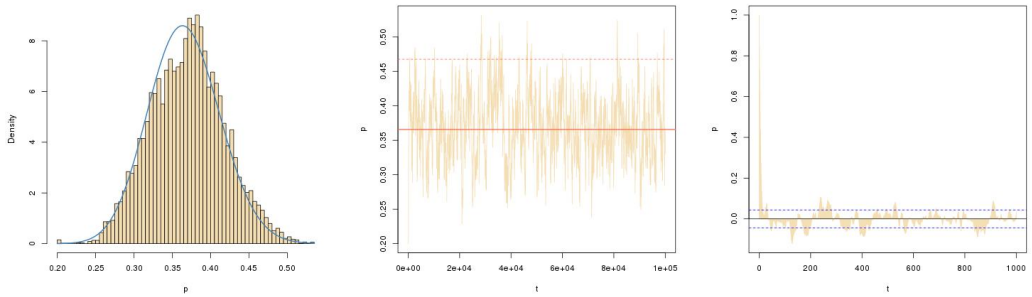


FIGURE 2.2: (left) Fit of a multiple-step Metropolis–Hastings algorithm applied to a Beta-binomial posterior distribution $p|x \sim Be(x+a, n+b-x)$ when $N = 100$, $x = 32$, $a = 7.5$ and $b = .5$. The binomial $\mathcal{B}(N, p)$ likelihood is replaced with a product of 100 Bernoulli terms and an acceptance step is considered for the ratio of each term. The histogram is based on 10^5 iterations, with an overall acceptance rate of 9% ; (centre) raw sequence of successive values of p in the Markov chain simulated in the above experiment ; (right) autocorrelogram of the above sequence.

confirms that the true posterior and the histogram resulting from such a simulated sample are in agreement.

In more complex settings, as for example in "Big Data" settings where the likelihood is made of a very large number of terms, the above principle also applies to any factorisation of the like of (2.1) so that each individual likelihood factor can be evaluated separately. This approach increases both the variability of the evaluation and the potential for rejection, but, if each term of the factored likelihood is sufficiently costly to compute, the decomposition brings some improvement in execution time. The graphs in Figure 2.2 illustrate an implementation of this perspective in the Beta-binomial case, namely when the binomial $\mathcal{B}(N, p)$ observation $x = 32$ is replaced with a sequence of N Bernoulli observations. The fit is adequate on 10^5 iterations, but the autocorrelation in the sequence is very high (note that the ACF is for the 100 times thinned sequence) while the acceptance rate falls down to 9%. (When the original $y = 32$ observation is (artificially) divided into 10, 20, 50, and 100 parts, the acceptance rates are 0.29, 0.25, 0.12, and 0.09, respectively.) The gain in using this decomposition is only appearing when each Bernoulli likelihood computation becomes expensive enough.

On one hand, the order in which the product (2.3) is explored determines the computational efficiency of the scheme, while, on the other hand, it has no impact on the overall convergence of the resulting Markov chain, since the acceptance of a proposal does require computing all likelihood values. It therefore makes sense to try to optimise this order by ranking the entries in a way that improves the execution speed of the algorithm (see Section 2.3.2).

We also stress that the Delayed Acceptance principle remains valid even when the likelihood function or the prior are not integrable over the parameter space. Therefore, the prior may well be improper. For instance, when the prior distribution is constant, a two-stage acceptance scheme reverts to the original Metropolis–Hastings one.

Finally, while the Delayed Acceptance methodology is intended to cater to complex likelihoods or priors, it does not bring a solution *per se* to the "Big Data" problem in that (a) all terms in the product must eventually be computed ; and (b) all terms previously computed (i.e., those computed for the last accepted value of the parameter) must be either stored for future comparison or recomputed. See, e.g., Scott et al., 2013 ; Wang et Dunson, 2013, for recent entries on different parallel ways of handling massive datasets.

The plan of the paper is as follows : in Section 2.2, we validate the decomposition of the acceptance step into a sequence of decisions, arguing about the computational gains brought by this generic modification of Metropolis–Hastings algorithms and further analysing the relation between the proposed method and the Metropolis–Hastings algorithm in terms of convergence properties and asymptotic variances of statistical estimates. In Section 2.4 we briefly state the relations between Delayed Acceptance and other methods present in the literature. In Section 2.3 we aim at giving some intuitions on how to improve the behaviour of Delayed Acceptance by ranking the factors in a given decomposition to achieve optimal computational efficiency and finally give some preliminary results in terms of optimal scaling for the proposed method. Then Section 2.5 studies Delayed Acceptance within three realistic environments, the first one made of logistic regression targets, the second one alleviating the computational burden from a Geometric Metropolis adjusted

Langevin algorithm and a third one handling an original analysis of a parametric mixture model via genuine Jeffreys priors. Section 2.6 concludes the paper.

2.2 Validation and convergence of Delayed Acceptance

In this section, we establish that Delayed Acceptance is a valid Markov chain Monte Carlo scheme and analyse on a theoretical basis the differences with the original version.

2.2.1 The general scheme

We assume for simplicity that the target distribution π and the proposal distributions $Q(x,\cdot)$ all admit densities w.r.t. the Lebesgue or counting measures. We also denote by π the target density and let $q(x,y)$ denote the proposal density.

Let $(X_n)_{n\geq 1}$ be a Markov chain evolving on X with Metropolis–Hastings Markov transition kernel P associated with q and π, i.e. for $A \in \mathcal{B}(\mathsf{X})$

$$P(x,A) := \int_A q(x,y)\alpha(x,y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x,y)\alpha(x,y)\mathrm{d}y\right)\mathbf{1}_A(x),$$

where

$$\alpha(x,y) := 1 \wedge r(x,y), \qquad r(x,y) := \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}.$$

Above, $\alpha(x,y)$ is known as the Metropolis–Hastings acceptance probability and $r(x,y)$ as the Metropolis–Hastings acceptance ratio.

We consider the class of "Delayed acceptance" Markov kernels associated with P, which are defined by factorisations of the function r as

$$r(x,y) = \prod_{k=1}^{d} \rho_k(x,y) \tag{2.4}$$

with all components in the product satisfying $\rho_k(x,y) = \rho_k(y,x)^{-1}$. The associated Delayed Acceptance Markov kernel is then defined as

$$\tilde{P}(x,A) := \int_A q(x,y)\tilde{\alpha}(x,y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x,y)\tilde{\alpha}(x,y)\mathrm{d}y\right)\mathbf{1}_A(x),$$

where

$$\tilde{\alpha}(x,y) := \prod_{k=1}^{d}\{1 \wedge \rho_k(x,y)\}.$$

We will denote by $(\tilde{X}_n)_{n\geq 1}$ the Markov chain associated with $\tilde{P}$.

The order in which the sequence of functions ρ_k appears in the factorisation (2.4) is important for algorithmic specification, as can be seen in Algorithm 2.1. It means that $\rho_1(x,Y)$ is evaluated first, $\rho_2(x,Y)$ second, and so on until $\rho_d(x,Y) = r(x,Y)/\prod_{k=1}^{d-1}\rho_k(x,Y)$ which is last, with the motivation that "early rejection" can allow computational savings by avoiding the computation of the subsequent $\rho_k(x,Y)$.

Algorithm 2.1 Delayed Acceptance

To sample from $\tilde{P}(x,\cdot)$:

1. Sample $Y \sim Q(x,\cdot)$.
2. For $k = 1,\ldots,d$:
   — With probability $1 \wedge \rho_k(x,Y)$ continue, otherwise stop and output x.
3. Output Y.

2.2.2 Validation

The first lemma is a standard representation leading to the validation of the Delayed Acceptance Markov chain :

Lemma 4. *For any Markov chain with transition kernel* Π *of the form*

$$\Pi(x,A) = \int_A q(x,y)a(x,y)\mathrm{d}y + \left(1 - \int_{\mathsf{X}} q(x,y)a(x,y)\mathrm{d}y\right)\mathbf{1}_A(x),$$

and satisfying detailed balance, the function $a(\cdot)$ *satisfies (for* π*-a.a.* x,y*)*

$$\frac{a(x,y)}{a(y,x)} = r(x,y).$$

Démonstration. This follows immediately from the detailed balance condition

$$\pi(x)q(x,y)a(x,y) = \pi(y)q(y,x)a(y,x).$$

□

The Delayed Acceptance Markov chain $(\tilde{X}_n)_{n\geq 1}$ is then associated with the intended target :

Lemma 5. $(\tilde{X}_n)_{n\geq 1}$ *is a* π*-reversible Markov chain.*

Démonstration. From Lemma 4 it suffices to verify that $\tilde{\alpha}(x,y)/\tilde{\alpha}(y,x) = r(x,y)$. Indeed, we have

$$\begin{aligned}\frac{\tilde{\alpha}(x,y)}{\tilde{\alpha}(y,x)} &= \frac{\prod_{k=1}^d\{1\wedge\rho_k(x,y)\}}{\prod_{k=1}^d\{1\wedge\rho_k(y,x)\}}\\ &= \prod_{k=1}^d \frac{1\wedge\rho_k(x,y)}{1\wedge\rho_k(y,x)}\\ &= \prod_{k=1}^d \rho_k(x,y) = r(x,y),\end{aligned}$$

since $\rho_k(y,x) = \rho_k(x,y)^{-1}$ and $(1\wedge a)/(1\wedge a^{-1}) = a$ for $a \in \mathbb{R}_+$. □

Remark 2. It is immediate to show that

$$\tilde{\alpha}(x,y) = \prod_{k=1}^d\{1\wedge\rho_k(x,y)\} \leq 1\wedge\prod_{k=1}^d\rho_k(x,y) = 1\wedge r(x,y) = \alpha(x,y),$$

since $(1\wedge a)(1\wedge b) \leq (1\wedge ab)$ for $a,b\in\mathbb{R}_+$.

2.2.3 Comparisons of the kernels P and $\tilde{P}$

Given a probability measure μ, let us denote

$$\mu(f) := \int_{\mathsf{E}} f(x)\mu(\mathrm{d}x)\,, \quad L^2(\mathsf{E},\mu) := \{f : \mu(f^2) < \infty\}$$
$$L_0^2\big(\mathsf{E},\mu\big) := \big\{f \in L^2(\mathsf{E},\mu) : \mu\big(f\big) = 0\big\}\,.$$

For a generic Markov kernel $\Pi : \mathsf{E} \times \mathcal{B}(\mathsf{E})$ with unique invariant probability measure μ, we define

$$\mathrm{var}(f,\Pi) := \lim_{n\to\infty} \mathrm{var}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[f(X_i) - \mu(f)\right]\right),$$

where $(X_n)_{n\geq 1}$ is a Markov chain with Markov kernel Π initialised with $X_1 \sim \mu$.

Remark 3. One can immediately conclude from the construction of $\tilde{P}$ that $\mathrm{var}(f,P) \leq \mathrm{var}(f,\tilde{P})$ for any $f \in L^2(\mathsf{X},\pi)$, using Peskun ordering (Peskun, 1973a; Tierney, 1998), since $\tilde{\alpha}(x,y) \leq \alpha(x,y)$ for any $(x,y) \in \mathsf{X}^2$.

For any $f \in L^2(\mathsf{E},\mu)$ we define the Dirichlet form associated with a μ-reversible Markov kernel $\Pi : \mathsf{E} \times \mathcal{B}(\mathsf{E})$ as

$$\mathcal{E}_\Pi(f) := \frac{1}{2}\int \mu(\mathrm{d}x)\Pi(x,\mathrm{d}y)\left[f(x) - f(y)\right]^2.$$

The (right) spectral gap of a generic μ-reversible Markov kernel has the following variational representation

$$\mathrm{Gap}\,(\Pi) := \inf_{f\in L_0^2(\mathsf{E},\mu)} \frac{\mathcal{E}_\Pi(f)}{\langle f, f\rangle_\mu} \quad .$$

which leads to the following comparison lemma :

Lemma 6 ((Andrieu, Lee et Vihola, 2013, Lemma 34))**.** *Let Π_1 and Π_2 be μ-reversible Markov transition kernels of μ-irreducible and aperiodic Markov chains, and assume that there exists $\varrho > 0$ such that for any $f \in L_0^2\big(\mathsf{E},\mu\big)$*

$$\mathcal{E}_{\Pi_2}\big(f\big) \geq \varrho\mathcal{E}_{\Pi_1}\big(f\big) \quad ,$$

then

$$\mathrm{Gap}\big(\Pi_2\big) \geq \varrho\mathrm{Gap}\big(\Pi_1\big)$$

and

$$\mathrm{var}\big(f,\Pi_2\big) \leq (\varrho^{-1}-1)\mathrm{var}_\mu(f) + \varrho^{-1}\mathrm{var}\,(f,\Pi_1) \quad f \in L_0^2(\mathsf{E},\mu).$$

We will need the following condition in the sequel, which imposes a uniform lower bound on each $\rho_k(x,y)$ when $\alpha(x,y) = 1$:

Condition 1. Defining $A := \{(x,y) \in \mathsf{X}^2 : r(x,y) \geq 1\}$, there exists a c such that $\inf_{(x,y)\in A} \min_{k\in\{1,\ldots,d\}} \rho_k(x,y) \geq c$.

Intuitively, this condition ensures that when the acceptance probability $\alpha(x,y)$ is 1 then the acceptance probability $\tilde{\alpha}(x,y)$ is uniformly lower bounded by a constant. Reversibility then implies that $\tilde{\alpha}(x,y)$ is uniformly lower bounded by a constant multiple of $\alpha(x,y)$ for all $x,y \in \mathsf{X}$. Ultimately, this allows one to show, using Lemma 6, that $\tilde{P}$ is not too different from P.

Proposition 1. *Assume Condition 1. Then Lemma 6 holds with* $\Pi_1 = P$, $\Pi_2 = \tilde{P}$, $\mu = \pi$ *and* $\varrho = c^{d-1}$.

Démonstration. Let $(x, y) \in A$, defined in Condition 1. Since $r(x, y) \geq 1$, we have $\alpha(x, y) = 1$. On the other hand, from Condition 1,

$$\begin{aligned}\tilde{\alpha}(x, y) &= \prod_{k=1}^{d} 1 \wedge \rho_k(x, y) \\ &= \prod_{k:\rho_k(x,y)<1} \rho_k(x, y) \geq c^{|\{k:\rho_k(x,y)<1\}|} \geq c^{d-1},\end{aligned}$$

since at least one $\rho_k(x, y) \geq 1$ whenever $r(x, y) \geq 1$.

From Lemma 4, when $(x, y) \in A$, we have

$$\tilde{\alpha}(y, x) = \tilde{\alpha}(x, y)/r(x, y) \geq c^{d-1}\alpha(x, y)/r(x, y) = c^{d-1}\alpha(y, x)$$

and thus $\tilde{\alpha}(x, y) \geq c^{d-1}\alpha(x, y)$ for any $(x, y) \in \mathsf{X}^2$. It follows that

$$\begin{aligned}\mathcal{E}_{\tilde{P}}(f) &= \int_{\mathsf{X}} \pi(\mathrm{d}x)\tilde{P}(x, \mathrm{d}y)\,(f(x) - f(y))^2 \\ &= \int_{\mathsf{X}} \pi(\mathrm{d}x)P(x, \mathrm{d}y)\frac{\tilde{\alpha}(x, y)}{\alpha(x, y)}\,(f(x) - f(y))^2 \\ &\geq c^{d-1}\int_{\mathsf{X}} \pi(\mathrm{d}x)P(x, \mathrm{d}y)\,(f(x) - f(y))^2 = c^{d-1}\mathcal{E}_P(f),\end{aligned}$$

and we conclude. □

The implication of this result is that, if P admits a right spectral gap, then so does $\tilde{P}$, whenever Condition 1 holds. Furthermore, and irrespective of whether or not P admits a right spectral gap, quantitative bounds on the asymptotic variance of MCMC estimates using $(\tilde{X}_n)_{n\geq 1}$ in relation to those using $(X_n)_{n\geq 1}$ are available.

2.2.4 Modification of a given factorisation

The easiest way to use the above result is to modify any candidate factorisation. Given a factorisation of the function r

$$r(x, y) = \prod_{k=1}^{d} \tilde{\rho}_k(x, y)\,,$$

satisfying the balance condition, we can define a sequence of functions ρ_k such that both $r(x, y) = \prod_{k=1}^{d} \rho_k(x, y)$ and Condition 1 holds. To that effect, take an arbitrary $c \in (0, 1]$ and define $b := c^{\frac{1}{d-1}}$. Then, if we set

$$\rho_k(x, y) := \min\left\{\frac{1}{b}, \max\left\{b, \tilde{\rho}_k(x, y)\right\}\right\}, \quad k \in \{1, \ldots, d-1\},$$

it then follows that one must define

$$\rho_d(x, y) := \frac{r(x, y)}{\prod_{k=1}^{d-1} \rho_k(x, y)}.$$

From this modification, we deduce the following result :

Proposition 2. *Using this scheme, Lemma 6 holds with* $\Pi_1 = P$, $\Pi_2 = \tilde{P}$, $\mu = \pi$ *and* $\varrho = c^2$.

Démonstration. We note that $\inf_{(x,y)\in \mathsf{X}^2} \prod_{k=1}^{d-1} 1 \wedge \rho_k(x,y) \geq b^{d-1} = c$ and that

$$\tilde{\rho}_d(x,y) = \frac{r(x,y)}{\prod_{k=1}^{d-1} \rho_k(x,y)} \geq b^{d-1} r(x,y) = c r(x,y).$$

With $A := \{(x,y) \in \mathsf{X}^2 : r(x,y) \geq 1\}$, it follows that $\inf_{(x,y)\in A} \tilde{\rho}_d(x,y) \geq c$, and so $\inf_{(x,y)\in A} \tilde{\alpha}(x,y) \geq c^2$. We conclude along the same lines as in the proof of Proposition 1. □

While this modification ensures that one can take $\varrho = c^2$ in Proposition 1, it is too general to suggest that using $\tilde{P}$ can be more computationally efficient than using P when the cost of evaluating each ρ_k is taken into account. Indeed, Proposition 2 holds when the functions $\tilde{\rho}_k$ are chosen completely arbitrarily. Of course in practice, one should choose $\tilde{\rho}_k$ and hence ρ_k so that they are in some sense in agreement with r.

We will show in the next example that a certain class of $\tilde{\rho}_k$'s are beneficial, namely those which correspond to Metropolis–Hastings acceptance ratios with "flattened" surrogate target densities. On the other hand, it is far from difficult to come up with surrogate target densities for which unmodified use of $\tilde{\rho}_k$ can lead to disastrous performance.

2.2.5 Example : unmodified surrogate targets

One common usage (Christen et Fox, 2005) of Delayed Acceptance is to substitute a surrogate target $\bar{\pi}$ for π in $\rho_1(x,y)$. We consider the case $d = 2$ and a random walk proposal to examine Condition 1 in this context. Here we have $q(x,y) = q(y,x)$ and so

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)},$$

while

$$\rho_1(x,y) = 1 \wedge \frac{\bar{\pi}(y)}{\bar{\pi}(x)}, \quad \rho_2(x,y) = 1 \wedge \frac{\pi(y)\bar{\pi}(x)}{\pi(x)\bar{\pi}(y)}.$$

Considering $(x,y) \in A = \{(x,y) \in \mathsf{X}^2 : r(x,y) \geq 1\}$ we require $c > 0$ satisfying simultaneously

$$\frac{\bar{\pi}(y)}{\bar{\pi}(x)} \geq c, \qquad \frac{\pi(y)\bar{\pi}(x)}{\pi(x)\bar{\pi}(y)} \geq c.$$

The first of these says that $\bar{\pi}(y)/\bar{\pi}(x)$ cannot be too small when $\pi(y) \geq \pi(x)$. The second says that $\bar{\pi}(y)/\bar{\pi}(x)$ should not be a large multiple of $\pi(y)/\pi(x)$. There are a large variety of choices of $\bar{\pi}$ that allow one to take $c = 1$. For example, $\bar{\pi}(x) = \pi(x) + C$ for some constant $C \geq 0$ and $\bar{\pi}(x) \propto \pi(x)^\beta$ for some $\beta \in [0,1]$. Note that $\beta = 0$ corresponds to $\bar{\pi}$ being a constant function and $\beta = 1$ corresponds to $\bar{\pi} \propto \pi$. In between, one can think of $\bar{\pi}$ as being a flattened version of π.

2.2.6 Counter-example : failure to reproduce geometric ergodicity

Consider the case $\pi(x) = \mathcal{N}(x; 0, 1)$ and $\bar{\pi}(x) = \mathcal{N}(x; 0, \sigma^2)$ with $Q(x, \cdot)$ a normal distribution with mean x and fixed variance for each $x \in \mathbb{R}$. Here we have

$$\rho_1(x, y) = \exp\left\{\frac{(x-y)(x+y)}{2\sigma^2}\right\}, \quad \rho_2(x, y) = \exp\left\{\frac{(1-\sigma^2)(y-x)(y+x)}{2\sigma^2}\right\}.$$

Mengersen et Tweedie (1996) showed that a random-walk Metropolis–Hastings chain for targets with super-exponential tails is geometrically ergodic. We now exploit this result to derive that, if $\sigma^2 < 1$, then the unmodified delayed acceptance Markov chain is not geometrically ergodic.

Proposition 3. *The unmodified Delayed Acceptance Markov chain using the factorisation into ρ_1 and ρ_2 as above is not geometrically ergodic when $\sigma < 1$.*

Intuitively, when x is large $P(x, (-\infty, x)) \approx \frac{1}{2}$ but $\lim_{x\to\infty} \tilde{P}(x, \{x\}) = 1$ because $\rho_1(x, y)$ takes on smaller and smaller values for $y > x$ and $\rho_2(x, y)$ takes on smaller and smaller values for $y < x$.

Démonstration. From Roberts et Tweedie (1996, Theorem 5.1), it suffices to show that $\pi-\text{ess}\inf_{x\in\mathsf{X}} \tilde{P}(x, \{x\}^{\complement}) = 0$, i.e. that for any $\tau \in (0, 1)$ we can find $A \subseteq \mathsf{X}$ such that $\pi(A) > 0$ and $\sup_{x\in A} \tilde{P}(x, \{x\}^{\complement}) \leq \tau$. Let $B_s(z)$ denote the ball of radius s around z. Given $\tau \in (0, 1)$, we define

$$r := \sup\{s > 0 : Q(x, B_s(x)) < \tau/3\},$$

and

$$A := \left\{x : x > \frac{r}{2} - \frac{\sigma^2 \log(\tau/3)}{r(1-\sigma^2)}\right\} \bigcap \left\{x : Q(x, \mathbb{R}_-) < \frac{\tau}{3}\right\}.$$

Then

$$\begin{aligned}
\tilde{P}(x, \{x\}^{\complement}) &= \tilde{P}(x, B_r(x) \setminus \{x\}) + \tilde{P}(x, B_r^{\complement}(x)) \\
&\leq \frac{\tau}{3} + \int_{B_r^{\complement}(x)} Q(x, \mathrm{d}y)\tilde{\alpha}(x, y) \\
&\leq \frac{2\tau}{3} + \int_{B_r^{\complement}(x)\cap\mathbb{R}_+} Q(x, \mathrm{d}y)\tilde{\alpha}(x, y) \\
&\leq \frac{2\tau}{3} + \sup_{y\in B_r^{\complement}(x)\cap\mathbb{R}_+} \tilde{\alpha}(x, y) \\
&= \frac{2\tau}{3} + \sup_{y\in B_r^{\complement}(x)\cap\mathbb{R}_+} [1 \wedge \rho_1(x, y)]\,[1 \wedge \rho_2(x, y)].
\end{aligned}$$

Now let $x \in A$, $y \in B_r^{\complement}(x) \cap \mathbb{R}_+$ and assume $y < x$. It follows that $\rho_2(x, y)$ attains its maximum when $y = x - r$ and therefore

$$\begin{aligned}
\rho_2(x, y) &\leq \exp\left\{\frac{(1-\sigma^2)r(r-2x)}{2\sigma^2}\right\} \\
&\leq \exp\left\{\frac{(1-\sigma^2)r}{2\sigma^2}\left[\frac{2\sigma^2\log(\tau/3)}{r(1-\sigma^2)}\right]\right\} = \frac{\tau}{3}.
\end{aligned}$$

Similarly, let $x \in A$, $y \in B_r^{\complement}(x) \cap \mathbb{R}_+$ and assume $y > x$. It follows that $\rho_1(x, y)$ attains its maximum when $y = x + r$ and therefore

$$\begin{aligned}
\rho_1(x,y) &\leq \exp\left\{-\frac{r(2x+r)}{2\sigma^2}\right\} \\
&\leq \exp\left\{-\frac{r}{2\sigma^2}\left(2r - \frac{2\sigma^2\log(\tau/3)}{r(1-\sigma^2)}\right)\right\} \\
&\leq \exp\left\{\frac{r}{2\sigma^2}\left(\frac{2\sigma^2\log(\tau/3)}{r(1-\sigma^2)}\right)\right\} \leq \frac{\tau}{3},
\end{aligned}$$

since $\log(\tau/3) < 0$ and $\sigma^2 < 1$. Therefore,

$$\sup_{y \in B_r^{\complement}(x) \cap \mathbb{R}_+} [1 \wedge \rho_1(x,y)]\,[1 \wedge \rho_2(x,y)] \leq \frac{\tau}{3}$$

so $\tilde{P}(x, \{x\}^{\complement}) \leq \tau$ and we conclude. □

The same argument can be made for much more general targets and proposals, albeit at the expense of brevity and clarity. We refrain from such a generalisation as our purpose here is to demonstrate that the DA chain may fail to inherit geometric ergodicity and that the simple proposed modification of the Delayed Acceptance kernel provided in Section 2.2.4 allows one to avoid this.

2.3 Optimisation

When considering Markov Chain Monte Carlo methods in practice, their efficiency as measured by mixing properties and computational cost is a fundamental issue. This section addresses both perspectives in connection with Delayed Acceptance. Section 2.3.1 examines the proposal distribution and derives its optimal scaling from standard random–walk Metropolis–Hastings theory. Then Section 2.3.2 covers the ranking of the factors ρ_i, which drives the total computational cost of the procedure.

2.3.1 Optimising the proposal mechanism

The explorative performances of a random–walk MCMC are strongly dependent on its proposal distribution. As exemplified in Roberts, Gelman et Gilks, 1997, finding the optimal scale parameter does lead to efficient 'jumps' in the state space and the overall acceptance rate of the chain is directly connected to the average jump distance and to the asymptotic variance of ergodic averages. This provides practitioners with an approach to 'auto-tune' the resulting random–walk MCMC algorithm. Extending this calibration to the Delayed Acceptance scheme is equally important, on its own ground towards finding a reasonable scaling for the proposal distribution and to avoid comparisons with the standard Metropolis–Hastings version.

The original framework of Roberts, Gelman et Gilks, 1997 is centered on estimating a collection of expected functionals, say g, where a plausible criterion

for the performances of the MCMC is the minimisation of the stationary integrated auto-correlation time (ACT) of the Markov chain, defined as

$$\tau_g = 1 + 2\sum_{i=1}^{\infty} \mathbb{C}\text{or}(g(X_0), g(X_i))$$

where the index g stresses the dependence on the considered functional, which is connected to the asymptotic variance through $\text{var}(P, g) = \tau_g \times \text{var}_\pi(g)$ whenever the chain is ϕ-irreducible, aperiodic, and reversible, $\text{var}_\pi(g)$ is finite and $g \in L^2(\pi)$.

Research on this optimisation focus on two main cases :

— Consider the limit in the dimension of the target distribution toward ∞, where Roberts, Gelman et Gilks, 1997 gave conditions under which each marginal chain converges toward a Langevin diffusion. Maximising the speed of that diffusion, say $h(\ell)$ where ℓ is a parameter of the scale of the proposal, implies a minimisation of the ACT and also that τ is *free* from the dependence on the functional, defining thus an independent measure of efficiency for the algorithm ;

— Sherlock et Roberts, 2009 focus on unimodal elliptically symmetric targets and show that a proxy for the ACT in finite dimensions is the Expected Square Jumping Distance (ESJD), defined as

$$\mathbb{E}\left[\|X' - X\|_\beta^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} \frac{1}{\beta_i^2}(X_i' - X)^2\right]$$

where X and X' are two successive points in the chain and $\|\cdot\|_\beta$ represent the norm on the principal axes of the ellipse rescaled by the coefficients β_i so that every direction contributes equally.

An interesting result in Sherlock et Roberts, 2009 is that, as $d \to \infty$, the ESJD on one marginal component of the chain converges with the same speed as the diffusion process described in Roberts, Gelman et Gilks, 1997. These authors furthermore show the asymptotic result holds for rather small dimension, roughly starting from $d = 5$.

Moreover, when considering efficiency for Delayed Acceptance, which is a technique tailored on costly computations, we need to focus on the execution time of the algorithm as well. We then proceed to define our measure of efficiency as

$$\textbf{Eff} := \text{ESJD}/\text{cost per iteration} \tag{2.5}$$

similarly to Sherlock et al., 2013 for Pseudo-Marginal MCMC.

Due to the complex acceptance ratio in Delayed Acceptance, an extension of the previous results requires rather stringent assumptions, albeit providing a proper guideline in practice. Section 2.5 will further demonstrate optimality extends beyond those conditions. Note that our assumptions are quite standard in the literature on the subject.

(H1) We assume for simplicity's sake that the Delayed Acceptance procedure operates on two factors only, i.e., that $r(x, y) = \rho_1(x, y) \times \rho_2(x, y)$. The acceptance probability of the scheme is thus

$$\tilde{\alpha}(x, y) = \prod_{i=1}^{2}(1 \wedge \rho_i(x, y)).$$

We also consider the ideal setting where a computationally cheap approximation $\tilde{f}(\cdot)$ is available for $\pi(\cdot)$ and precise enough so that $\rho_2(x,y) = r(x,y)/\rho_1(x,y) = \pi(y)/\pi(x) \times \tilde{f}(x)/\tilde{f}(y) = 1$.

(H2) We further assume that the target distribution satisfies (A1) and (A2) in Roberts, Gelman et Gilks, 1997, which are regularity conditions on π and its first and second derivatives, and that $\pi(x) = \prod_{i=1}^{n} f(x_i)$.
(H3) We consider a random walk proposal $y = x + \sqrt{\ell^2/d}\, Z$, where $Z \sim \mathcal{N}(0, I_d)$. Note that Gaussianity can be easily relaxed to distributions with finite fourth moment and similar results are available for more heavy-tailed distributions (Neal et Roberts, 2011).

(H4) Finally we assume that the cost of computing $\tilde{f}(\cdot)$, say c, is proportional to the cost of computing $\pi(\cdot)$, named C, with $c = \delta C$.

Normalising costs by setting $C = 1$, the average total cost per iteration of the Delayed Acceptance chain is $\delta + \mathbb{E}[\tilde{\alpha}]$ and the efficiency of the proposed method under the above conditions is

$$\mathbf{Eff}(\delta, \ell) = \frac{ESJD}{\delta + \mathbb{E}[\tilde{\alpha}]}$$

Lemma 7. *Under the above conditions (H1–H4) on the target $\pi(x)$, on the proposal $q(x,y)$ and on the factorised acceptance probability $\tilde{\alpha}(x,y) = \prod_{i=1}^{2}(1 \wedge \rho_i(x,y))$ we have that*

$$\tilde{\alpha}(x,y) = (1 \wedge \rho_1(x,y))$$

and that as $d \to \infty$

$$\mathbf{Eff}(\delta, \ell) = \frac{h(\ell)}{\delta + \mathbb{E}[\tilde{\alpha}]} = \frac{2\ell^2 \Phi(-\frac{\ell\sqrt{I}}{2})}{\delta + 2\Phi(-\frac{\ell\sqrt{I}}{2})}$$

$$a(\ell) = \mathbb{E}[\tilde{\alpha}] = 2\Phi(-\frac{\ell\sqrt{I}}{2})$$

where $I := \mathbb{E}\left[\left(\frac{(\pi(x))'}{\pi(x)}\right)^2\right]$ as defined in Roberts, Gelman *et* Gilks, *1997.*

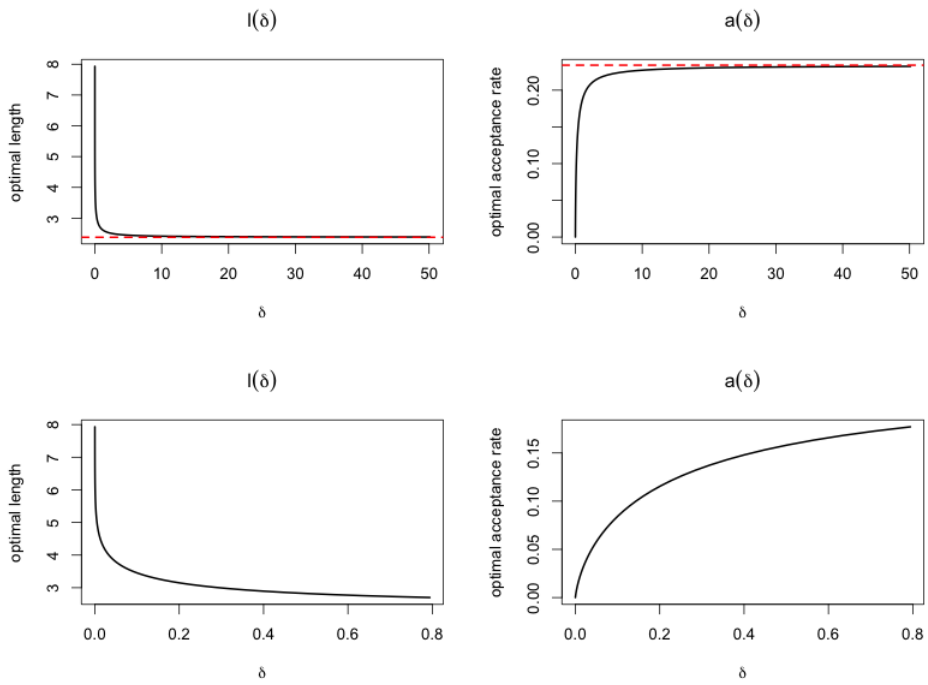
Démonstration. It is easy to see that (H1) implies $\tilde{f}(\cdot) = \pi(\cdot)$ and so $\rho_1(x,y) = r(x,y)$. Moreover, by definition, $\rho_2(x,y) = r(x,y)/\rho_1(x,y) = 1$ and hence the second test is always accepted. The acceptance rate reduces then to just the ratio $\tilde{f}(y)/\tilde{f}(x) = \rho_1(x,y)$.

The second part of the lemma follows directly from Theorem 1.1 in Roberts, Gelman et Gilks, 1997. □

Let us stress that almost all assumptions in the above Lemma can be relaxed and that performances are robust against small deviances from those assumptions, as shown by the literature on standard Metropolis–Hastings. Obtaining analytical results without such conditions, while possible, requires however a considerable mathematical effort.

We now state the main practical implication of Lemma 7.

FIGURE 2.3: Two top panels : behaviour of $\ell^*(\delta)$ and $\alpha^*(\delta)$ as the relative cost varies. Note that for $\delta >> 1$ the optimal values converges towards the values computed for the standard Metropolis–Hastings (dashed in red). Two bottom panels : close–up of the interesting region for $0 < \delta < 1$.



Proposition 4. *If the conditions of Lemma 7 holds, the optimal average acceptance rate $\alpha^*(\delta)$ is independent of I.*

Démonstration. Consider $\mathbf{Eff}(\delta, \ell)$ in terms of $(\delta, a(\ell))$:

$$a = g(\ell) = 2\Phi\left(-\frac{\ell\sqrt{I}}{2}\right) \quad ; \quad \ell = g^{-1}(a) = -\Phi^{-1}\left(\frac{a}{2}\right)\frac{2}{\sqrt{I}}$$

$$\mathbf{Eff}(\delta, a) = \frac{\frac{4}{I}\left[\Phi^{-1}\left(\frac{a}{2}\right)^2 a\right]}{\delta + a} = \frac{4}{I}\left\{\frac{1}{\delta + a}\left[\Phi^{-1}\left(\frac{a}{2}\right)^2 a\right]\right\}$$

where we dropped the dependence on ℓ in a for notation's sake. It is now evident that to maximise $\mathbf{Eff}(\delta, a)$ in a we only need maximise $\left\{\frac{1}{\delta+a}\left[\Phi^{-1}\left(\frac{a}{2}\right)^2 a\right]\right\}$, which is independent of I. □

The optimal scale of the proposal $\ell^*(\delta)$ and the optimal acceptance rate $a^*(\delta)$ are thus given as functions of δ. In particular, as the relative cost of computing $\rho_1(x, y)$ with respect to $\rho_2(x, y)$ decreases, the proposed moves become bolder, in that ℓ^* increases and a^* decreases, since rejecting costs the algorithm little in terms of time, while every accepted move results in an almost independent sample. On the contrary when δ grows larger the chain rapidly approaches a Metropolis–Hastings behaviour, as it is no longer convenient to reject early. Figure 2.3 helps visualise the result.

2.3.2 Ranking the Blocks

As mentioned at the end of Section 2.1, the order in which the factors $\rho_i(x, y)$ are tested has a strong influence on the performance of the algorithm. Delayed Acceptance was first developed in Fox et Nicholls, 1997; Christen et Fox, 2005 to speed up computations using a cheap approximation $\tilde{\pi}(\cdot)$ of the target distribution $\pi(\cdot)$ as a first step before computing the actual, and costly, Metropolis–Hastings ratio $\pi(y)/\pi(x)$ only in the cases where the acceptance test based on the approximation $\tilde{\pi}$ was satisfied. The main idea, namely to avoid the computation of the most costly parts as often as possible, remains relevant even for factorisations composed of more than two terms.

Consider an i.i.d. framework; the target (in x) is given by

$$\pi(x|Z) \propto p(x) \times L(Z|x) = p(x|Z) \times \prod_{i=1}^{n} f(z_i|x)$$

where $Z = (z_1, \ldots, z_n)$ is an i.i.d. sample from $f(z|x)$ and $p(x)$ is the prior distribution for x, we can always consider the decomposition

$$r(x, y) = \prod_{i=1}^{K} \xi_i(x, y) \tag{2.6}$$

where each $\xi_i(x, y)$ is made of a small number of density ratio terms, with one including the prior and proposal ratios. In the limit, it is feasible if not necessarily efficient to consider the case $K = n + 1$ with

$$\xi_i(x, y) = \frac{f(z_i|y)}{f(z_i|x)} \quad i = 1, \ldots, n \quad \text{and} \quad \xi_{n+1}(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)}.$$

Assuming the computing cost is comparable for all terms, a solution for optimising the order of these factors ranks the entries according to the success rates observed so far, starting with the least successful values. Alternatively, the factorisation can start with the ratio that has the highest variance, since it is the most likely to be rejected. (Note however that poor factorisations (2.6) lead to very low acceptance rates, as for instance when picking only outliers in a given group of observations.) Lastly, we can rank factors by their correlation with the full Metropolis–Hastings ratio; taking the argument to the limit, if the first factor has a perfect correlation with $r(\cdot, \cdot)$ then all the successive terms must be accepted and their order is hence of no interest.

This later setting is akin to considering the hypothetical optimal solution introduced in Section 2.3.1 with only two terms in the decomposition. Let a small number of the best scoring terms be merged to form ρ_1 and let the remaining factors become ρ_2. $\rho_1(x, y) = \tilde{\pi}(y)/\tilde{\pi}(x)$ is then highly correlated to $r(x, y)$, $\rho_2(x, y) \sim 1$ for every (x, y) and hence $\tilde{\pi}(x)$ is a close approximation of the target, albeit probably flattened, which is exactly what we want (see Section 2.2).

As all these features can be evaluated for each subsample while running a chain with acceptance ratio factored as in (2.6), an implementation based on this intuition is then to take

$$\tilde{\pi}_{Z^*}(x) \propto p(x)^{m/n} \prod_{i=1}^{m} f(z_i^*|x) \quad \text{or} \quad \tilde{\pi}_{Z^*}(x) \propto \prod_{i=1}^{m} f(z_i^*|x)$$

with $m < n$, where $Z^* = (z_1^*, \ldots, z_m^*)$ is a subsample of Z. At each iteration t of the Markov chain we compute all the $\xi_i(x^{(t-1)}, y)$ and $Z^{*(t)}$ is chosen as the subset that maximise the observed correlation between the values of $\frac{\tilde{\pi}_{Z^*}(x^{(j)})}{\tilde{\pi}_{Z^*}(x^{(j)})} \quad (j = 1, \ldots, t-1)$ and the full Metropolis–Hastings ratio (or whatever other selected criterion). As computing the real $\operatorname{argf\,max}_{Z^* \subset Z}$ is expensive, in our practical implementation we resort to a *forward* selection scheme; starting with the factor ξ_i with the maximal correlation we build $Z^{*(t)}$ merging one term at a time until a desired correlation level is achieved, the observed correlation after including another term does not grow more than a small $\epsilon > 0$ or the size of $Z^{*(t)}$ has reached a critical point for computational purposes (e.g. 10% of the whole sample Z).

A relevant warning is that if we rearrange terms during the run, not only reordering but also merging them, in accordance to their correlation with the unmodified ratio, the resulting method has no theoretical guarantee since the kernel is potentially changing at each iteration depending on properties of previous samples (Gelfand et Sahu, 1994).

As with standard adaptive MCMC (Roberts et Rosenthal, 2005) we resort thus to a *finite adaptation scheme*; we start with a fixed number of iterations to rank and rearrange the factors, followed by a fixed ordering to achieve ergodicity of the chain. We test this procedure in Section 2.5.1 on a simulated example.

Finally note that while we focused on the i.i.d. setting, in more complex cases where the ratio is factored and Delayed Acceptance can be applied, it is often the case that the optimal ordering of such factors is already known.

2.4 Relation with other methods

2.4.1 Delayed Acceptance and Prefetching

Prefetching, as defined by Brockwell, 2006, is a programming method that accelerates the convergence of a single MCMC chain by guessing future states in the path of a random walk Metropolis–Hastings Markov chain in order to use any additional computing power available, in the form of extra parallel processors, to calculate in advance necessary quantities (like the Metropolis–Hastings ratio) so that when the chain reaches a given state the computationally-heavy part of that iteration are ready.

Clearly the usefulness of this technique depends on our ability to guess the path of the chain correctly and hence many advanced prefetching strategies make use of the observed acceptance rate of the chain or even of a fast approximation $\tilde{\pi}$ of the target distribution to select the most likely future outcomes.

Since an in-depth exploration of prefetching is outside the scope of this work the reader is referred to Strid, 2010 and citations therein for a complete discussion of the argument.

As mentioned above and demonstrated in Strid, 2010; Angelino et al., 2014 if a cheap approximation $\tilde{\pi}$ of the target density is available, it can be used to select more likely future paths of the chain and this results in an efficient prefetching algorithm.

In our case the master process sequentially samples from the (Delayed Acceptance) chain by checking only the (assumed) inexpensive first approximation

$\rho_1(x,y) = \tilde{\pi}(y)/\tilde{\pi}(x)$ while the other additional processors provide him the more expensive $\rho_2(x,y) = \pi(y)\tilde{\pi}(x)/\pi(x)\tilde{\pi}(y)$ computed beforehand thanks to prefetching. The theoretical properties of the chain are unchanged while the achievable speed-up may be substantial, especially for the first few additional processors.

2.4.2 Alternative procedure for Delayed Acceptance

In the case that every factor $\rho_i(x,y)$ has roughly the same computational cost, Philip Nutzman suggested (personal communication) that Delayed Acceptance can be slightly modified by taking the overall acceptance probability

$$\prod_{i=1}^{d} \min\left\{\rho_i(x,y), 1\right\} \quad \text{to be instead} \quad \min_{k=1,\dots,d} \left\{ \prod_{i=1}^{k} \rho_i(x,y), 1 \right\}.$$

Such a decomposition follows from the same idea that one would like to compute as few factors as possible once one realizes that the proposal is likely to be rejected. Under this modification the associated Markov chain still achieves the correct target in the stationary regime and the procedure satisfies detailed balance, provided the ordering of the terms is uniformly random.

An interesting consequence of this modification is that, as the number of factor increases, the acceptance rate eventually stabilises, while for the method described in Section 2.1 the acceptance rate decreases to zero. This property is indeed appealing, even thought this procedure logically takes longer to complete when compared with the standard Delayed Acceptance (albeit less than the reference Metropolis–Hastings procedure).

The evident disadvantage of the modification in a general setting is that detailed balance implies that the factors are computed in a random order at each iteration, making vain any attempt to adapt in terms of the ordering (Section 2.3.2) or to set the order based on respective computational costs.

This drawback can be somewhat reduced by combining the above two approaches ; consider the decomposition

$$\pi(y)\, q(y,x)/\pi(x)q(x,y) = \left[\prod_{i=1}^{d_1} \rho_i(x,y)\right] \times \left[\prod_{j=1}^{d_2} \phi_j(x,y)\right]$$

where $d_1 + d_2 = d$ and the factors ρ_i and ϕ_j represent respectively cheap factors and costly factors. By taking now

$$\min_{(m=1,\dots,d_1)} \left\{ 1, \prod_{i=1}^{m} \rho_i(x,y) \right\} \times \min_{(k=1,\dots,d_2)} \left\{ 1, \prod_{j=1}^{k} \phi_i(x,y) \right\}$$

the algorithm computes cheap factors first and expensive factors last, applying the symmetry requirement to satisfy detail balance inside each of both subsets. Clearly the above can be generalised to a larger number of subsets, each with d_i factors in it. Intuitively, this last modification can be explained as an early rejection of each of the intermediate acceptance/rejection steps inside a Delayed Acceptance scheme.

Remark 4. Interestingly if $d_l = 1\ \forall l$ (l being the number of subsets considered) this procedure reduces to Delayed Acceptance, and for l that increases and $d_l > 1\ \forall l$ this combined technique will have a even lower overall acceptance rate than standard Delayed Acceptance.

2.4.3 Delayed Acceptance and Slice Sampling

As a final remark, we stress another analogy between our Delayed Acceptance algorithm and slice sampling (Neal, 1997 ; Robert et Casella, 2004). Based on the same decomposition (2.1), slice sampling proceeds as follows

1. simulate $u_1, \ldots, u_n \sim \mathcal{U}(0,1)$ and set $\lambda_i = u_i \ell(\theta|x_i)$ $(i = 1, \ldots, n)$;
2. simulate θ' as a uniform under the constraints $\ell_i(\theta'|x_i) \geq \lambda_i$ $(i = 1, \ldots, n)$.

to compare with Delayed Acceptance which conversely

1. simulate $\theta' \sim q(\theta'|\theta)$;
2. simulate $u_1, \ldots, u_n \sim \mathcal{U}(0,1)$ and set $\lambda_i = u_i \ell(\theta|x_i)$ $(i = 1, \ldots, n)$;
3. check that $\ell_i(\theta'|x_i) \geq \lambda_i$ $(i = 1, \ldots, n)$.

The differences between both schemes are thus that (a) slice sampling always accepts a move, (b) slice sampling requires the simulation of θ' under the constraints, which may prove infeasible, and (c) Delayed Acceptance re-simulates the uniform variates in the event of a rejection. In this respect, Delayed Acceptance appears as a "poor man's" slice sampler in that values of $\theta' s$ are proposed until one is accepted.

2.5 Examples

To illustrate the improvement brought by Delayed Acceptance, we study three different realistic settings that reflect on the generality of the method. First, in Section 2.5.1 we consider a Bayesian analysis of a logistic regression model, to assess the computational gain brought by our approach in a "Big-Data" environment where obtaining the likelihood is the main computational burden. Secondly (Section 2.5.2) we examine a high dimensional toy Normal-Normal model, sample with a geometric Metropolis adjusted Langevin algorithm where the main computational cost comes from the proposal distribution which is position specific and involves derivatives of the density up till the third level, which are computed numerically at each iteration. Finally in Section 2.5.3 we investigate a mixture model where a formal Jeffreys prior is used, as it is not available in closed-form and does require an expensive approximation by numerical or Monte Carlo means. The latter example comes as a realistic setting where the prior itself is a burdensome object, even for small datasets.

2.5.1 Logistic Regression

While a simple model, or due to its simplicity, logistic regression is widely used in applied statistics, especially in classification problems. The challenge in the Bayesian analysis of this model is not generic, since simple Markov Chain Monte Carlo techniques providing satisfactory approximations, but stems from the data-size itself. This explains why this model is used as a benchmark in some of the recent accelerating papers (Korattikara, Chen et Welling, 2013 ; Neiswanger, Wang et Xing, 2013 ; Scott et al., 2013 ; Wang et Dunson, 2013). Indeed, in "big Data" setups, MCMC is deemed to be progressively inefficient and researchers are striving to keep simulation effective, focusing mainly on parallel computing and on sub-sampling but also on replacing the classic Metropolis scheme itself.

We tested the proposed method against the standard Metropolis–Hastings algorithm on 10^6 simulated data with a 100-dimensional parameter space. The proposal

Algorithm	**relative ESS** (aver.)	**relative ESJD** (aver.)	**relative Time** (aver.)
DA-MH over MH	1.1066	12.962	0.098

Algorithm	**relative Eff gain (ESS)** (aver.)	**relative Eff gain (ESJD)** (aver.)
DA-MH over MH	5.47	56.18

TABLE 2.1: Comparison between MH and MH with Delayed Acceptance on a logistic model. **ESS** is the effective sample size, **ESJD** the expected square jumping distance, **time** is the computation time.

distribution is Gaussian : $y|x \sim \mathcal{N}(x, \Sigma)$ with Σ initialised to be $0.2 \times I_d$ (d being the dimension of the parameter space) and then adapted. The Metropolis–Hastings benchmark was made adaptive by targeting the asymptotic optimal acceptance rate of $\alpha^* = 0,234$ (ROBERTS, GELMAN et GILKS, 1997).

Delayed Acceptance was optimised first against the ordering of the factors as explained in Section 2.3 ; we split the data into subsamples of 10 elements and computed their empirical correlation with the full Metropolis–Hastings ratio as a criterion. Once these estimates were stable we merged into the surrogate target $\tilde{f}$ the smallest number of subsamples needed to achieve a ≥ 0.85 correlation with $r(x, y)$. As soon as the ordering was fixed we computed δ, the relative cost of the obtained ρ_1 with respect to the full ratio, and run the chain for the remaining iterations optimising Σ against the optimal acceptance rate found through (2.5). We also added the modification explained in Section 2.2.4 with c set such that b was slightly lower than the optimal acceptance rate above.

We collected 100 repetitions of the experiment and the results are presented in Table 2.1. Before commenting the results we highlight the fact that this situation may seem not particularly appealing for Delayed Acceptance and in fact straight application of the method by randomly choosing the composition of ρ_1 and ρ_2 may lead to variable results. Further coding effort is required here in order to choose adaptively how to split the MH ratio. Borrowing from both Section 2.3.2 and the end of Section 2.3.1, i.e. by choosing during the brief burn-in of the chain which subset best represents the whole likelihood and then, based on how populated that subset is, targeting a specific acceptance ratio, produces both a completely automated MCMC version for this kind of data (*iid*) and better results under a time constraint.

As shown in Table 2.1, while the assumption made in Section 2.3 not completely satisfied, the relative efficiency of Delayed Acceptance is higher that for MH by a factor of almost 6. We measured efficiency trough *effective sample size* (ESS, from the **coda** R package (PLUMMER et al., 2006)) or *expected square jumping distance* (ESJD). By choosing the first subsample to be *informative* on the whole ratio there is practically no loss on ESS (while the estimated ESJD actually increased) and, given the significantly reduced acceptance rate, the computing time is usually less then a fourth of the computing time of the corresponding optimal MH, taking into account the first part of chain used to determine the blocks ranking.

2.5.2 G-MALA with Delayed Acceptance

MALA and Geometric MALA :

Random walk Metropolis–Hastings, while generic and popular, can struggle with posterior distributions in high dimensions or in the presence of high correlation between some components. In such cases it is inefficient, with low acceptance rate, poor mixing and highly correlated samples. Metropolis adjusted Langevin algorithm (MALA, see for instance Roberts et Stramer, 2002) has been devised to overcome these difficulties by taking advantage of the gradient of the target distribution in the proposal mechanism, making the Markov chain more robust with respect to the dimension of the problem and proposing broader moves with higher probability. MALA is based on a Langevin diffusion, with the target (the posterior distribution $\pi(\theta|y)$ in our case) as a stationary distribution, defined by the SDE

$$\frac{d\theta}{dt} = \nabla_\theta \log(\pi(\theta|y))\frac{dt}{2} + \frac{dB}{dt}$$

where B is a Brownian motion. Using a first-order discretisation the diffusion gives the following proposal mechanism :

$$\theta' = \theta^{(i-1)} + \varepsilon^2 \nabla_\theta \log(\pi(\theta^{(i-1)}|y))/2 + \varepsilon Z$$

where ε is the step-size for the Euler's integration and $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This discretisation is then compensated by introducing an accept/reject probability similar to a Metropolis–Hastings algorithm.

This diffusion is isotropic and will hence still be inefficient for highly correlated components or with very different scales, as the step size ε is fixed across dimensions. Roberts et Stramer, 2002 propose to alleviate the issue using a pre-conditioning matrix A so that the proposal becomes

$$\theta' = \theta^{(i-1)} + \varepsilon^2 A^T A \nabla_\theta \log(\pi(\theta^{(i-1)}|y))/2 + \varepsilon A Z.$$

Christensen, Roberts et Rosenthal, 2003 demonstrate however that defining this matrix in general can be difficult and that tuning on the go may result in an inappropriate asymptotic behaviour.

In a recent work Girolami et Calderhead, 2011 propose the Geometric-MALA in order to overcome this difficulty, advising the use of a position specific metric for the matrix A, which takes advantage of the geometry of the target space that the chain is exploring. They suggest in particular the Fisher-Rao metric tensor. In terms of Bayesian inference, where the target distribution is the posterior density, this choice translates into $A^T A$ being the expected Fisher information matrix plus the negative Hessian of the log-prior.

This theoretically efficient solution also performs well in practice but comes with a serious computational burden in the fact that at every evaluation of the Metropolis–Hastings ratio derivatives up till the third order of our log-target distribution are needed and, in the event of them being analytically not available, expensive numerical approximations are to be computed (see equation (10) of Girolami et Calderhead, 2011).

Sampling with Delayed Acceptance and GMALA :

Geometric-MALA represent a perfect application for Delayed Acceptance since we can naturally divide its acceptance ratio into the product of the posterior ratio and the ratio of the proposals, the latter to be only computed when the proposed point is associated with a relatively large posterior probability.

As described above, the computational bottleneck of the G-MALA lays in the computation of the third derivative of our log-target at the proposed point, while the computation of the posterior itself has usually a low relative cost. Moreover even with a non-symmetric efficient proposal mechanism (the discretised Langevin diffusion) G-MALA is still close to a random walk and we expect the ratio of the proposal to be near 1, especially at equilibrium especially when ε is small. Therefore, the first ratio is inexpensive, relative to the second one, while the decision reached at the first stage should be consistent with the overall acceptance rate.

Given that optimal scaling for MALA in terms of the dimension d of the target differs from the random-walk setting (see Roberts et Rosenthal, 2001), we set the variance of the random-walk normal component as $\sigma_d^2 = \frac{\ell^2}{d^{1/3}}$. Borrowing from Section 2.3.1, we can obtain the optimal acceptance rate for the DA-MALA, through Equation (2.5), by maximising

$$\mathbf{Eff}(\delta, \ell) = \frac{h(\ell)}{\delta + \mathbb{E}\left[\tilde{\alpha}\right] - \delta \times \mathbb{E}\left[\tilde{\alpha}\right]} = \frac{2\ell^2 \Phi(-\frac{K\ell^3}{2})}{\delta + 2\Phi(-\frac{K\ell^3}{2}) \times (1-\delta)}$$

or equivalently

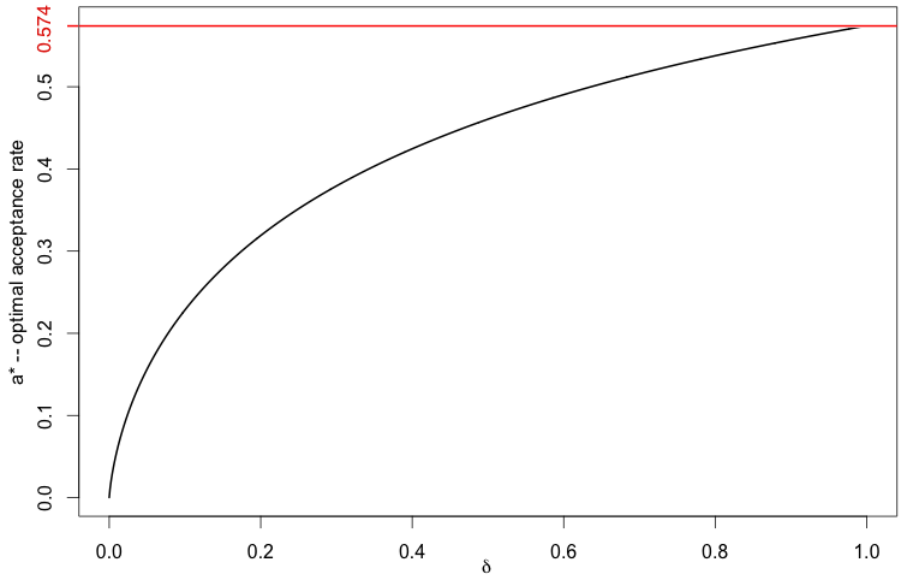
$$\mathbf{Eff}(\delta, a) = -\left(\frac{2}{K}\right)^{\frac{2}{3}} \left[\frac{a\Phi^{-1}\left(\frac{a}{2}\right)^{\frac{2}{3}}}{\delta + a(1-\delta)}\right].$$

In the above the computational cost per iteration is taken to be $c = \delta C$ for the posterior ratio, $C = 1$ for the proposal ratio (and hence $c + \mathbb{E}\left[\tilde{\alpha}\right](C - c)$ for the whole kernel), $h(\ell)$ is again the speed of the limiting diffusion process and K is a measure of roughness of the target distribution, depending on its derivatives. Since the optimal a^* is independent from K, we do not define it more rigorously, referring to Roberts et Rosenthal, 2001. Figure 2.4 shows that a^* decreases with δ, as is the case with random-walk Metropolis–Hastings. It reaches the known optimum for the standard MALA when $\delta = 1$.

Simulation study :

To test the above assumptions we ran a toy MALA example where we drew 100 samples from a $\mathcal{N}_d(\theta, I)$, with $d = 10$; $\pi(\theta)$ was set to be $\mathcal{N}_d(0, 100)$. Figure 2.5 presents an example run. We then repeated the experiment 100 times and computed an average efficiency gain, defined either as ESS or as the ESJD, over the computing time. We computed δ at each run by averaging a few computed derivatives, required by the proposal ratio. We then adapt ε to get the optimal acceptance rate, being conservative in order to avoid overflow issues with the first-order numerical integrator. Results are presented in Table 2.2. Delayed Acceptance exhibits improvement by a factor of 10 in this example, obtained almost for free in terms of coding time.

FIGURE 2.4: Optimal acceptance rate for the DA-MALA algorithm as a function of δ. In red, the optimal acceptance rate for MALA obtained by ROBERTS et ROSENTHAL, 2001 is met for $\delta = 1$.

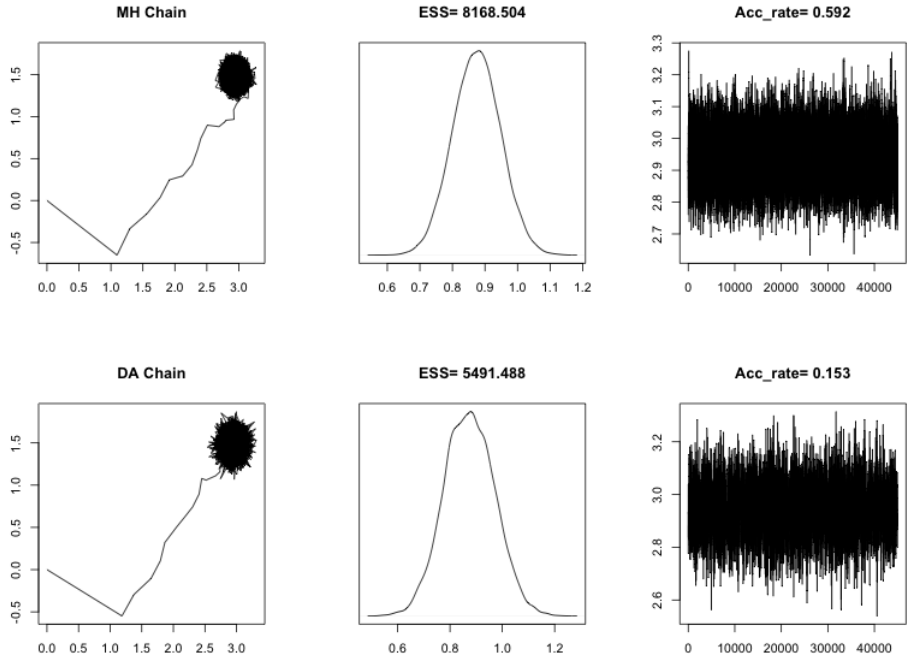


Algorithm	**ESS** (aver.)	**ESS** (sd)	**ESJD** (aver.)	**ESJD** (sd)	**time** (aver.)	**time** (sd)
MALA	7504.486	107.21	5244.946	983.473	176078	1562.3
DA-MALA	6081.023	121.42	5373.253	2148.761	17342.91	6688.3

Algorithm	**a** (aver.)	**ESS/time (aver.)**	**ESJD/time (aver.)**
MALA	0.661	0.04	0.03
DA-MALA	0.09	0.35	0.31

TABLE 2.2: Comparison between standard geometric MALA and geometric MALA with Delayed Acceptance, with **ESS** the effective sample size, **ESJD** the expected square jumping distance, **time** the computation time and **a** the observed acceptance rate.

FIGURE 2.5: Comparison between geometric MALA (top panels) and geometric MALA with Delayed Acceptance (bottom panels) : marginal chains for two arbitrary components (left), estimated marginal posterior density for an arbitrary component (middle), 1D chain trace evaluating mixing (right).



HMC with Delayed Acceptance :

As a side note, while the reasoning applied to MALA does theory apply to Hamiltonian Monte Carlo (HMC), the computational gain obtained through Delayed Acceptance is only connected with avoiding some proposal computations. In a general HMC though (with both point-dependent and independent pre-conditioning matrices), proposing a new value still involves the computation of $L-1$ (with L the number of steps in the discretised–Hamiltonian integration) derivatives, as only the starting point is recovered from the previous iteration, while computing the final Metropolis–Hastings ratio involves just the extra computation at the end point. Therefore, in this setting, the computational gain is much reduced.

2.5.3 Mixture Model

Non-Informative inference on a Mixture Model :

Consider a standard mixture model (McLACHLAN et PEEL, 2000) with a fixed number of components

$$\sum_{i=1}^{k} w_i \, f(x|\theta_i)\,, \quad \text{with} \quad \sum_{i=1}^{k} w_i = 1\,. \tag{2.7}$$

This standard setting nonetheless offers a computational challenge in that the reference objective Bayesian approach based on the Fisher information and the associated Jeffreys prior (JEFFREYS, 1939 ; ROBERT, 2007) is not readily available for computational reasons and has thus not been implemented so far. Proxys using

Jeffreys priors on the components of (2.7) have been proposed instead, with the drawback that since they always lead to improper posteriors, ad hoc corrections have to be implemented (Diebolt et Robert, 1994 ; Roeder et Wasserman, 1997 ; Stephens, 1997).

When relying instead on dependent improper priors, it is not always the case that the posterior distribution is improper. For instance, Robert et Titterington, 1998 provide a location-scale representation that allows for some improper prior. In the current paper, we consider instead the genuine Jeffreys prior for the complete set of parameters in (2.7), derived from the Fisher information matrix for the whole model. While establishing the analytical properness of the associated posterior is beyond the goal of the current paper, we handle large enough samples to posit that a sufficient number of observations is allocated to each component and hence the likelihood function dominates the prior distribution. (In the event the posterior remains improper, the associated MCMC algorithm should exhibit a transient behaviour.)

Therefore, this is an appropriate and realistic example for evaluating Delayed Acceptance since the computation of the prior density is clearly costly, relying on many integrals of the form :

$$-\int_{\mathcal{X}} \frac{\partial^2 \log\left[\sum_{i=1}^{k} w_i\, f(x|\theta_i)\right]}{\partial\theta_h \partial\theta_j} \left[\sum_{i=1}^{k} w_i\, f(x|\theta_i)\right] \mathrm{d}x\,. \tag{2.8}$$

Indeed, these integrals cannot be computed analytically and thus their derivation involve numerical or Monte Carlo integration. This setting is such that the prior ratio—as opposed to the more common case of the likelihood ratio—is the costly part of the target evaluated in the Metropolis–Hastings acceptance ratio. Moreover, since the Jeffreys prior involves a determinant, there is no easy way to split the computation in more parts than "prior × likelihood". Hence, the Delayed Acceptance algorithm can be applied by simply splitting between the prior $p^J(\psi)$ and the likelihood $\ell(\psi|x)$ ratios, the later being computed first. Moreover, since the proposed prior is "non informative", its influence on the definition of the posterior distribution should be small with respect to the likelihood function and, then, computing the likelihood ratio first should not have a substantial impact on the acceptance rate. However, the improper nature of the prior means using a second acceptance ratio solely based on the prior can create trapping states in practice, even though the method remains theoretically valid. We therefore opted for stabilising this second step by saving a small fraction of the likelihood, corresponding to 5% of the sample, to regularise this second acceptance ratio. This choice translates into Algorithm 2.2.

Simulation study :

An experiment comparing a standard Metropolis–Hastings implementation with a Metropolis–Hastings version relying on Delayed Acceptance is summarised in Table 2.3. Data were simulated from the following Gaussian mixture model :

$$f(y|\theta) = 0.10\mathcal{N}(-10, 2) + 0.65\mathcal{N}(0, 5) + 0.25\mathcal{N}(15, 7). \tag{2.9}$$

Both the standard Metropolis–Hastings and the Delayed Acceptance version are adapted against their respective optimal acceptance rate, which is computed to be

Algorithm 2.2 Metropolis–Hastings with Delayed Acceptance for Mixture Models

Set $\ell_2(\cdot|x) = \prod_{i=1}^{\lfloor pn \rfloor} \ell(\cdot|x_i)$ and $\ell_1(\cdot|x) = \prod_{i=\lfloor pn \rfloor+1}^{n} \ell(\cdot|x_i)$ where $p \in (0,1)$

1. Simulate $\psi' \sim q(\psi'|\psi)$;
2. Simulate $u_1, u_2 \sim \mathcal{U}(0,1)$ and set $\lambda_1 = u_1 \ell_1(\psi|x)$;
3. **if** $\ell_1(\psi'|x) \leq \lambda_1$, repeat the current parameter value and return to 1 ;
   else set $\lambda_2 = u_2 \ell_2(\psi|x) p^J(\psi)$;
4. **if** $\ell_2(\psi'|x) p^J(\psi') \geq \lambda_2$ accept ψ' ;
   else repeat the current parameter value and return to 1.

Algorithm	**ESS** (aver.)	**ESS** (sd)	**ESJD** (aver.)	**ESJD** (sd)	**time** (aver.)	**time** (sd)
MH	1575.963	245.96	0.226	0.44	513.95	57.81
MH + DA	628.767	87.86	0.215	0.45	42.22	22.95

TABLE 2.3: Comparison using different performance indicators in the example of mixture estimation, based on 100 replicas of the experiments according to model (2.9) with a sample size $n = 500$, 10^5 MH simulations and 500 samples for the prior estimation. ("ESS" is the effective sample size,"time" is the computational time). The actual averaged gain ($\frac{ESS_{DA}/ESS_{MH}}{time_{DA}/time_{MH}}$) is 9.58, higher than the "double average" that the table above suggests as being around 5.

2%, given that δ is empirically established to be 0.01 using 500 samples for the Monte Carlo estimation of the prior. As a consequence the MH+DA algorithm will produce less unique samples in the total 10^5 iterations of the chain, as reflected in the lesser ESS in Table 2.3, but this is counterbalanced by the impressive decrease in computing time, leading again to an overall gain in terms of ESS/t of about 9.

2.6 Conclusion

We introduced in this paper Delayed Acceptance, a generic and easily implemented modification of the standard Metropolis–Hastings algorithm that splits the acceptance rate into more than one step in order to increase the computational efficiency of the resulting MCMC, under the sole condition that the Metropolis–Hastings ratio can be factorised this way.

The choice of splitting the target distribution into parts ultimately depends on the respective costs of computing the said parts and of reducing theoretically the overall acceptance rate and expected square jump distance (ESJD). Still, this generic alternative to the standard Metropolis–Hastings approach can be considered on a customary basis, since it both requires very little modification in programming and can be easily tested against the basic version, both empirically and theoretically by the results of (2.2). The Delayed Acceptance algorithm presented in (2.1) can significantly decrease the computational time *per se* as well as the overall acceptance rate. Nevertheless, the examples presented in Section 2.5 suggest that the gain in terms of computational time is not linear in the reduction of the acceptance rate, especially in the presence of optimisation techniques like (2.3).

Furthermore, our Delayed Acceptance algorithm does naturally merge with the

widening range of prefetching techniques, in order to make use of parallelisation and reduce the overall computational time even more significantly. Most settings of interest are open to take advantage of the proposed method, if mostly in the situation of Bayesian statistics where the target density and/or the Metropolis–Hastings ratio always allow for a natural factorisation. The case when the likelihood function can be factorised in an useful way represents the best gain brought by our solution, in terms of computational time, and it may easily improve even more by exploiting parallelisation techniques.

Acknowledgements

Thanks to Christophe Andrieu for a very helpful discussion on an earlier version of the manuscript. The massive help provided by Jean-Michel Marin and Pierre Pudlo towards an implementation on a large cluster has been fundamental in the completion of this work. Thanks to Samuel Livingstone for suggesting the Geometric MALA example and finally thanks to Philip Nutzman for the interesting conversation and for the suggestion of the method proposed in (2.4.2). Christian P. Robert research is partly financed by Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) on the 2012–2015 ANR-11-BS01-0010 grant "Calibration" and by a 2010–2015 senior chair grant of Institut Universitaire de France. Marco Banterle PhD is funded by Université Paris Dauphine.

References

Andrieu, Christophe, Anthony Lee et Matti Vihola (2013). "Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers". In : *arXiv preprint arXiv:1312.6432*.

Angelino, E. et al. (2014). "Accelerating MCMC via Parallel Predictive Prefetching". In : *arXiv preprint arXiv:1403.7265*.

Brockwell, A.E. (2006). "Parallel Markov chain Monte Carlo Simulation by Pre-Fetching". In : *J. Comput. Graphical Stat.* 15.1, p. 246–261.

Christen, J.A. et C. Fox (2005). "Markov chain Monte Carlo using an approximation". In : *Journal of Computational and Graphical Statistics* 14.4, p. 795–810.

Christensen, Ole F., Gareth O. Roberts et Jeffrey S. Rosenthal (2003). *Scaling Limits for the Transient Phase of Local Metropolis–Hastings Algorithms*.

Diebolt, J. et Christian P. Robert (1994). "Estimation of Finite Mixture Distributions by Bayesian Sampling". In : 56, p. 363–375.

Fox, C. et G. Nicholls (1997). "Sampling conductivity images via MCMC". In : *The Art and Science of Bayesian Image Analysis*, p. 91–100.

Gelfand, A.E. et S.K. Sahu (1994). "On Markov chain Monte Carlo acceleration". In : *J. Comput. Graph. Statist.* 3.3, p. 261–276.

Girolami, M. et B. Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, p. 123–214.

Golightly, A., D. A. Henderson et C. Sherlock (2014). "Delayed acceptance particle MCMC for exact inference in stochastic kinetic models". In : *ArXiv e-prints*. arXiv : `1401.4369 [stat.CO]`.

Jeffreys, H. (1939). *Theory of Probability*. Oxford, U.K. : Clarendon Press.

Korattikara, A., Y. Chen et M. Welling (2013). "Austerity in MCMC land: Cutting the Metropolis-Hastings budget". In : *arXiv preprint arXiv:1304.5299*.

McLachlan, G. J. et D. Peel (2000). *Finite Mixture Models*. New York : J. Wiley.

Mengersen, K.L. et R.L. Tweedie (1996). "Rates of convergence of the Hastings and Metropolis algorithms". In : 24, p. 101–121.

Neal, Peter et Gareth Roberts (2011). "Optimal Scaling of Random Walk Metropolis Algorithms with Non–Gaussian Proposals". English. In : *Methodology and Computing in Applied Probability* 13.3, p. 583–601. issn : 1387-5841. doi : `10.1007/s11009-010-9176-9`. url : `http://dx.doi.org/10.1007/s11009-010-9176-9`.

Neal, R.M. (1997). *Markov chain Monte Carlo methods based on 'slicing' the density function*. Rapp. tech. University of Toronto.

Neiswanger, W., C. Wang et E. Xing (2013). "Asymptotically Exact, Embarrassingly Parallel MCMC". In : *arXiv preprint arXiv:1311.4780*.

Peskun, P. H. (1973a). "Optimum Monte-Carlo sampling using Markov chains". In : *Biometrika* 60, p. 607–612.

Peskun, P.H. (1973b). "Optimum Monte Carlo sampling using Markov chains". In : 60, p. 607–612.

Plummer, Martyn et al. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC". In : *R News* 6.1, p. 7–11. URL : http://CRAN.R-project.org/doc/Rnews/.

Robert, C.P. (2007). *: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York : Springer.

Robert, C.P. et G. Casella (2004). *Monte Carlo Statistical Methods*. second. Springer-Verlag.

Robert, C.P. et M. Titterington (1998). "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation". In : *Statistics and Computing* 8.2, p. 145–158.

Roberts, G. O., A. Gelman et W. R. Gilks (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". In : *Ann. Appl. Probab.* 7.1, p. 110–120. ISSN : 1050-5164.

Roberts, G. O. et O. Stramer (2002). "Langevin Diffusions and Metropolis–Hastings Algorithms." In : *Methodology and Computing in Applied Probability* 4.4, p. 337–358. ISSN : 1387-5841. DOI : 10.1023/A:1023562417138.

Roberts, Gareth O. et Jeffrey S. Rosenthal (2001). "Markov Chains and deinitializing processes". In : *Scandinavian Journal of Statistics* 28, p. 489–504.

Roberts, G.O. et J.S. Rosenthal (2005). "Coupling and Ergodicity of Adaptive MCMC". In : *J. Applied Proba.* 44, p. 458–475.

Roberts, G.O. et R.L. Tweedie (1996). "Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms". In : 83, p. 95–110.

Roeder, K. et L. Wasserman (1997). "Practical Bayesian density estimation using mixtures of Normals". In : 92, p. 894–902.

Scott, S.L. et al. (2013). "Bayes and big data: The consensus Monte Carlo algorithm". In : *EFaBBayes 250 conference* 16.

Sherlock, C. et al. (2013). "On the efficiency of pseudo-marginal random walk Metropolis algorithms". In : *ArXiv e-prints*. arXiv : 1309.7209 [stat.CO].

Sherlock, Chris et Gareth Roberts (2009). "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". In : *Bernoulli* 15.3, p. 774–798. DOI : 10.3150/08-BEJ176. URL : http://dx.doi.org/10.3150/08-BEJ176.

Shestopaloff, A. Y. et R. M. Neal (2013). "MCMC for non-linear state space models using ensembles of latent sequences". In : *ArXiv e-prints*. arXiv : 1305.0320 [stat.CO].

Stephens, M. (1997). "Bayesian Methods for Mixtures of Normal Distributions". Thèse de doct. Department of Statistics.

Strid, I. (2010). "Efficient parallelisation of Metropolis–Hastings algorithms using a prefetching approach". In : *Computational Statistics & Data Analysis* 54.11, p. 2814–2835.

Tierney, L. et A. Mira (1998). "Some adaptive Monte Carlo methods for Bayesian inference". In : *Statistics in Medicine* 18, p. 2507–2515.

Tierney, Luke (1998). "A note on Metropolis-Hastings kernels for general state spaces". In : *Ann. Appl. Probab.* 8.1, p. 1–9. issn : 1050-5164.

Wang, X. et D.B. Dunson (2013). "Parallel MCMC via Weierstrass Sampler". In : *arXiv preprint arXiv:1312.4605*.

Chapitre 3

Bayesian Dimension Expansion

Modelling nonstationary spatial processes

This is joint work with Nicolas Chopin and Luke Bornn.

3.1 Introduction

Hierarchical spatial models for environmental processes are receiving more and more interest as geostatistical data availability grows thanks to modern Geographical Information and Global Positioning Systems. The majority of these models make use of the extremely flexible framework of Gaussian processes to carry out estimation of interesting quantities but often still rely on the simplifying assumption of stationarity.

The aim of this work is to introduce a general and intuitive technique that explicitly model nonstationarity, allowing us to account for hidden environmental effects whether they correspond to a physical phenomenon or not, while retaining the simplicity of familiar stationary Gaussian processes.

A number of existing methods deal already with modelling nonstationarity and usually fall into one of two categories : non-trivial convolution of locally stationary processes or 'image–warping' techniques. In the former assumes that when observing the process under study *locally* the effect of nonstationarity is negligible and hence a local stationary model is used. These local processes are later 'convoluted' (in a general sense) toward the global nonstationary process. The strand of literature known as process–convolution, originated in Higdon, 1998 ; Higdon, Swall et Kern, 1999 shows in fact that some nonstationary Gaussian processes can be represented by as the convolution of local kernels via Brownian motions even when the kernel are let to vary spatially to model the nonstationarity ; Xia et Gelfand, 2005 ; Paciorek, 2007 later extended the idea to accommodate a larger class of processes. Similar in conception and principles are low-rank splines (Ruppert, Wand et Carroll, 2003 ; Lin et al., 2000) and more recently the predictive process approach (Banerjee et al., 2008 ; Finley et al., 2009 ; Eidsvik et al., 2012). The latter comprehend all the derivations of the work of Sampson et Guttorp, 1992, whose idea is to deform the geography such that the observed process looks stationary in the resulting space. Multi–dimensional scaling methods are generally used to define the deformed locations and this plate splines generate the map between the original and the war-

ped space. This idea was later extended to the Bayesian framework by Damian, Sampson et Guttorp, 2001 ; Schmidt et O'Hagan, 2003.

Following the results in Perrin et Meiring, 2003 and Perrin et Schlather, 2007, namely the existence of a higher dimensional stationary representation for any non-stationary field, Bornn, Shaddick et Zidek, 2012 developed instead an optimization method called Dimension Expansion (DE) to model a non-stationary process $Y(\mathbf{X})$. The expanded (latent) dimensions $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_s)$ are learned such that the process $Y([\mathbf{X}, \mathbf{Z}])$, where $[\mathbf{X}, \mathbf{Z}]$ represent the concatenation between observed and latent locations, exhibit stationarity ; to be precise $\mathbf{Z}$ are found so that the theoretical assumed parametric variogram has minimum distance from the empirical one in the higher-dimensional space.

Without a proper inferential framework though the method lacks a way to observe the *uncertainty* of the inferred latent space and properly *estimate* the parameters involved in modelling the covariance structure ; this is the focus of this work.

While similar to 'image–warping' at a first glance, this method differ fundamentally in that the original space is retained and not necessarily warped, but we add flexibility through the extra dimensions. Moreover it is not susceptible to one of the major drawback of image–warping methods, which is the possibility of folding the space, that result in a non-injective transformation between original and warped manifold. When this occur two different locations overlap and they become essentially repetition of the process, whose variance is controlled only by independent measurement error rather than being, more logically, heavily correlated.

3.2 The Bayesian dimension expansion model

Consider a process $\{R(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}$ is an observed, potentially non-stationary, process. Assume that we can decompose R as

$$R(\mathbf{X}) = Y(\mathbf{X}) + \mu(\mathbf{X}) + \varepsilon(\mathbf{X}) \tag{3.1}$$

where $\mu(\mathbf{X})$ is a mean function, that could depends on some covariates as well, and $\varepsilon(\mathbf{X})$ is an independent measurement error process, sometimes called *the nugget*, independent from $Y(\mathbf{X})$ which captures the spacial association of the process R and is the main focus of inference in this work. Modelling R is usually a trivial extension of modelling Y as the mean function μ is usually assumed to be a deterministic function of the spacial locations $\mathbf{X}$ and eventually some covariates, often times estimated non-parametrically (see Rue, Martino et Chopin, 2009 for example). $\varepsilon(\mathbf{X})$ is instead assumed to have a Gaussian distribution with diagonal covariance. Sometimes the nugget might be modelled as a smooth process itself, allowing for spatially–dependent measurement errors, which is just a slight increase in parameter dimension.

Assume now that $\{Y(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}, \quad \mathcal{S} \subseteq \mathbb{R}^d$ is a zero-mean univariate Gaussian process with Covariance function $\Sigma_{\theta_y}(h)$, h being the distance between two points. Extensions to multivariate Gaussian processes for Y exists and easily apply to the following at the cost of an increased notation burden. We refer the reader to (e.g.) Genton et Kleiber, 2015 and the relative discussion for an introduction on the

cross–covariance operator, which extends the covariance function, that allows for such a generalisation.

The stationary structure above (Cressie, 1993) can be unreasonable as it explicitly assumes that there is no association between nearby location, which is often unrealistic. In order to model this *nonstationarity*, in the form of spacial association, we will expand the locations by adding a latent process $\mathbf{Z} \in \mathcal{Q}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ such that $Y([\mathbf{X}, \mathbf{Z}])$, now observed on a subset of $\mathbb{R}^{p+d}$, exhibit stationarity.

To sensibly infer the latent process as a smooth function of the original space we will elicitate a prior on $\mathbf{Z}$ such that $\{\mathbf{Z}_i(\mathbf{X}), \mathbf{X} \in \mathcal{S}\}_{i=1,\dots,p}$ is again as a zero-mean univariate Gaussian process with Covariance function $\Sigma_{\theta_z}(h)$. Assuming prior knowledge on the correlation within such a multivariate latent process seems unrealistic and hence once again we will simply rely on its one dimensional marginals.

To retain some flexibility without any too strong assumption on the smoothness of the processes we will assume here on Matérn Covariance functions for both Y and $\mathbf{Z}$, and thus for $\theta = (\sigma^2, \phi)$ (both being non-negative real numbers) we will have

$$\Sigma_{\theta,\nu}(h) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}\frac{h}{\phi}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{h}{\phi}\right) \tag{3.2}$$

where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the Bessel function of the second kind.

Note that while in general the function $K_\nu(\cdot)$ is computationally expensive, notable examples of Matérn covariances for fixed ν have simplified expressions, like :

$$\begin{aligned}
\Sigma_{\theta,\nu}(h) &= \sigma^2 \exp\left(-\frac{h}{\phi}\right), \qquad \nu = \frac{1}{2} \\
\Sigma_{\theta,\nu}(h) &= \sigma^2 \left(1 + \frac{\sqrt{3}h}{\phi}\right) \exp\left(-\frac{\sqrt{3}h}{\phi}\right), \qquad \nu = \frac{3}{2} \\
\Sigma_{\theta,\nu}(h) &= \sigma^2 \left(1 + \frac{\sqrt{5}h}{\phi} + \frac{5h^2}{3\phi^2}\right) \exp\left(-\frac{\sqrt{5}h}{\phi}\right), \qquad \nu = \frac{5}{2} \\
\Sigma_{\theta,\nu}(h) &= \sigma^2 \exp\left(-\frac{h^2}{2\phi^2}\right), \qquad \nu \to \infty
\end{aligned}$$

As ν increase, it is clear to see that the smoothness of the associated process increase as well. We will hence assume $\nu = \frac{1}{2}$ for Y and $\nu \in \left(\frac{1}{2}, \frac{5}{2}\right)$ for $\mathbf{Z}$ as we would like to imply a smooth function for $\mathbf{Z}$ and at the same time accommodate quicker variations on the covariance function for Y.

Note that the smoothness assumption in the prior for $\mathbf{Z}$ is what protect us from producing *random noise* for the latent dimensions, effectively avoiding overfitting thanks to the natural penalisation for increasing extra–dimensions of $\mathcal{Q}$.

To complete the Bayesian formulation we now require the elicitation of the prior for the parameters σ_y^2, σ_z^2, which represent the variability of the processes (i.e. the diagonal elements of the covariance matrix), and ϕ_y, ϕ_z, which control the smoothness of the covariance function or rather the decay rate of the correlation with respect to the distance.

As they all live in the non-negative half line we will usually, without any prior information, assume them to be diffuse $\mathcal{G}amma$ distributions. Defining no constraints

for σ_z^2 though can be dangerous, as the effect of $\mathbf{Z}$ in

$$\Sigma_{\sigma_y^2,\phi_y}(d([\mathbf{X},\mathbf{Z}]))$$

can potentially obscure completely the effect of $\mathbf{X}$ as the correlations depend only on the distance between points. Whatever the latent phenomenon we are implicitly measuring, we would like to assume that it lies in a comparable space to what we are observing in $\mathbf{X}$; hence we usually force the variation of the latent process to be compatible with the variation of $\mathbf{X}$ by adjusting the hyper–parameters $\alpha_{\sigma_z^2}$ and $\beta_{\sigma_z^2}$.

To summarize, the complete Dimension Expansion Bayesian model can be written as :

$$Y|\mathbf{X},\mathbf{Z},\sigma_y^2,\phi_y \sim GP(\mathbf{0},\Sigma_{\sigma_y^2,\phi_y}(d([\mathbf{X},\mathbf{Z}]))) \tag{3.3a}$$
$$\mathbf{Z}_i|\mathbf{X},\sigma_z^2,\phi_z \sim GP(\mathbf{0},\Sigma_{\sigma_z^2,\phi_z}(d(\mathbf{X}))) \tag{3.3b}$$
$$\sigma_y^2 \sim \Gamma(\alpha_{\sigma_y^2},\beta_{\sigma_y^2}),\ \phi_y \sim \Gamma(\alpha_{\phi_y},\beta_{\phi_y}) \tag{3.3c}$$
$$\sigma_z^2 \sim \Gamma(\alpha_{\sigma_z^2},\beta_{\sigma_z^2}),\ \phi_z \sim \Gamma(\alpha_{\phi_z},\beta_{\phi_z}) \tag{3.3d}$$

where $d(\cdot)$ is the Euclidean distance function and $\alpha_{\sigma_y^2},\beta_{\sigma_y^2},\alpha_{\sigma_z^2},\beta_{\sigma_z^2},\alpha_{\phi_y},\beta_{\phi_y},\alpha_{\phi_z},\beta_{\phi_z}$ are all positive real numbers.

3.2.1 The dimension of the latent space p

The dimension p, in $\mathbf{Z} \in \mathcal{Q}$, $\mathcal{Q} \subset \mathbb{R}^p$, was not mentioned until now.

In situations where the latent space represents some unaccounted for covariates, p is generally unknown and even if we are modelling some real world phenomena we might not know how many of them we need to consider.

By specifying a prior $\pi(p)$, moreover, the baseline model which correspond to *stationarity* hence $p = 0$, may have now a positive probability and we would thus be encoding a prior similar in principles to the one proposed in Simpson et al., 2014 where we are guaranteed not to force a nonstationary model for Y when instead the stationarity assumption is supported by the data.

On the other hand, this generality comes with a few sampling drawbacks as some form of Reversible Jump (Green, 1995) has to be included in our procedure. While devising move between the different model sizes in Reversible Jump might not be extremely difficult in our case, as for example $\mathbf{Z} = (\mathbf{Z}_1,\mathbf{Z}_2)$ has equal likelihood with respect to $\mathbf{Z}' = (\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{0})$, this often induces mixing problem as Reversible Jump often cause the acceptance rate to drop significantly.

An attractive property of Sequential Monte Carlo is though the possibility of getting as an algorithm by-product the *model evidence* ; If we define the sequence of target distribution as $\pi_j(\cdot) = \pi(p)\pi(\theta_z)\pi(\theta_y)\pi(\mathbf{Z}|\theta_z,\mathbf{X})\ell(Y|\mathbf{X},\mathbf{Z},\theta_y)^{t_j}$, with t_j decreasing from ∞ to 1, we can then run the algorithm for a fixed p multiple times (once for every p with $\pi(p) > 0$) and finally, thanks to Bayes factors, decide for the preferred latent process dimension and formally test for nonstationarity.

While potentially a very expensive procedure, it has to be noted that seldom we will need more than a few extra-dimensions since as suggested by both Perrin et Meiring, 2003 ; Perrin et Schlather, 2007 and experimental results in Bornn, Shaddick et Zidek, 2012, $\mathcal{Q}$ is likely to be comparable in size with $\mathcal{S}$.

This is the strategy we will adopt here on in this work, but we would still like to mention that, when unable to run the procedure for multiple values of p, one could

obtain a prior that avoid 'overfitting' toward nonstationarity and satisfy the principles in Simpson et al., 2014 by substituting σ_z^2 with its inverse τ_z, that represent the prior precision, and eliciting a prior whose moments are undefined so that very high (potentially *infinite*) values are possible. For a similar situation Simpson et al., 2014 advice on a type-2 Gumbel distribution. This would reduce the process to a flat surface, which won't effectively extend the manifold where Y lies and the process would then be defined on its original space. This causes computational problems however in our case as ϕ_z would then become unidentifiable, so care is advised in analysing the output of such a method.

3.3 Dimension Expansion Sampler

There are a few difficulties in sampling from the above model, and $\mathbf{Z}$ in particular. First and foremost both the likelihood and the posterior for $\mathbf{Z}$ are invariants to some *isometric* transformations, and hence the model is not completely identifiable, and secondly the elements within each extra dimension $\mathbf{Z}_i$ are, by design, very highly correlated and hence sampling can be tricky.

The first point is actually less critical then expected as we are mostly interested in inferring the Covariance matrix or function $\Sigma_{\sigma_y^2,\phi_y}$ of the observed process, rather than $\mathbf{Z}$ itself, which is left untroubled by this identifiability problem; think for example of flip around zero for a one dimensional latent process, or just swapping the indexes of the marginals in a bivariate $\mathbf{Z}$; in both cases the distances between points are left unchanged.

This could still affect the mixing of an MCMC chain though, so particular care must be made in choosing the sampler; while SMC is inherently more robust to this particular problem, we might need to implement a strategy similar to the one explained in Nobile, 1998 if we were to sample from a MCMC procedure, where we would periodically propose a move to an equal-posterior probability point in order to properly explore all the space $\mathcal{Q}$; note though that identifying the transformations needed to carry out such a strategy becomes harder and harder as p increases.

If the latent field express a real-life phenomenon, e.g. the missing altitude in some environmental data, we could however be interested in recovering a posterior estimate of the $\mathbf{Z}$ process; to achieve that we potentially just need at the end of the sampling strategy to perform a post-processing step in order to choose only one particular *shape* for the latent processes and superimposing all the samples to that by applying the above posterior-invariant transformations.

Procrustes analysis tools (Dryden et Mardia, 1998) may also be used to give some powerful insights on the interpretation of the latent process, at the cost of sacrificing the accuracy of the estimates uncertainty; not all *isometric* transformation that a general procrustes routine might try to apply, like translation, are in fact posterior-invariant (even though they are likelihood-invariant).

An example on how the posterior uncertainty is affected by procrustes can be seen in Figure 3.1.

To tackle the second problem instead a plethora of solutions are available in the literature, from over-relaxation (Neal, 1995) to Hamiltonian Monte Carlo (Neal, 2012; Girolami et Calderhead, 2011). Most of them need to be tuned by the user though and, given that environmental problems may vary a lot in their formulation and we want our method to be general enough, we opted for another recent

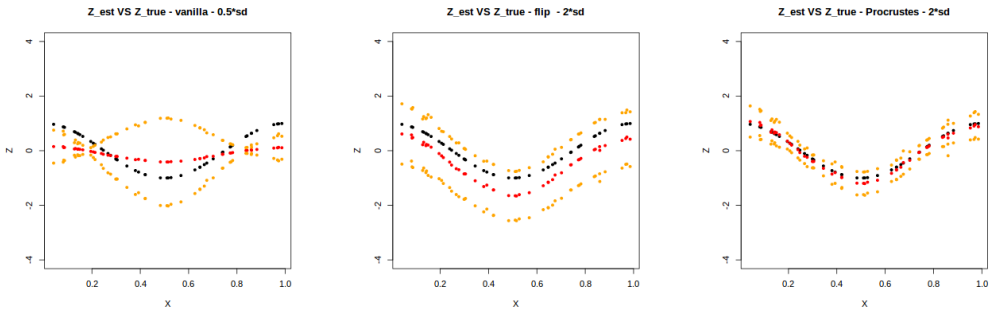


FIGURE 3.1: Unprocessed output (left), Posterior-equivalent postprocessing (centre) and Procrustes postprocessing (right) on a 1-dimensional latent process. In black the real latent process, in red its posterior expectation and in orange ± 0.5 the posterior standard deviation (left) or $\pm$ 2 the posterior standard deviation (centre and right).

procedure that takes advantage of both over-relaxation and slice sampling (NEAL, 2003), proposed to solve this issue specifically for Gaussian processes, which is the Elliptical Slice Sampler from MURRAY, PRESCOTT ADAMS et MACKAY, 2010. Note moreover that while in principle we can't control the number of likelihood computation that will occur in the Elliptical Slice Sampler, we have observed empirically that even at low temperature we don't perform more than a dozen ; in the HMC literature though, even for adaptive methods like HOFFMAN et GELMAN, 2014 ; CARPENTER et al., 2015, hundreds of leap-frog steps are common and this would result overall in a more costly procedure. Outside Gaussian processes thought the performances of Elliptical Slice sampling have not yet been explored thoughtfully (NISHIHARA, MURRAY et ADAMS, 2012) so HMC might there become a more appealing contender.

Sampling from $\sigma_y^2, \phi_y, \sigma_z^2$ and ϕ_z (given the latent process) is done jointly by using a multivariate Gaussian random walk Metropolis-Hastings in the logarithmic space. This is pretty standard in the literature and it does not need further justifications.

The preferred sampler is thus a Sequential Monte Carlo algorithm, with the sequence of distribution defined as sequence starting from the prior and terminating on the posterior distribution via *tempering* of the likelihood, that runs for a fixed p and update the other parameters via partial Gibbs-type steps using *Elliptical Slice Sampling* for $\mathbf{Z}$ given the rest and Metropolis–Hastings with *joint log-normal proposal* for the other parameters given $\mathbf{Z}$. We *repeat the procedure* for every p with $\pi(p) > 0$ and finally, thanks to Bayes factors, decide for the preferred latent process dimension.

As we have chosen to work with an SMC algorithm, we specify the tempering schedule for the sequence of targets in an adaptive way, by selecting the next temperature so that the conditional Effective Sample Size (cESS defined in ZHOU, JOHANSEN et ASTON, 2013), defined as

$$cESS = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}, \tag{3.4}$$

with $W_j^{(i)}$ and $w_j^{(i)}$ the normalised and unnormalised weights respectively at iteration j for particle i, drops by a specific percentage. Intuitively cESS consider how good

a proposal the current particle system represent for the estimation of expectations under the next target.

Similarly we define the Effective Sample Size (ESS) as

$$ESS = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}, \tag{3.5}$$

which corresponds to a measure of variation of the current normalized importance weights, and we resample, via Residual Resampling, if the particle system falls below a chosen threshold. The cESS in (3.4) is equivalent to (3.5) if resampling is performed at each iteration.

The sampler is here presented in algorithm 3.1.

Algorithm 3.1 SMC Sampler for the Bayesian Dimension Expansion model

For a chosen p, set $t = \infty$, $j = 0$;
for $i \in 1, \dots, N$ **do**
  $\theta_0^{(i)} \sim \pi(\cdot)$; $Z_0^{(i)} \sim f_0(\cdot)$; $W_0^{(i)} = \frac{1}{N}$.
end for
while $t > 1$ **do**
  $j \leftarrow j + 1$; Compute t^* such that $cESS_{t^*} = \rho \times cESS_t$;
  for all i **do**
    – Compute $W_j^{(i)} \propto W_{j-1}^{(i)} \times \frac{\ell(Y|\theta_{j-1}^{(i)}, Z_{j-1}^{(i)})^{1/t^*}}{\ell(Y|\theta_{j-1}^{(i)}, Z_{j-1}^{(i)})^{1/t}}$;
    – $(\theta_j, Z_j, W_j)^{(i)} \leftarrow (\theta_{j-1}, Z_{j-1}, W_j)^{(i)}$; $t \leftarrow t^*$;
  end for
  Compute $e = ESS(t)$ and resample if $e < (\psi \times N)$;
  for $i \in 1, \dots, N$ **do**
    – $\theta_j^{(i)} \sim K_\theta^{(j)}(\cdot|\theta_j^{(i)})$;
    – $Z_j^{(i)} \sim K_{Ell}(\cdot|Z_j^{(i)})$;
  end for
end while

where ρ and ψ are positive real constants with value in $(0, 1)$ that govern respectively the temperature decrease and the threshold for the resampling step; π is the prior for $\theta = (\sigma_y^2, \phi_y, \sigma_z^2, \phi_z)$, f_0 is the prior for $\mathbf{Z}$ and $\ell(Y|\theta, Z)$ the likelihood.

K_{Ell} and K_θ are of course the above described Elliptical slice sampler and Metropolis–Hastings Markov kernels to perturb the particles. Notice moreover how we noted $K_\theta^{(j)}$ as dependent on the iteration index j; SMC allows us in fact to adapt the kernels over time and at each iteration we make use of the past set of particles to tune the covariance matrix of the log–normal proposal.

As a last note, the computational cost per iteration is $\mathcal{O}(N \times s^3)$, which can be very high if s is only moderately large. The next section will describe two modelling strategies to overcome this problem in order for our method to scale up to reasonable dimensions for the modern days' applications.

This algorithm was is implemented in C++ and makes use of Armadillo (Sanderson, 2010) for the linear algebra, OpenMP (`http://omp.org`) to make use of additional available CPUs, NVBLAS when available to deflect burdensome matrix computations toward the GPUs (`http://docs.nvidia.com/cuda/nvblas/`) and has an R (R Core Team, 2015) interface thanks to the Rcpp package (Eddelbuettel et François, 2011).

The code is available for testing at `https://bitbucket.org/marcobanterle/DE/`.

3.4 Computational Challenges

As usual with Gaussian processes, our approach scales poorly with the number of observed locations, as the likelihood computation needs $\mathcal{O}(n \times s^3)$ operations and this becomes problematic for even a moderate number of locations. Various low-rank approximations to covariance matrices have been proposed in the literature to overcome this problem (Smola et Bartlett, 2001 ; Seeger, Williams et Lawrence, 2003 ; Schwaighofer et Tresp, 2002 ; Quinonero-Candela et Rasmussen, 2005 ; Paciorek, 2007 ; Snelson et Ghahramani, 2005), but in the following we will focus in particular two of these techniques ; we will first introduce the work of Banerjee et al., 2008 and Datta et al., 2014, that propose two different type of approximations, and then explain the adaptation needed for them to work in our Dimension Expansion framework.

3.4.1 Nearest–Neighbour Gaussian processes approximation

Let $W = (w_1, \ldots, w_s)$ be the set of location where we observe a Gaussian process $Y(w) \sim GP(0, C_\theta(w, \cdot))$, where $w \in \mathcal{W} \subset \mathbb{R}^d$ and assume for ease of explanation that $Y(w) \in \mathcal{Y} \subset \mathbb{R}$, in which case C_θ is a covariance function between two locations in the space that depends on parameter $\theta \in \Theta$. The extension via cross-covariance functions to multivariate spacial processes is just slightly more burdensome in notation, albeit much more complicated computationally and will not be explored in this work.

Now, if Y is observed on the set W we can model its joint distributions as $Y_W \sim \mathcal{N}(0, C_{W,\theta})$ where $C_{W,\theta}$ is a positive definite $s \times s$ matrix with element i, j being $C_\theta(w_i, w_j)$. When s is large the inferential procedure becomes burdensome (or overall unfeasible on a time constraint) due to the need to compute determinant and inverse of the $C_{W,\theta}$ matrix, both of which require $\sim \mathcal{O}(s^3)$ operations.

In order to introduce this first computationally efficient model approximation let us rewrite the joint density of Y_W as a chain-product of full conditional densities

$$p(Y_W) = p(Y(w_1)) \times p(Y(w_2)|Y(w_1)) \times \ldots \times p(Y(w_s)|Y(w_{s-1}), \ldots, Y(w_1)). \quad (3.6)$$

Datta et al., 2014 proposed to reduce the conditioning sets to have *at most* m elements, $m \ll s$, so that effectively the covariance structure becomes sparse and we are able to compute (3.6) using only $m \times m$ matrices in the product. More precisely, they propose to construct for every w_k a conditioning set $N(w_k) \subset W \setminus w_k$ so that

we can approximate (3.6) as

$$\tilde{p}(Y_W) = \prod_{k=1}^{s} p(Y(w_k)|Y_{N(w_k)}) \tag{3.7}$$

with $Y_{N(w_k)} = \{Y(w_i)\ ;\ w_i \in N(w_k)\}$.

If we now define $N_W = \{N(w_k)\ ;\ k = 1, \ldots, s\}$, the collection of all the conditioning sets on the observed locations, the pair (W, N_W) defines respectively the set of nodes and the set of edges of a directed graph G. Note that $N(w_k)$ defines the set of *directed neighbours* of w_k in the graph, hence the name of the procedure, even when the sets are not explicitly populated by neighbours in the original space.

Datta et al., 2014 show that in our case, when $p(Y_W)$ is Gaussian, if the factorisation produces a valid joint density, each factor in the decomposition follows again a normal density and that (3.7) is a multivariate Gaussian density that has a sparse covariance matrix $\tilde{C}_{W,\theta'}$ with at most $sm(m+1)/2$ non-zero elements.

In order to be sure that the decomposition in (3.7) is a proper joint density we must ensure acyclicity of the associated graph G. We thus restrict any conditioning set $N(w_k)$ to be a subset of $\{w_1, \ldots, w_{k-1}\}$ only, condition sufficient for our factorisation to yields a valid probability density. To explicitly specify the sets $N(w_k)$ for $k = 1, \ldots, s$ we need to assume an ordering of the locations first and subsequently a rule to populate them.

First, as in spatial statistics locations have no intrinsic order, we are free to rearrange W as needed, say as W' which is made up by the same elements but potentially in a different order. As for how to populate the neighbours set for a given location w'_k Several proposal have been made in the literature (Stroud, Stein et Lysen, 2014; Stein, Chi et Welty, 2004; Gramacy et Apley, 2013) but we are going to follow the simple strategy proposed by Vecchia, 1988 and followed by Datta et al., 2014 as well, to choose the elements in $N(w'_k)$ as the m_k nearest-neighbours in euclidean distance among $\{w'_1, \ldots, w'_{k-1}\}$, with $m_k = \min\{m, k-1\}$.

Finally from (3.7) we can easily see how each of the factor involves at most an $m \times m$ matrix, and hence we can compute $\tilde{p}(Y_W)$ in just $\mathcal{O}(s \times m^3)$ operations, which is a massive improvement from $\mathcal{O}(s^3)$ as $m \ll s$; there is no need in fact to even store any $s \times s$ matrix.

Nearest–Neighbour Dimension Expansion processes

A few complications arise in using this procedures in combination with Dimension Expansion. In order to retain the computational advantages gained in the likelihood computation, we will have to approximate the prior process for the latent locations in the same way, or rather to assign to $\mathbf{Z}$ a sparse prior, by defining the sets $N_W^{(0)}$ that correspond to the nearest–neighbours for the latent processes $\mathbf{Z}_k(\mathbf{X})$, for $k = 1, \ldots, p$, so that the prior density can be approximated via

$$\tilde{p}(\mathbf{Z}_k|\mathbf{X}, \theta_{\mathbf{Z}}) = \prod_{j=1}^{s} p(\mathbf{Z}_k(x_j)|\mathbf{Z}_{k,N^{(0)}(x_k)}) \qquad k = 1, \ldots, p. \tag{3.8}$$

Moreover the very nature of Dimension Expansion makes so that for Y_W, with $W = [\mathbf{X}, \mathbf{Z}]$, the euclidian neighbourhood of each location is changing with $\mathbf{Z}$ and

would thus probably be sensible to consider this when constructing the sets N_W in (3.7).

Datta et al., 2014 report that the order in which we consider the locations and their neighbours seems to matter relatively little for the resulting inference and at the same time we do not expect Dimension Expansion to completely change the topography of the space, but rather to push points father apart or closer together *locally*. This suggests a first solution, which is to take $N_W = N_W^{(0)}$, i.e. consider only the original locations $\mathbf{X}$ to build our approximation.

This defeats a little the principles behind the procedure though so we introduce two other alternatives to adapt this choice to the changes in $\mathbf{Z}$.

(I) We might consider for each particle $\mathbf{Z}_l$ a different neighbourhood, effectively computing N_W on the warped space $[\mathbf{X}, \mathbf{Z}_l]$. This would implies that we are implicitly targeting our original model but computing the likelihood in an approximate way, as described in (3.7) ;

(II) We might instead consider to start from the same $N_W^{(0)}$ set as a starting point and then update N_W, not necessarily at each iteration, by computing an average distance matrix $\bar{D}_{\mathbf{X},\mathbf{Z}}$, where $\left[\bar{D}_{\mathbf{X},\mathbf{Z}}\right]_{i,j}$ is the distance between two locations i and j in the warped space $[\mathbf{X}, \mathbf{Z}]$ averaged over the current set of particles, and deduce the neighbours sets from $\bar{D}_{\mathbf{X},\mathbf{Z}}$; by correctly re-weighting the particles for the change in model we can guarantee the convergence of the SMC toward the *approximated*[1] model.

If we update the conditioning sets at each iteration for approach(II) both the approaches have similar computational costs, as the most expensive operation is the computation of the N $s \times s$ distance matrices $D_{\mathbf{X},\mathbf{Z}_l}$ that are needed in both the procedures, but (II) becomes cheaper as the updates get more rare since for a given set N_W only a sparse version of $D_{\mathbf{X},\mathbf{Z}_l}$ is needed to compute the likelihood.

We would like to note that we could correct for the inexactness in (I) by formally defining the Nearest-Neighbours structure in the Gaussian process likelihood as a conditioning to a Directed Acyclic Graph (DAG) $G = (W, N_W)$ that would encode the conditional independence (the sparsity) in the covariance explicitly. By assigning a prior to G and devising a proper Markov move to use in the SMC procedure we would indeed exactly target an expanded model from which we could marginalise out G and perform inference as usual. Several work have been published on MCMC methods on likelihood conditioned on DAGs, see for example Eaton et Murphy, 2012. Given the additional complication of constraining the graph space to a fixed level of sparsity (as we want at most m neighbours in each set) and the difficulties related to proposing effective moves in the high-dimensional graph space (which is likely the case in situations where we are applying the sparse model) we won't pursue this strategy here and the development of this exact SMC is deferred to future work as we are for the moment only interested in a computationally feasible, albeit approximate, solution to our problem.

Considering that both methods have similar performances in practice, even though we must be aware that the parameters connected with the two approaches have different interpretations (as they are conditioned on a specific approximation in (II) and *pseudo*-marginal with respect to the approximation/graph in approach

1. note that here we use the word 'approximated' as we expect the sparsity assumption to be restrictive with respect to the real underlying process, but the SMC would generate samples from a valid, albeit sparse, model.

(I)), we will always use (II) when referring to the Nearest-Neighbours process in the following Sections.

Formally, call $W = (w_1, \ldots, w_s)$ the set of current locations under focus (i.e. we could have $W = \mathbf{X}$ for the prior or $W = [\mathbf{X}, \mathbf{Z}']$ for the likelihood connected with some latent process $\mathbf{Z}'$) and respectively W' the reordered locations as described in the previous Section ; we will select the first point as the point with minimal average distance to every other point :

$$w'_1 = \min \left\{ i \; ; \; \frac{1}{s} \sum_{j \neq i} \left[\bar{D}_W \right]_{i,j} \right\} \tag{3.9}$$

where $\bar{D}_W = \frac{1}{N} \sum_{l=1}^{N} D_{W_l}$, W_l being the locations in the (possibly) warped space for particle l, D_{W_l} the distance matrix relative to W_l, and of course $N(w'_1) = \varnothing$. The choice to initialise the decomposition in a clumped area is dictated by the fact that subsequent point will be guaranteed to be conditioned on close-by points rather than possibly very far ones for which the correlation function would be close to null.

Every subsequent point is then chosen as the one closest on average to the already chosen points :

$$w'_k = \min \left\{ i \; ; \; \frac{1}{|\mathcal{P}_k|} \sum_{j \in \mathcal{P}_k} \left[\bar{D}_W \right]_{i,j} \right\} \qquad k = 2, \ldots, s \tag{3.10}$$

where $\mathcal{P}_k$ is the set of all the already chosen indexes at the k^{th} iteration, $|\mathcal{P}_k|$ its cardinality and $N(w'_k)$ is composed by the m_k points closest to w'_k in $\mathcal{P}$, with $m_k = \min(m, k-1)$. Again $N_W = \{N(w'_i) \; ; \; i = 1, \ldots, s\}$.

In the sampler 3.1, after re–computing the ordering of the locations W', we need thus to reweight the particles using simply the ratio of the likelihood computed using respectively the new and the previous order (remember the prior is unchanged as the $\mathcal{X}$ space is not changing).

3.4.2 Predictive Process approximation

As in the previous Section, call $W = (w_1, \ldots, w_s)$ the set of s locations where our process $Y(w) \sim GP(0, C_\theta(w, \cdot))$ is observed ; $w \in \mathcal{W} \subset \mathbb{R}^d$, $Y(w) \in \mathcal{Y} \subset \mathbb{R}$.

Consider now a *reference set* of m knots $W^* = (w_1^*, \ldots, w_m^*)$ where each point w_k^* may or may not belong to the observed set. The Gaussian process specification for $Y(w)$ gives

$$Y^* = [Y(w_k^*)]_{k=1}^m \sim \mathcal{N}_m(0, C_\theta^*), \tag{3.11}$$

with C_θ^* is a $m \times m$ covariance matrix with element i, j defined by $C_\theta(w_i^*, w_j^*)$.

The main idea in Banerjee et al., 2008 is essentially to evaluate the process onto the knots while the rest is approximated via deterministic extrapolation (also known as 'krieging') as in

$$\tilde{Y}(w) = \mathbb{E}[Y(w)|Y^*] = c_\theta(w|W^*)^T {C^*}_\theta^{-1} Y^* \tag{3.12}$$

where $c_\theta(w|W^*)$ is a vector whose elements are $C_\theta(w, w_j^*)$ for $j = 1, \ldots, m$.

It is easy to see in fact that $\tilde{Y}(w)$ follow a Gaussian process itself, with covariance function $\tilde{C}_\theta(w, w') = c_\theta(w|W^*)^T {C^*}_\theta^{-1} c_\theta(w'|W^*)$ and we will refer to it as the *predictive process* obtained from its parent $Y(w)$ and the knots W^*.

Banerjee et al., 2008 propose to directly use the predictive process instead of the complete Y to fit the data; the reduction in dimension is evident in the fact that we would now be dealing only with m random effects and $m \times m$ matrices. While being similar to other low-rank approximations (for example Higdon, 2002; Paciorek, 2007 among other convolution / projection techniques) Banerjee et al., 2008 show that the predictive process is optimal amongst projection methods in that being essentially a conditional expectation it minimises $\mathbb{E}\big[Y(w) - f(Y^*)|Y^*\big]$ over all possible real functions $f(Y^*)$.

Finley et al., 2009; Eidsvik et al., 2012 later acknowledge the over-smoothing that occur by approximating $Y(x)$ with its low-dimension projection $\tilde{Y}(w)$ and correct for it by defining a modified predictive process as

$$\tilde{Y}_\epsilon(w) = \tilde{Y}(w) + \epsilon(w) \tag{3.13}$$

$$\epsilon(w) \overset{iid}{\sim} \mathcal{N}\left(0, C_\theta(w, w) - c_\theta(w|W^*)^T {C^*}_\theta^{-1} c_\theta(w|W^*)\right) \tag{3.14}$$

This way $\mathbb{V}ar\big(\tilde{Y}_\epsilon(w)\big) = \mathbb{V}ar\big(Y(w)\big)$ while retaining the attractive properties of the predictive process since $\mathbb{E}\big[\tilde{Y}_\epsilon(w)\big] = \mathbb{E}\big[\tilde{Y}(w)\big]$.

Extensions to multivariate processes are possible but won't be explored in this work.

Banerjee et al., 2008 proceed then by integrating out Y^* and sample from the full-conditional of the parameters θ. The computational costs is reduced thanks to the Sherman-Woodbury-Morrison formula that allows to efficiently compute both the inverse matrices and the determinants needed by using only operations on the $m \times m$ matrices defined through the knots; the resulting cost is once again $\mathcal{O}(s \times m^3)$.

Predictive Processes and Dimension Expansion

In the Dimension Expansion model though, as it was the case for Section 3.4.1, we will need to approximate the prior process in a similar way if we want to retain the computational gains as even $\mathbf{Z}$ is defined as being (un-)observed on s location, warping thus the original space to a dimension where Y is stationary.

This is achieved by simply assigning a prior now to $\mathbf{Z}_k^* \sim \mathcal{N}_m\big(0, C^*_{0,\theta_\mathbf{Z}}\big)$ for $k = 1, \ldots, p$ where the elements of $C^*_{0,\theta_\mathbf{Z}}$ are defined as $C_{\theta_\mathbf{Z}}(\mathbf{X}_i^*, \mathbf{X}_j^*)$ for $i, j = 1, \ldots, m$ with $\mathbf{X}^*$ being the knots on the $\mathcal{X}$ space. We could rely on the Matrix-Normal distribution or on a Cross-Covariance function to define a joint distribution over all the latent dimensions $\mathbf{Z}$ at the same time, but we find fruitless the more general notation, especially for the prior distribution, when the latent dimension would probably be supposed independent *a priori* anyway. By interpolating it on the observed locations $\mathbf{X}$ similarly to (3.12) we will then obtain $\tilde{\mathbf{Z}}$.

We can now define a predictive process $\tilde{Y}$ as before by conditioning on its realisation Y^* over the knots $W^* = [\mathbf{X}^*, \mathbf{Z}^*]$ and *predict* it on $\tilde{W} = [\mathbf{X}, \tilde{\mathbf{Z}}]$.

Dimension Expansion moreover complicate the likelihood to the point that integrating out with respect to Y^* is unfeasible and instead we update it using its full conditional distribution, which is simply a multivariate normal in our case (3.11), we will hence have no need for Woodbury type formulae as all the operations simply involve $m \times m$ matrices or diagonal $S \times s$ ones for the variance correction in

Eq. (3.13). This result thus again in a cost of $\mathcal{O}(s \times m^3)$ operations for the likelihood computation using this procedure.

Algorithm 3.1 changes thus a bit as we need to introduce the sampling for Y^* and the prior is shifted over the knots. The pseudo–code can thus be written as in Algorithm 3.2.

Algorithm 3.2 Sampler for the Predictive approximation on the Bayesian DE model

For a chosen p, set $t = \infty$, $j = 0$;
for $i \in 1, \dots, N$ **do**
  $\theta_0^{(i)} \sim \pi(\cdot)$; $Z_0^{*(i)} \sim f_0(\cdot)$; $W_0^{(i)} = \frac{1}{N}$;
  sample $Y_0^{*(i)}$ from (3.11) and obtain $\tilde{\mathbf{Z}}_0^{(i)}$ as in (3.12) .
end for
while $t > 1$ **do**
  $j \leftarrow j + 1$; Compute t^* such that $cESS_{t^*} = \rho \times cESS_t$;
  for all i **do**
    – Compute $W_j^{(i)} \propto W_{j-1}^{(i)} \times \frac{\ell(Y|\theta_{j-1}^{(i)}, \tilde{\mathbf{Z}}_{j-1}^{(i)}, Y_{j-1}^{*(i)})^{1/t^*}}{\ell(Y|\theta_{j-1}^{(i)}, \tilde{\mathbf{Z}}_{j-1}^{(i)}, Y_{j-1}^{*(i)})^{1/t}}$;
    – $(\theta_j, Z_j^*, Y_j^*, W_j)^{(i)} \leftarrow (\theta_{j-1}, Z_{j-1}^*, Y_{j-1}^*, W_j)^{(i)}$; $t \leftarrow t^*$;
  end for
  Compute $e = ESS(t)$ and resample if $e < (\psi \times N)$;
  for $i \in 1, \dots, N$ **do**
    – $\theta_j^{(i)} \sim K_\theta^{(j)}(\cdot|\theta_j^{(i)})$;
    – $Z_j^{*(i)} \sim K_{Ell}(\cdot|Z_j^{*(i)}$;
    – $Y_j^{*(i)}$ from (3.11);
    – $\tilde{\mathbf{Z}}_j^{(i)}$ as in (3.12);
  end for
end while

Note that the interpretation of the parameters θ associated with $\tilde{Y}$ in general changes with respect to their meaning in the parent process Y.

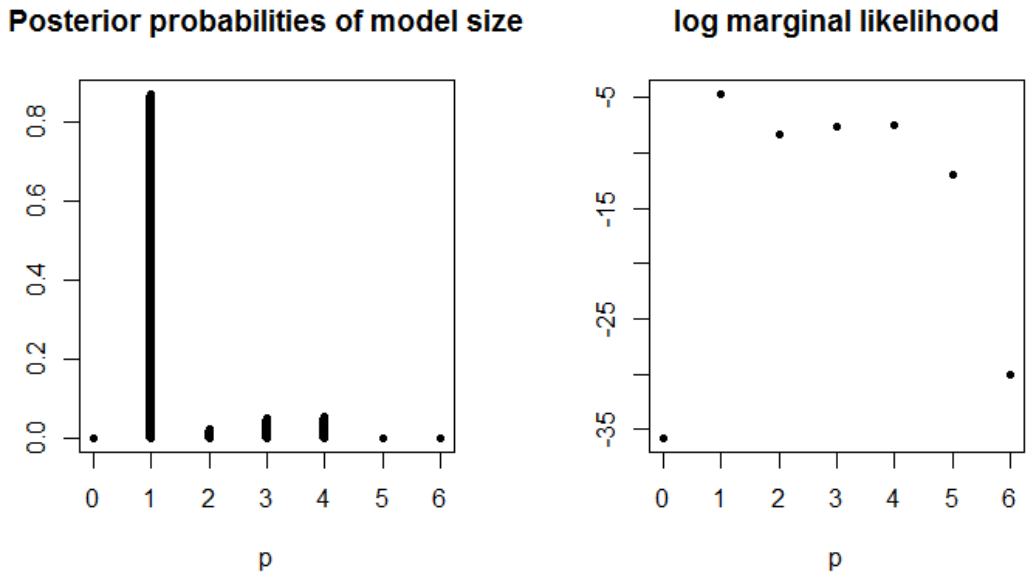
We'd like to conclude this Section on computationally advantageous approximations by noting that nothing prevent us from defining the prior over the latent dimensions $\mathbf{Z}$ as a sparse process like in Section 3.4.1 and then approximate Y as $\tilde{Y}$ like we did in this last Section 3.4.2, but while the prior specification might seem more elegant we don't expect it to make a big difference in practice, while potentially complicate the implementation of the method due to the need to both define knots and neighbours structure at the same time. We want moreover to stress how we think of both methods as computationally feasible alternatives for the complete model (3.3), and hence not the focus of the present work.

3.5 Experimental Results

3.5.1 1–dimensional latent process

We start by testing the procedure on a small example that will help us in illustrating the technique.

FIGURE 3.2: Posterior probabilities for model size p, simulated paraboloid data with n=10, s=25



We defined a grid of 25 points $\mathbf{X} = (X_1, X_2) \in [0,1] \times [0,1]$ and defined a latent process $\mathbf{Z} = 1.2 - (4(X_1 - 0.5)^2 + 4 * (X_2 - 0.5)^2$; see the first panel in Figure 3.3. Following (3.3a) we drew a small sample of $n = 10$ points from a zero-mean Gaussian process with covariance function $\Sigma_y(h) = 5 \exp\left(-\frac{h}{0.5}\right)$, h being the distance between points. We proceeded to run our procedure for $p \in 1, \ldots, 6$ where each SMC was run with $N = 5000$ particles until convergence, when the adaptive temperature ladder reaches 1.

Figure 3.2 shows that even with few data, the posterior for p is in this case very concentrated on the true model and can be correctly picked up by our procedure. Stationarity is strongly rejected and while there is a mild support for $p > 1$ the prior penalisation prevent us from choosing overly–complicated latent structures.

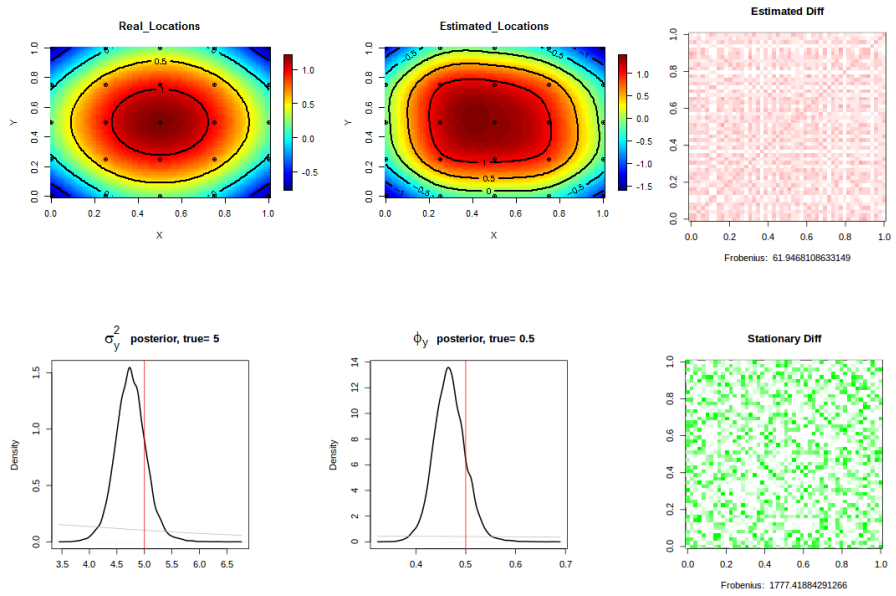
The (procrustes–corrected, see Section 3.3) inferred latent space and posterior densities for θ_y are shown in Figure 3.3. While not perfect remember that we only had 10 data points to work with. Posterior simulation with more observation tend to show a perfect replica of the original latent field and a degenerate posteriorfor p on the true value, reducing the amount of possible discussion.

3.5.2 Solar Radiation Data

HAY, 1983 first presented a set of measurement of solar radiation taken in Canada, in the Vancouver surroundings, that has since become a classic test-bed data set used to study nonstationary models. Out of the $s = 12$ stations one is positioned in a mountain environment but only latitude and longitude are taken into account when considering their location on the space and this cause the data to exhibit nonstationarity. As shown already in BORNN, SHADDICK et ZIDEK, 2012 and other, the station associated with Grouse mountain should be pulled further apart from the other locations to produce a stationary field.

We test this assumption by running out method on $N = 5000$ particles on the

FIGURE 3.3: Real and estimated (posterior expectation) latent dimension $\mathbf{Z}$, posterior distributions for θ_y and Frobenius norm for the difference between the real and estimated or stationary covariance matrix.



data set made up by 4 years worth of measurements for the twelve stations.

First, notice how in Figure 3.4 again stationarity on the original space is clearly rejected, but a one dimensional latent process is enough to make the data stationary in the extended space. It is very interesting as well to note how the estimated univariate $\mathbf{Z}$ (Figure 3.7) closely resemble the elevation contours of the mountains of the area, suggesting in accordance with the literature that the missing altitude plays an important role. In our case is rather estimated as a depression and the 'zero–elevation' point is somewhere in between the sites on the plain and the mountain site, as the prior for the latent field is centred on zero, but we decided not to postprocess the output any more as it let us stress once again how the non–identifiability in the likelihood (and in the posterior to some extent) make necessary particular care when analysing the results.

The result in this case is a significant difference in the estimated covariance (shown as heat–maps in Figure 3.6) most notably for all the elements associated with the last site (row and column 12 in the plot); another point of interest seems to be the second most northern station, which is 'farther', in terms of covariance, from the group of clustered stations on the left that what would appear looking only to the original locations. From Figure 3.6 is thus apparent that the estimated covariance is overall very close to the empirical one, which once again confirms that the nonstationarity is accounted for thanks to our general framework without needing any prior data exploration or analysis, but still retaining good interpretability of the associated parameters and the familiar structure of a stationary process.

We present the posterior distributions for θ_y in Figure 3.5 to confirm that their values are in accordance with the literature (BORNN, SHADDICK et ZIDEK, 2012).

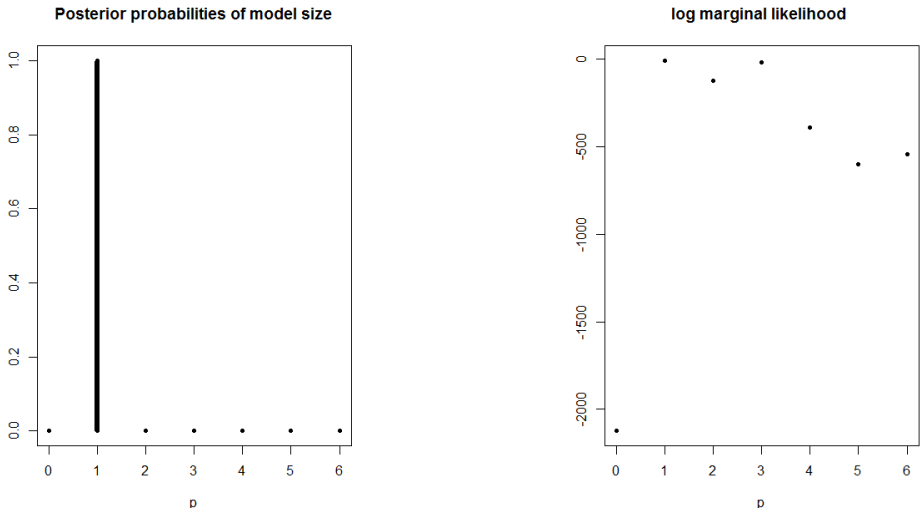


FIGURE 3.4: Posterior probabilities for model size p, solar radiation data (HAY, 1983).

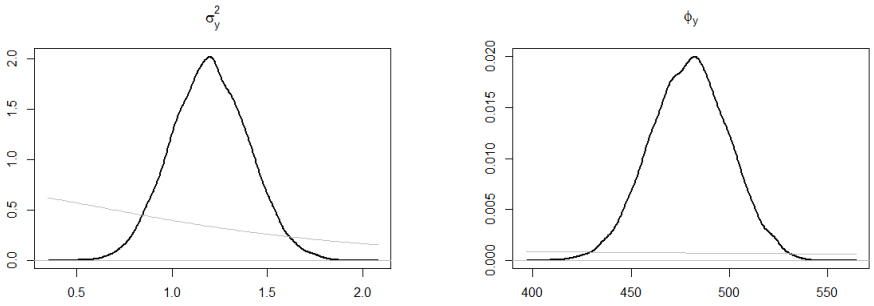


FIGURE 3.5: Posterior distributions for θ_y, solar radiation data.

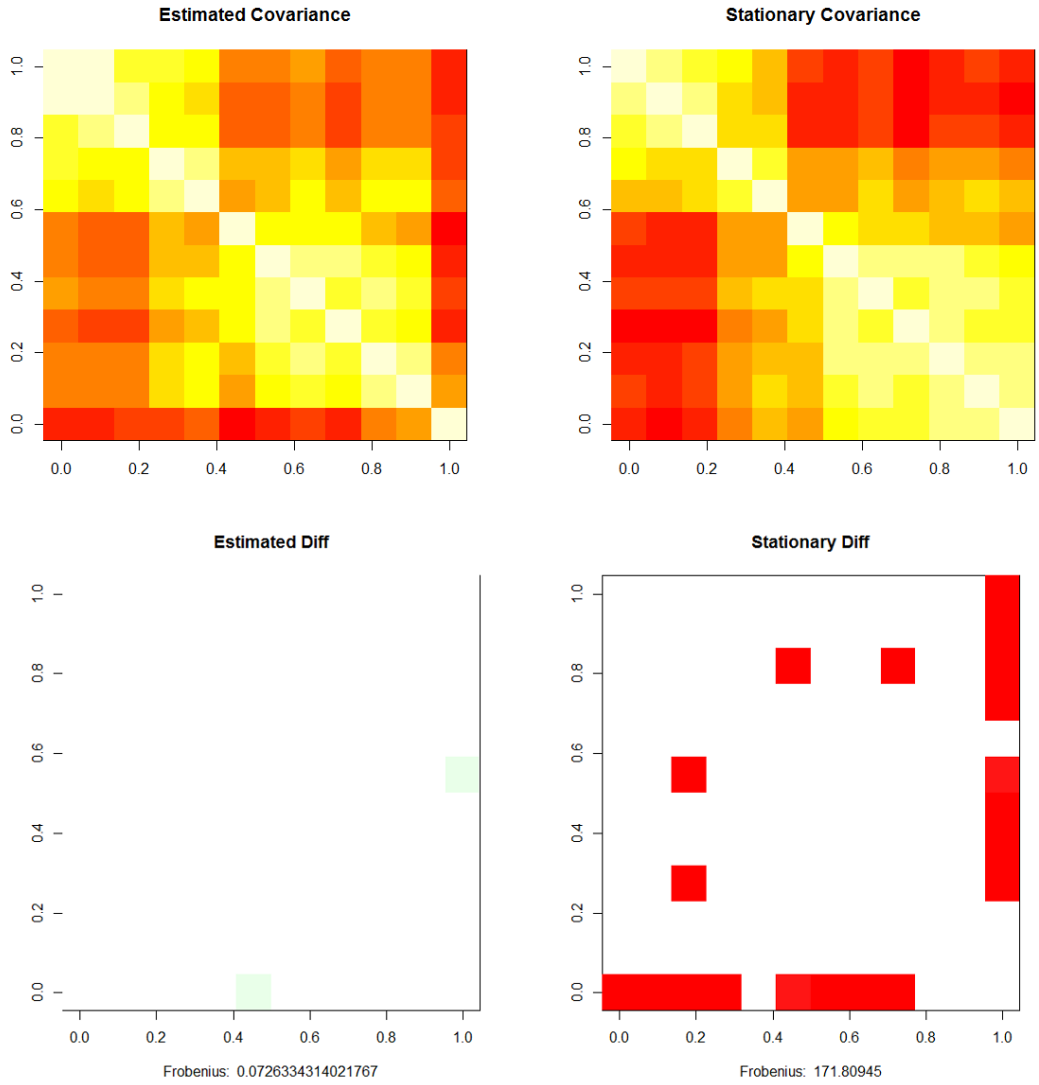


FIGURE 3.6: Top row : Heat maps for the estimated and (estimated for $p = 0$) stationary covariance matrix. Bottom row : Heat maps for the difference between empirical and estimated covariance matrices, nonstationary and stationary respectively (Frobenius norm shown as well). Solar radiation data.

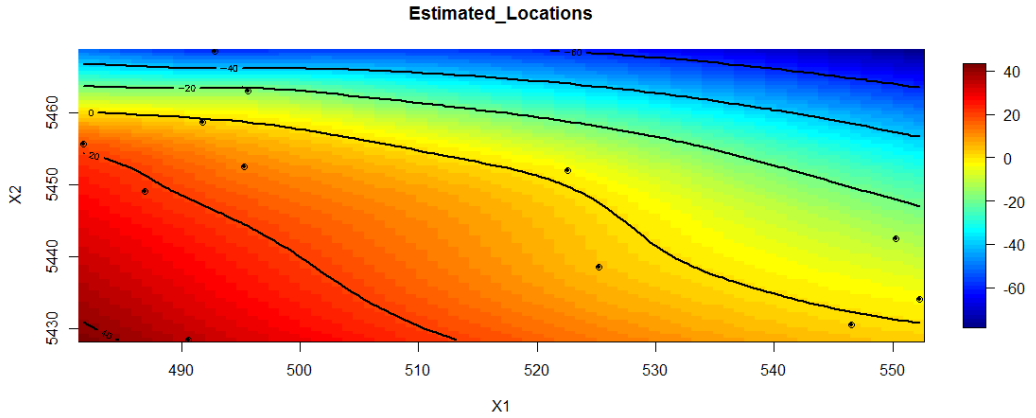


FIGURE 3.7: Estimated latent process, solar radiation data.

3.5.3 Bivariate latent field in high dimensions

We move now onto a more computationally challenging example of a process observed in $s = 250$ locations uniformely distributed in the $[0, 1] \times [0, 1]$ rectangle. We augment the space with $\mathbf{Z} = Z_1, Z_2$, with each margin sampled from a zero-mean Gaussian processes with covariance function $\Sigma_z(h) = 1 \exp\left(-\frac{h}{0.5}\right)$ with h a distance on the (X_1, X_2) space. $n = 100$ observations from (3.3a) are then sampled and $\mathbf{Z}$ is finally hidden.

Running the exact procedure in Algorithm 3.1 for a fixed p takes slightly over 4 days a modern quad core laptop, while in 2–3 hours we can get the results for both the approximate models (2 hours and 26 minutes for the predictive model and 3 hours and 04 minutes for the nearest-neighbours due probably to a less optimised code).

As the procedure is very costly we decided to avoid unnecessary run and decided to implement a *greedy* procedure that sample from each model size $p = 0, 1, 2, \dots$ until the marginal likelihood decreases under a certain threshold depending on the previous run. This is because thanks to the prior overly–complex models should be penalised more and more and hence, once the support from the data does not increase by adding another latent process, we expect the estimated marginal likelihood to drop significantly. Both approximation were then run multiple times ($p = 0, 1, 2, \dots$) and correctly select as the max *a posteriori* model the correct dimension $p = 2$. The exact model is too expensive in this situation and was then just tested for the correct $p = 2$ against the ground truth to validate the findings.

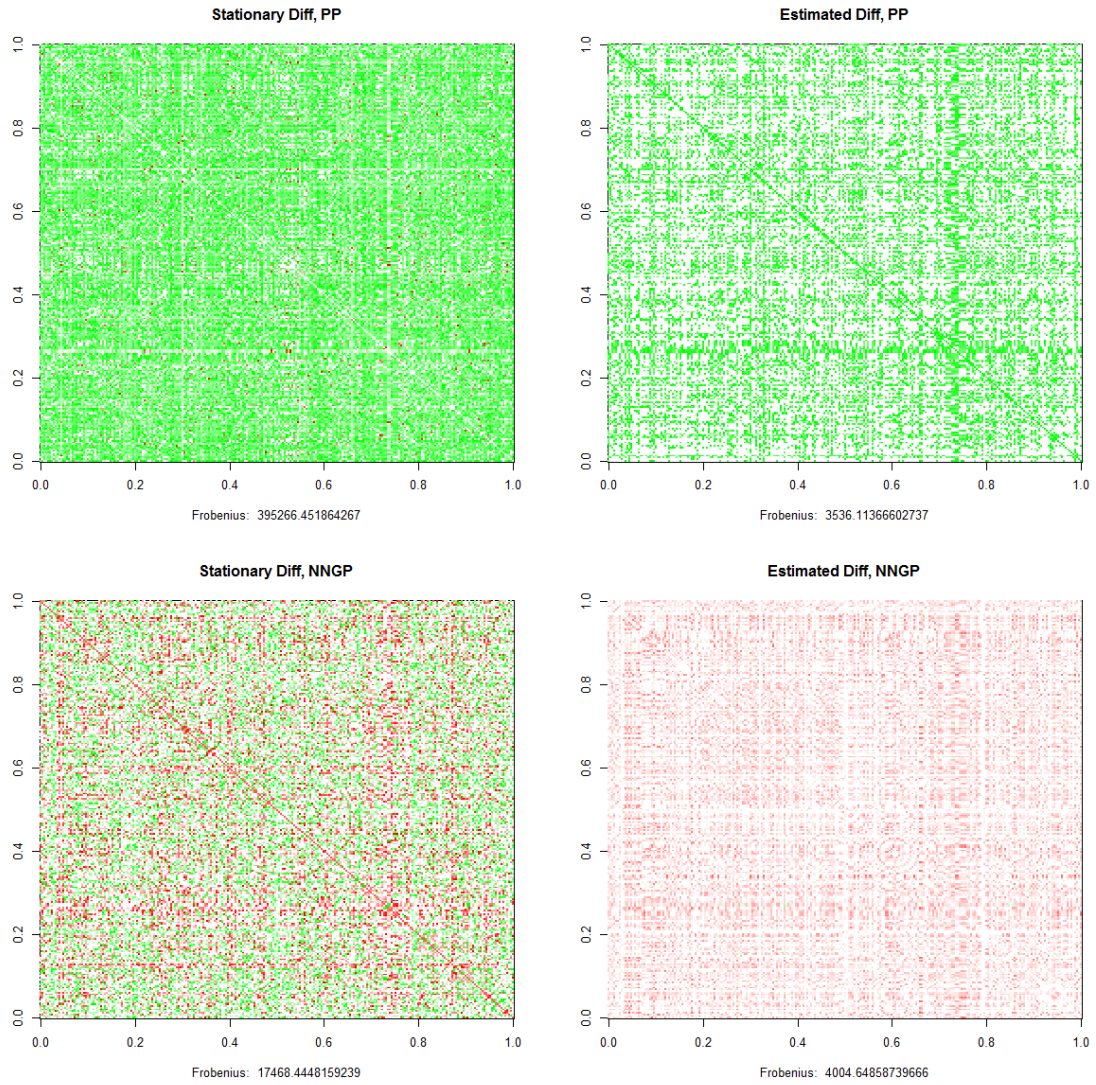


FIGURE 3.8: Heat–maps representing the difference between true Covariance matrix and estimated one produced by Predictive Process (top) and the Nearest–Neighbours (bottom)

Here we present the result for both the techniques as heat–map differences from the ground truth (Figure 3.8), and posterior mean for the latent process (Figures 3.9–3.10). For the predictive process both the posterior average of $\mathbf{Z}^*$ and $\tilde{\mathbf{Z}}$ are shown. Overall, even though the predictive process solution is prone to over–smooth the surfaces, either method produces satisfying results ; note however that the Predictive process approximation consistently produced bad results for the estimated stationary model.

3.6 Conclusion

We have introduced a pure Bayesian version of Dimension Expansion (BORNN, SHADDICK et ZIDEK, 2012), a technique that allows to model nonstationarity in environmental processes by expanding the original observational space through latent effects. This allows for a flexible procedure capable of fitting diverse and challenging data while retaining the simplicity of the stationarity assumption and allowing

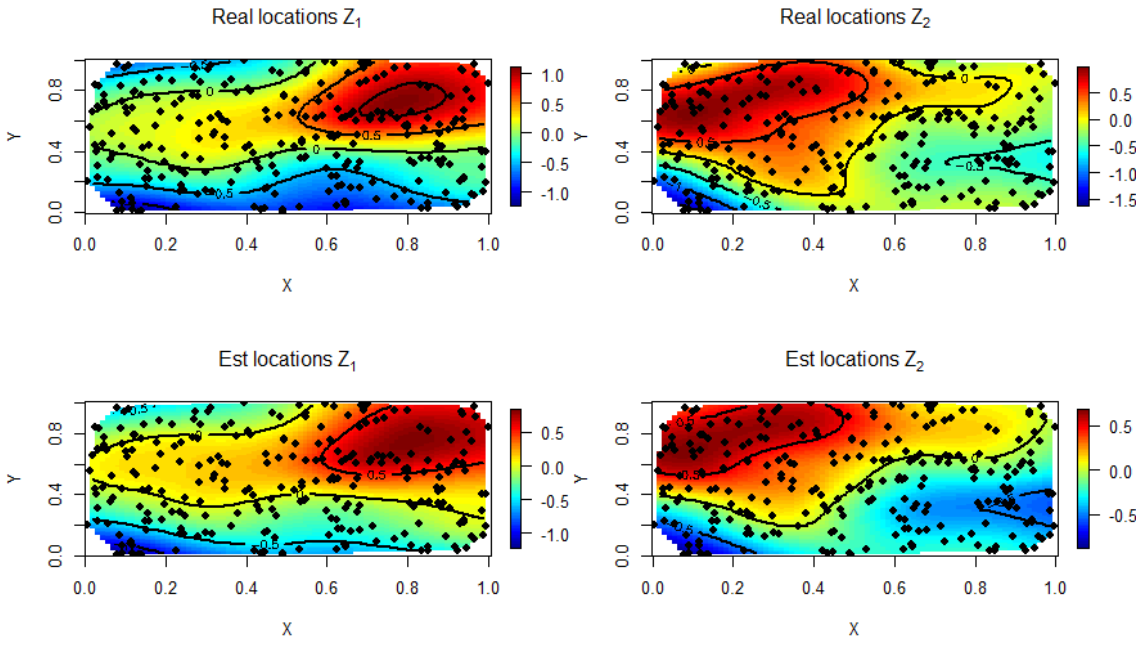


FIGURE 3.9: Results for a 2-dimensional latent process estimated through *nearest-neighbours approximation* where $Z_1(X_1, X_2)$ and $Z_2(X_1, X_2)$; we used 10000 particles, adaptive temperature schedule, $n = 100$, $s = 250$.

for the use of classic covariance functions. We have focused moreover on detailing a proper sampler for this model and some computationally cheaper approximate alternatives that allow this method to scale to higher dimension to what currently reported in the Bayesian literature.

The examples on both simulated and real data shows that the model is able to adapt to a variety of situation without the need for a extensive analysis on the causes of the observed nonstationarity. The ability to accommodate staionarity as a special case in more than one way and the intrinsic ability of the Bayesian paradigm to penalise to over–complicate models is another guarantee of the generality of the procedure.

Many extensions of the present work are possible and require possibly little effort, like allowing for a spatially varying *nugget* in the covariance function or simultaneously allowing for modelling of the mean function, possibly on an the same extended space. The extensions to non–Gaussian observed processes or to multidimensional processes, while not necessarily trivial, seem quite possible with little modifications, especially keeping in mind that we are not able to make use of the Gaussian likelihood and its conjugate prior to integrate any of the parameters out anyway (see for example Section 3.4.2), and there is nothing inherently Gaussian in the definition of the method.

Variational Bayes methods and other approximate techniques that do not rely on sampling might also be explored to push higher the upper limit in dimension due to computational feasibility might also be attractive. These and are left for future works.

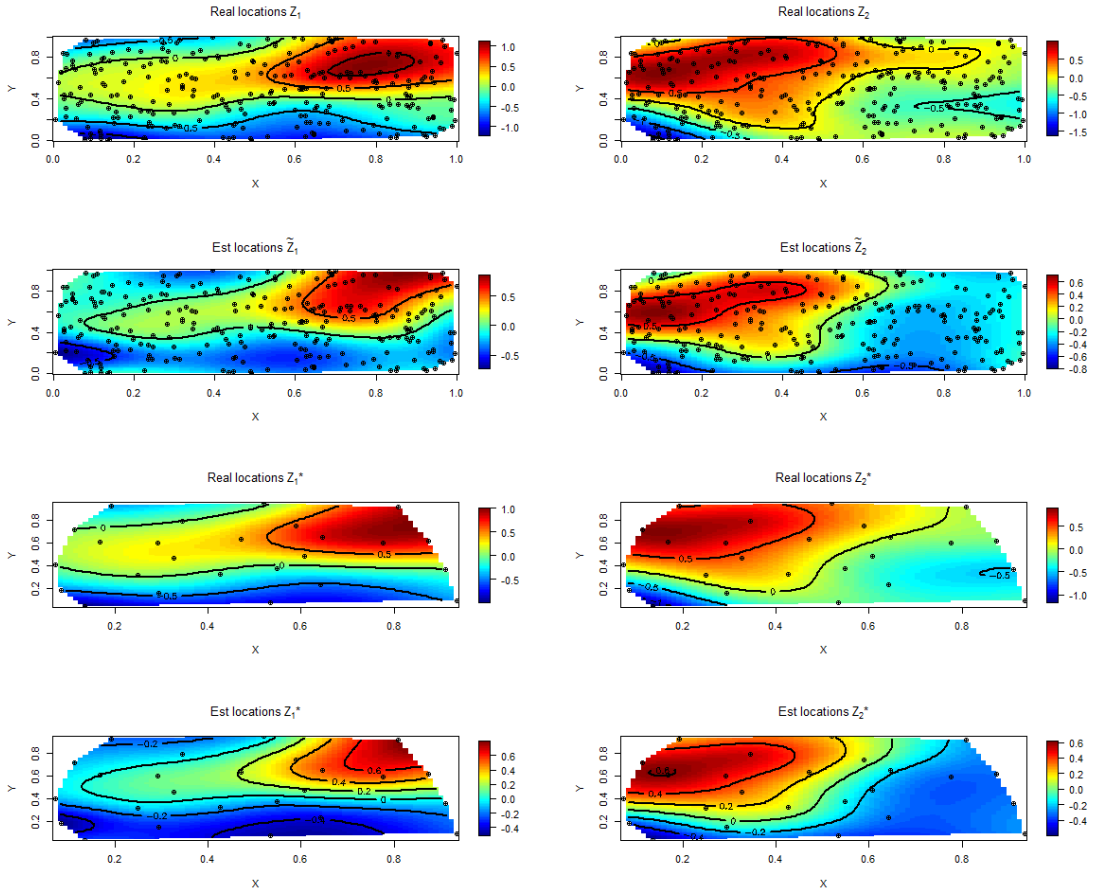


FIGURE 3.10: Results for a 2-dimensional latent process estimated through *predictive processes approximation* where $Z_1(X_1, X_2)$ and $Z_2(X_1, X_2)$; we used 10000 particles, adaptive temperature schedule, $n = 100$, $s = 250$.

References

Banerjee, Sudipto et al. (2008). "Gaussian predictive process models for large spatial data sets". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, p. 825–848.

Bornn, Luke, Gavin Shaddick et James V. Zidek (2012). "Modeling Nonstationary Processes Through Dimension Expansion". In : *Journal of the American Statistical Association* 107.497, p. 281–289. doi : 10.1080/01621459.2011.646919.

Carpenter, Bob et al. (2015). "Stan: a probabilistic programming language". In : *Journal of Statistical Software*.

Cressie, N. (1993). *Spatial Statistics*. New York.

Damian, Doris, Paul D Sampson et Peter Guttorp (2001). "Bayesian estimation of semi-parametric non-stationary spatial covariance structures". In : *Environmetrics* 12.2, p. 161–178.

Datta, A. et al. (2014). "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets". In : *ArXiv e-prints*. arXiv : 1406.7343 [stat.ME].

Dryden, Ian L et Kanti V Mardia (1998). *Statistical shape analysis*. T. 4. J. Wiley Chichester.

Eaton, Daniel et Kevin Murphy (2012). "Bayesian structure learning using dynamic programming and MCMC". In : *arXiv preprint arXiv:1206.5247*.

Eddelbuettel, Dirk et Romain François (2011). "Rcpp: Seamless R and C++ Integration". In : *Journal of Statistical Software* 40.8, p. 1–18.

Eidsvik, Jo et al. (2012). "Approximate Bayesian inference for large spatial datasets using predictive process models". In : *Computational Statistics & Data Analysis* 56.6, p. 1362–1380.

Finley, Andrew O et al. (2009). "Improving the performance of predictive process modeling for large datasets". In : *Computational statistics & data analysis* 53.8, p. 2873–2884.

Genton, Marc G, William Kleiber et al. (2015). "Cross-covariance functions for multivariate geostatistics". In : *Statistical Science* 30.2, p. 147–163.

Girolami, M. et B. Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, p. 123–214.

Gramacy, R. B. et D. W. Apley (2013). "Local Gaussian process approximation for large computer experiments". In : *ArXiv e-prints*. arXiv : 1303.0383 [stat.ME].

Green, Peter J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In : *Biometrika* 82.4, p. 711–732. issn : 0006-3444.

Hay, John E (1983). “Solar energy system design: The impact of mesoscale variations in solar radiation”. In : *Atmosphere-Ocean* 21.2, p. 138–157.

Higdon, Dave, J Swall et J Kern (1999). “Non-stationary spatial modeling”. In : *Bayesian statistics* 6.1, p. 761–768.

Higdon, Dave et al. (2002). “Space and space-time modeling using process convolutions”. In : *Quantitative methods for current environmental issues* 3754.

Higdon, David (1998). “A process-convolution approach to modelling temperatures in the North Atlantic Ocean”. In : *Environmental and Ecological Statistics* 5.2, p. 173–190.

Hoffman, Matthew D et Andrew Gelman (2014). “The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo”. In : *The Journal of Machine Learning Research* 15.1, p. 1593–1623.

Lin, Xiwu et al. (2000). “Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV”. In : *Annals of Statistics*, p. 1570–1600.

Murray, I., R. Prescott Adams et D. J. C. MacKay (2010). “Elliptical slice sampling”. In : *ArXiv e-prints*. arXiv : `1001.0175 [stat.CO]`.

Neal, R.M. (1995). *Suppressing random walks in MCMC using ordered overrelaxation*. Rapp. tech. University of Toronto, Canada : Dept. of Statistics.

— (2003). “Slice sampling (with discussion)”. In : 31, p. 705–767.

— (2012). “MCMC using Hamiltonian dynamics”. In : *arXiv preprint arXiv:1206.1901*.

Nishihara, Robert, Iain Murray et Ryan P Adams (2012). “Generalizing elliptical slice sampling for parallel mcmc”. In : *Neural Information Processing Systems (NIPS), Big Learning Workshop on Algorithms, Systems, and Tools for Learning at Scale*. T. 3. 3, p. 4.

Nobile, A. (1998). “A hybrid Markov chain for the Bayesian analysis of the multinomial probit model”. In : *Statistics and Computing* 8, p. 229–242.

Paciorek, Christopher J (2007). “Computational techniques for spatial logistic regression with large data sets”. In : *Computational statistics & data analysis* 51.8, p. 3631–3653.

Perrin, Olivier et Wendy Meiring (2003). “Nonstationarity in Rn is second-order stationarity in R2n”. In : *Journal of applied probability*, p. 815–820.

Perrin, Olivier et Martin Schlather (2007). “Can any multivariate gaussian vector be interpreted as a sample from a stationary random process?” In : *Statistics & probability letters* 77.9, p. 881–884.

Quinonero-Candela, Joaquin et Carl Edward Rasmussen (2005). “A unifying view of sparse approximate Gaussian process regression”. In : *The Journal of Machine Learning Research* 6, p. 1939–1959.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : `http://www.R-project.org/`.

Rue, H., S. Martino et N. Chopin (2009). “Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations”. In : 71, p. 319–392.

Ruppert, David, Matt P Wand et Raymond J Carroll (2003). *Semiparametric regression*. 12. Cambridge university press.

Sampson, Paul D et Peter Guttorp (1992). "Nonparametric estimation of nonstationary spatial covariance structure". In : *Journal of the American Statistical Association* 87.417, p. 108–119.

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Rapp. tech. NICTA.

Schmidt, Alexandra M et Anthony O'Hagan (2003). "Bayesian inference for nonstationary spatial covariance structure via spatial deformations". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.3, p. 743–758.

Schwaighofer, Anton et Volker Tresp (2002). "Transductive and inductive methods for approximate Gaussian process regression". In : *Advances in Neural Information Processing Systems*, p. 953–960.

Seeger, Matthias, Christopher Williams et Neil Lawrence (2003). "Fast forward selection to speed up sparse Gaussian process regression". In : *Artificial Intelligence and Statistics 9*. EPFL-CONF-161318.

Simpson, D. P. et al. (2014). "Penalising model component complexity: A principled, practical approach to constructing priors". In : *ArXiv e-prints*. arXiv : 1403.4630 [stat.ME].

Smola, Alex J et Peter Bartlett (2001). "Sparse greedy Gaussian process regression". In : *Advances in Neural Information Processing Systems 13*. Citeseer.

Snelson, Edward et Zoubin Ghahramani (2005). "Sparse Gaussian processes using pseudo-inputs". In : *Advances in neural information processing systems*, p. 1257–1264.

Stein, Michael L, Zhiyi Chi et Leah J Welty (2004). "Approximating likelihoods for large spatial data sets". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.2, p. 275–296.

Stroud, J. R., M. L. Stein et S. Lysen (2014). "Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice". In : *ArXiv e-prints*. arXiv : 1402.4281 [stat.CO].

Vecchia, Aldo V (1988). "Estimation and model identification for continuous spatial processes". In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 297–312.

Xia, Gangqiang et Alan E Gelfand (2005). "Stationary process approximation for the analysis of large spatial datasets". In : *ISDS, Duke University*.

Zhou, Y., A. M Johansen et J. A. Aston (2013). "Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach". In : *ArXiv e-prints*. arXiv : 1303.3123 [stat.ME].

Chapitre 4

Bayesian tests for Conditional Independence

Bayesian inference on dependency structure via the Gaussian Copula graphical model

This is joint work with Christian P. Robert.

4.1 Introduction

Detecting dependence between random variables is a long studied problem in statistics and in the last decade in particular nonlinear measures of dependence have emerged as fundamental tools in many applied fields, when assuming a Gaussian distribution for the observed data is unrealistic. A large number of methods have been proposed like information theoretic quantities, mostly based on mutual information (see for example Kinney et Atwal, 2014 and references therein), or kernel methods (Fukumizu et al., 2007 ; Zhang et al., 2012) ; in the Bayesian literature recent developments on the subject can be found for example in Kunihama et Dunson, 2014 and Filippi et Holmes, 2015.

The difficulty with most methods is that they need to consider each pair of random variables separately in order to infer the whole dependency structure and most of them lack a proper correction for multiple testing.

Alternatively, graphical models provide an elegant way to express the full dependence structure of a set of random variables, which makes them appealing for tasks like dimension reduction in a regression setting. However they usually either rely on unconvincing linearity assumptions (like the Gaussian graphical model) or need to resort to approximations in the estimation procedure to accommodate more realistic models (Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015).

Copula models have been introduced exactly to provide a flexible tool to study multivariate data and they have been extensively studied notably for their ability to separate the modelling of the marginal distributions from the estimation of the dependence structure between them. See for example Joe, 2014 for a recent review on the subject.

The Gaussian Copula graphical model was firstly introduced in statistics by

Liu, Lafferty et Wasserman, 2009 ; Liu et al., 2012, permitting both a flexible representation of multivariate data and precise inference on the dependence structure through a conditional graph. In a Bayesian perspective it was further explored by Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015 by exploiting the G-Wishart distribution (Roverato, 2002) as a conjugate prior for the precision matrix Λ of the Gaussian Copula.

Until recently, the Bayesian literature had either to focus on decomposable graphs, in order to compute the normalizing constant of the G-Wishart, or to estimate it, introducing approximations in the procedure. Making use of the recent literature on the G-Wishart distribution (Lenkoski, 2013 ; Uhler, Lenkoski et Richards, 2014), we devise an exact MCMC estimation procedure for the Gaussian Copula graphical model that does not share these limitations. We propose as well a fully Bayesian procedure that explicitly models the marginals in the Copula in a nonparametric fashion with no assumption on their shape via a Dirichlet process prior. All the algorithms are written in C++ and make use of available CPUs and of the GPUs for linear algebra operations.

The paper proceeds as follows : in Section 4.2 we re-introduce the Gaussian Copula graphical model and fix notation ; in Section 4.3 we propose our novel estimation procedure and finally in Section 4.4 we test our method in two different simulated scenarios.

4.2 Gaussian Copula Graphical model

4.2.1 Gaussian Copula model

Consider a random vector $X = (X_1, X_2, \dots, X_d)$ with marginal distribution functions $F_1(X_1), \dots, F_d(X_d)$ and joint distribution $\pi(X)$. A Copula model is a particular way of reconstructing $\pi(X)$ thanks to Sklar's theorem, that ensures the existence of a function $C : [0,1]^d \to [0,1]$ such that

$$\pi(X) = C\left(F_1(X_1), \dots, F_d(X_d)\right).$$

This approach is quite compelling because it allows for the dependence structure to be modelled separately from the margins. In particular note that Copula modelling does not attempt at estimating the function C but rather, given a family of distributions for the copulas and the margins, to reconstruct the joint $\pi(X)$.

The Gaussian Copula is a distribution over the unit hypercube $[0,1]^d$ which, given a correlation matrix R, can be written as

$$C_R(u_1, \dots, u_d) = \Phi_d\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)|R\right) \tag{4.1}$$

where $u_1, \dots, u_d \sim \mathcal{U}_{[0,1]}$, Φ^{-1} is the inverse cdf of a standard normal and $\Phi_d(\cdot|R)$ is the cdf of a d-variate normal with zero mean and covariance matrix R. Its density can therefore be written, simply by deriving (4.1), as (Pitt, Chan et Kohn, 2006) :

$$|R|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} z'(R^{-1} - I_d)z\right) \tag{4.2}$$

where $z = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$ and I_d is the identity matrix.

It is also worth point it out that, regardless of the estimator chosen on the margins, R contains all the necessary information on the dependence structure in X.

4.2.2 Gaussian graphical model

Let $\{X_v : v \in V\}$, with $V = \{1, \dots, d\}$, be a random vector with distribution $\pi(X)$. For simplicity of exposition let $V \subseteq \mathbb{R}^d$. The conditional independence structure of X is encoded in a graph $G = (V, E)$ where each vertex $v \in V$ correspond to a random variable X_v and $E \subseteq V \times V$ is its set of edges. Due to the pairwise Markov property relative to G

$$X_{v_i} \perp\!\!\!\perp X_{v_j} | X_{V \setminus \{v_i, v_j\}} \iff (v_i, v_j) \notin E$$

Now, assume X follows a d-variate normal distribution $\mathcal{N}_d(0, \Lambda^{-1})$, where Λ is a precision matrix. We constrain some off-diagonal elements of Λ to be equal to zero and encode these constraints into a graph G, *i.e.* $E = \{(v_i, v_j) : v_i \neq v_j, \Lambda_{v_i, v_j} \neq 0\}$. A natural and conjugate prior for such $\Lambda|G$'s is the G-Wishart distribution $W_G(\delta, D)$ (Roverato, 2002), with probability density

$$p(\Lambda|\delta, D, G) = \frac{1}{I_G(\delta, D)} |\Lambda|^{(\delta-2)/2} \left(-\frac{1}{2} \mathrm{tr}(\Lambda' D) \right) \tag{4.3}$$

which reduces to the Wishart distribution when the graph is complete. The normalizing constant $I_G(\delta, D)$ is otherwise available in closed-form in a restricted number of cases or iteratively obtainable, starting from the nearest (in the graph space) known case, which is usually a chordal graph (Uhler, Lenkoski et Richards, 2014).

The posterior distribution of Λ, given n samples $X_v^{(i)}$ with $i = 1, \dots, n$, $v \in V$ and a graph G, is therefore $W_G(\delta + n, D + U)$ where $U = \sum_{i=1}^{n} X^{(i)} X^{(i)'}$.

4.2.3 Gaussian Copula graphical model

Let F_v be the univariate distribution function of X_v, F_v^{-1} its (pseudo-)inverse and f_v its density.

Instead of resorting to a correlation matrix R, the Gaussian Copula model can be parametrized through

$$R_{v_i, v_j}(\Lambda) = \frac{(\Lambda^{-1})_{v_i, v_j}}{\sqrt{(\Lambda^{-1})_{v_j, v_j} \times (\Lambda^{-1})_{v_j, v_j}}} \tag{4.4}$$

where Λ is a precision matrix conditioned on a graph G that encodes the dependence relations. We effectively enforce the dependence structure contained in the graph G into the Copula model and thus free ourselves from the Gaussian assumption on the data. Copula models have, in fact, been shown to be quite flexible ; see for example Liu, Lafferty et Wasserman, 2009.

The formal specification of the model is :

$$\begin{aligned} Z|\Lambda, G &\sim \mathcal{N}_d(0, \Lambda^{-1}), \\ \tilde{Z}|\Lambda, G &\sim \mathcal{N}_d(0, R(\Lambda)), \\ X_v^{(i)} &= F_v^{-1}\left(\Phi(\tilde{Z}_v^{(i)})\right), \\ \Lambda|G &\sim W_G(\delta, D), \\ G &\sim \pi_{\mathcal{G}}, \end{aligned} \tag{4.5}$$

$v \in V$, $i \in \{1, \dots, n\}$; we defined

$$\tilde{Z}_v^{(i)} = Z_v^{(i)} / \sqrt{[\Lambda^{-1}]_{v,v}}$$

which we can see, following (4.4), allows us to focus directly on $R(\Lambda)$ through $\Lambda|G$. $\pi_{\mathcal{G}}$ is a probability distribution over $\mathcal{G}$, the space of possible graphs.

To complete the Bayesian formulation for this model we need though to specify how the margins will be treated.

In order to perform inference on this model Hoff, 2007 suggested to essentially integrate out the margins by devising a set of constraints $\mathcal{D}$, given the data X_v, on the ordering of the latent variables Z_v; he called this approach the extended rank likelihood. He shows how the likelihood factorizes into $p(X|R, \{F_v : v \in V\}) \propto p(\mathcal{D}|R) \times p(X|R, \mathcal{D}, \{F_v : v \in V\})$ and that $p(\mathcal{D}|R)$ is now the only relevant part of the likelihood required for inferring on R which in turn is free from the dependence on the margins, see Section 4.3. This idea has been extended to graphical modelling by, e.g., Dobra et Lenkoski, 2011; Mohammadi et al., 2015.

The approach we pursue was introduced by Pitt, Chan et Kohn, 2006, who instead proposed to directly model the (assumed parametric) margins via their parameters $\Theta = \{\theta_v : v \in V\}$. The likelihood of the copula model, marginals included, can be written as

$$\begin{aligned} f(X|\Theta, R) = |R|^{-\frac{n}{2}} \times \\ \prod_{i=1}^{n} \left(\exp\left\{ -\tfrac{1}{2} \tilde{Z}^{(i)\prime} (R^{-1} - I_d) \tilde{Z}^{(i)} \right\} \right. \\ \left. \prod_{j=1}^{d} f_{v_i}(X_{v_j}^{(i)}, \theta_{v_j}) \right) \end{aligned} \tag{4.6}$$

where

$$\tilde{Z} = \left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d) \right)$$

and

$$(u_1, \dots, u_d) = (F_{v_1}^{-1}(X_{v_1}), \dots, F_{v_d}^{-1}(X_{v_d})).$$

The drawback of this formulation clearly is the need to assume a parametric family on F_v, a task that could be daunting for most realistic settings. We could choose instead to model these marginal distributions nonparametrically, almost without changing our formulation but allowing for a more flexible framework, by assuming that f_v is a mixture of distributions h_k, with k from a (potentially infinite) set $\mathcal{K}$ whose mixing proportions are driven by a Dirichlet process. This nonparametric model is common in the literature and named Dirichlet process mixture model. We can alternatively formalize it by saying that $X_v^{(i)} \sim h(X_v^{(i)}; \theta_v^{(i)})$, where the model parameters come from a random distribution $\theta_v^{(i)} \sim P$ drawn from a Dirichlet process $P \sim DP(\alpha, P_0)$ with base distribution P_0.

The Bayesian formulation of this extended Gaussian Copula graphical model

work can thus be described as follows :

$$\begin{aligned} Z|\Lambda, G \sim & \quad \mathcal{N}_d(0, \Lambda^{-1}), \\ \tilde{Z}|\Lambda, G \sim & \quad \mathcal{N}_d(0, R(\Lambda)), \\ X_v^{(i)}|\theta_v^{(i)} = & \quad F_v^{-1}\left(\Phi(\tilde{Z}_v^{(i)}); \theta_v^{(i)}\right), \\ \theta_v^{(i)} \sim & \quad P, \\ P \sim & \quad DP(\alpha, P_0), \\ \Lambda|G \sim & \quad W_G(\delta, D), \\ G \sim & \quad \pi_{\mathcal{G}}, \end{aligned} \tag{4.7}$$

$v \in V$, $i \in \{1, \dots, n\}$.

To complete the specification we mention versions of $\pi_{\mathcal{G}}$ that have been proposed in the literature. Assuming a uniform distribution over the space $\mathcal{G}$ (i.e. $\pi_{\mathcal{G}}(G) \propto 1$) helps in simplifying expressions but as pointed out in Jones et al., 2005 this uniform prior's mass with respect to the number of edges present in the graph peaks around $|V|(|V|-1)/4$, with $|V|$ the number of vertices, which favors intermediate-size graphs ; we might instead insist on a prior that encourages sparsity. More involved families of priors include priors on the number of edges in the graph (Wong, Carpendale et Greenberg, 2003), priors that encourage sparsity (Dobra et al., 2004 ; Jones et al., 2005) or priors having properties connected to multiple testing corrections (Scott et Berger, 2006). Since Bayesian inference previously focussed on decomposable graphs, as pointed out by Bornn et Caron, 2011 ; Armstrong et al., 2009, some extra care is needed as the space of decomposable graphs is quite different from the full graph space. Our work does not face this limitation and we are hence able to chose $\pi_{\mathcal{G}}$ as a binomial prior with parameter β on the number of edges :

$$\pi_{\mathcal{G}} \propto \beta^{|E|}(1-\beta)^{\binom{|V|}{2}-|E|}. \tag{4.8}$$

This prior's size peaks around $\beta \times \binom{|V|}{2}$ and we can control the desired sparsity of the graph through the parameter β.

At last, the most common Dirichlet process mixture model (DPmm) and the one that we use in this work is the DP Gaussian Mixture model, where each component of the mixture has a Gaussian density ; this implies $\theta_v^{(i)} = (\mu_v^{(i)}, \sigma_v^{(i)})$ and P_0 will then follow a Normal-Gamma or Normal-InverseGamma distribution, depending on the parametrization. However, note that the Copula model is flexible enough to accommodate discrete ad mixed data, by changing the formulation of the marginals within the DPmm.

When all margins are continuous, the Markov properties connected with this model are guaranteed to translate into Markov properties for the observed random variables. If some of the marginals were discrete, additional dependencies among the Xs could be introduced, but they are thought to have only a secondary relevance as they emerge only from the marginals (Liu, Lafferty et Wasserman, 2009).

4.2.4 Testing dependence in the Bayesian Gaussian Copula graphical model

There are two major advantages in using the Gaussian Copula graphical model when compared with other non-linear measures of dependence, as highlighted in

Dobra et Lenkoski, 2011. Both are connected with the result of the inference being a sample from the posterior distribution of G. First, the whole dependence structure is available, instead of solely pair-wise measures. Second, contrary to most other methods, uncertainty can be assessed by HPD intervals in addition to the production of an estimate of the strength of the dependence.

The probability of two variables v_i and v_j being dependent can be estimated in two ways :

— via Bayes factors involving the correlation matrix, considering hypotheses of the type $H_0 : |R_{v_i,v_j}| < \varepsilon$ versus alternatives like $H_1 : |R_{v_i,v_j}| \geq \varepsilon$ for which

$$B_{v_i,v_j} = Pr(H_1|X)/Pr(H_0|X) \tag{4.9}$$

is estimated simply by the ratio of the proportion of samples that fall respectively in the alternative or in the null hypotheses ; the threshold ε helps us control the degree of certainty needed ;

— via the posterior probability of edge inclusion, i.e. $Pr(\{v_i, v_j \in V : (v_i, v_j) \in E\})$, computed simply as the proportion of graph samples that contain the edge (v_i, v_j).

4.3 MCMC Inference

4.3.1 Bayesian modelling of the Gaussian graphical Copula

If we exhaustively model the marginals, as in (4.7), our approach is quite similar to Pitt, Chan et Kohn, 2006. The posterior $\pi(\Theta, \Lambda, G|X)$ is explored via an MCMC algorithm with sweeps between full conditionals in a Gibbs-like algorithm, sampling iteratively from $\pi(\Lambda, G|X, \Theta)$ and from $\pi(\theta_v|\{\Theta \setminus \theta_v\}, \Lambda, G, X)$, $v \in V$. We however improve the above on several points.

Considering the likelihood in (4.6) and the full model (4.7), it is clear that if the prior on (Λ, G) is $W_G(\delta, D) \times \pi_{\mathcal{G}}(G)$, the conditional distribution for Λ is

$$\Lambda|G, X, \Theta \sim W_G(\delta + n, D + U) \times \pi_{\mathcal{G}}$$

where $U = \sum_{i=1}^{n} Z^{(i)} Z^{(i)'}$.

We can sample directly from this distribution following Lenkoski, 2013, where the author proposes to sample a standard Wishart, connected with a complete graph, and then restrict this variate to the correct space through a variation of Iterative Proportion Scaling (IPS, see Dempster, 1972).

Sampling from the graph space is slightly more involved as every move in $\mathcal{G}$ implies a move on Λ as well, which needs to be restricted to the new correct space. Roverato, 2002 shows that all the elements of the upper triangular matrix Ψ, with $\Psi'\Psi = \Lambda$, aside for the ones on the diagonal and the ones that corresponds to the edges E of the graph are non-free and can be obtained from the others. Making use of this Dobra et Lenkoski, 2011 devise an algorithm that proposes a new neighbour graph G^*, meaning selecting randomly two vertices and either adding or removing the corresponding edge from the current G, restrict the Cholesky decomposition of the current precision matrix based on the above and finally recover the proposed Λ^*.

Wang et Li, 2012 and Lenkoski, 2013 noted how this algorithm relies heavily on approximating the prior normalizing constant for the G-Wishart, needed for computing the Metropolis-Hastings ratio of the proposal, and showed how this approximation may fail in high dimension. They propose instead to rely on a variation of the exchange algorithm (Murray, Ghahramani et MacKay, 2012) (named double reversible jump in Lenkoski, 2013) to obviate the problem.

Theorem 3.7 in Uhler, Lenkoski et Richards, 2014 does provide a way to compute the ratio of the normalizing constants for two precision matrices associated with neighbour graphs, namely given $G = (V, E)$ and $G^e = (V, E^e)$, the graph with an additional edge e (i.e. $E^e = E \cup e$),

$$I_G(\delta, I_d) = \pi^{-\frac{1}{2}} \frac{\Gamma(\delta + \frac{1}{2}(d+2))}{\Gamma(\delta + \frac{1}{2}(d+3))} I_{G^e}(\delta, I_d) \tag{4.10}$$

where d denote the number of triangles formed by the edge e and two other edges in G^e.

This suffices to adapt the proposal of Dobra et Lenkoski, 2011 into an exact algorithm for a generic undirected graph and the corresponding restriction of the associated precision matrix Λ.

The only other difference between the original paper and our proposal is that instead of a uniform prior over the graph space we encourage sparsity via the prior specified in (4.8), inserting its expression as needed in the final Metropolis-Hastings ratio.

For the second part, $\pi(\theta_v | \{\Theta \setminus \theta_v\}, \Lambda, G, X)$, extra care is needed. Even though the likelihood in (4.6) seems to factorize well even with respect to Θ, $\tilde{Z}$ depends implicitly on the marginals and hence change as Θ moves despite Λ remaining fixed.

A number of samplers are available for Dirichlet process mixture models and essentially all do work with due precautions. We opted for the one introduced by Ge et al., 2015, based on the slice sampler of Walker, 2007 ; Kalli, Griffin et Walker, 2011 for its ability to be easily and effectively parallelized.

In this case each time new components are constructed through stick-breaking, $\tilde{Z}$ has to be updated. When the latent component-assignment variables are resampled for each data point $X_v^{(i)}$, the factor

$$\left(\prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} \tilde{Z}^{(i)\prime} (R^{-1} - I_d) \tilde{Z}^{(i)} \right\} \right) f_v^{(i)}(X_v^{(i)}, \theta_v^{(i)})$$

has to be considered, instead of just the density $f_v^{(i)}$ of the proposed new component. Finally, when a new sample is proposed for the parameters θ_v, by the standard procedure of Ge et al., 2015, we need to compute the corresponding proposed $\tilde{Z}^*$ and accept/reject based on

$$\frac{\prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} \tilde{Z}^{*(i)\prime} (R(\Lambda)^{-1} - I_d) \tilde{Z}^{*(i)} \right\}}{\prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} \tilde{Z}^{(i)\prime} (R(\Lambda)^{-1} - I_d) \tilde{Z}^{(i)} \right\}}.$$

For this last step we still use conjugate priors to ease computations but, since we need to include the factors connected to the Copula, different priors could be potentially

considered with little effort. Banterle et al., 2015 for example implies in fact that if the preferred updates were coming from a Metropolis-Hastings type of sampler, computing later this second acceptance/rejection step still preserves the correct target distribution.

It is worth stressing that the explicitly modelling of the marginals explained above allows us not only to infer the dependence structure of X (see Section 4.2.4) but as a by-product we obtain an estimation of its joint distribution as well.

4.3.2 Semi-Parametric modelling of the graphical Gaussian Copula

If the marginals or the joint distribution of X are not of direct interest we can effectively integrate them out of the sampler by noting (Hoff, 2007) that, since F_v^{-1} and Φ are nondecreasing, the model in (4.5) implies the following relations between X and Z (or equivalently on $\tilde{Z}$) :

$$\begin{aligned} X_v^{(i)} < X_v^{(j)} &\implies Z_v^{(i)} < Z_v^{(j)}, \\ Z_v^{(i)} < Z_v^{(j)} &\implies X_v^{(i)} \leq X_v^{(j)}. \end{aligned} \tag{4.11}$$

In general, given X_v, the latent samples Z_v are constrained to the set

$$A(X) = \{Z \in \mathbb{R}^{n \times d} : L_v^j(Z) < Z_v^{(j)} < U_v^j(Z)\} \tag{4.12}$$

where

$$\begin{aligned} L_v^j(Z) = &\max \left\{ Z_v^{(k)} : X_v^{(k)} < X_v^{(j)} \right\}, \\ U_v^j(Z) = &\min \left\{ Z_v^{(k)} : X_v^{(j)} < X_v^{(k)} \right\}. \end{aligned} \tag{4.13}$$

We can also accomodate missing values in the observations by setting $L_v^j(Z)$ and $U_v^j(Z)$ respectively to $-\infty$ and ∞ if the value of $X_v^{(j)}$ is not present.

Now, calling $\mathcal{D}$ the event $\{Z \in A(X)\}$, we can rewrite the likelihood as (see Hoff, 2007)

$$f(X|\Lambda, G) = f(\mathcal{D}|\Lambda, G) \times f(X|\mathcal{D}, R(\Lambda), \{F_v : v \in V\}).$$

$f(\mathcal{D}|\Lambda, G)$ is the only relevant factor for inferring (Λ, G) and hence Hoff, 2007 (and later Dobra et Lenkoski, 2011 among others, in a graphical setting) constructs a Gibbs sampler on the posterior distribution

$$\pi(\Lambda, G|\mathcal{D}) \propto f(\mathcal{D}|\Lambda, G)\pi_\Lambda(\Lambda|G)\pi_\mathcal{G}(G).$$

At each iteration we need to resample the latent variables Z first and then update the precision matrix and its associated graph.

Dobra et Lenkoski, 2011 show that the full conditional distribution for $Z_v|Z_{V\setminus v}$ for each latent data point $Z_v^{(j)}$ is normal

$$\mathcal{N}\left(- \sum_{v' \in S_G(v)} \frac{\Lambda_{v',v}}{\Lambda_{v,v}} Z_{v'}, \frac{1}{\Lambda_{v,v}} \right) \tag{4.14}$$

where $S_G(v) = \{v' \in V : (v, v') \in E\}$, truncated between L_v^j and U_v^j as defined in (4.13).

Once Z is updated, we must resample Λ and G. Once again, due to the lack of a direct sampler for G-Wishart variates, the sampler in Dobra et Lenkoski, 2011 involved approximations on the normalizing constant for the prior distribution of (Λ, G). We proceed exactly as in the previous subsection, adapting their proposal for neighbour graphs to use the exact ratio of normalizing constants as in (4.10).

With this model we can include categorical, ordinal and binary variables without any adjustment to the algorithm. Finally, as sampling from the truncated normals in (4.14) is generally computationally cheaper than exploring the space of Θ, this version of the algorithm will be slightly faster, at the expenses of the inference on the marginals and consequently of the full joint distribution of X.

4.3.3 Implementation details

All the procedures highlighted above were implemented through C++ code making use of the OMP API (`http://openmp.org`) for the parallel parts. Armadillo (Sanderson, 2010) was used as a library for linear algebra, linked to NVBLAS (`http://docs.nvidia.com/cuda/nvblas/`) to potentially deviate computations to available GPUs.

Interfaces with **R** (R Core Team, 2015) are included for each method thanks to the Rcpp package (Eddelbuettel et François, 2011).

The G-Wishart sampler used is the one implemented in the *BDgraph* **R** package (Mohammadi et Wit, 2015).

4.4 Experiments

Applications for the above method include but are not limited to network determination in social science, economics and especially in biology where a wealth of high-throughput data to be analysed is emerging, whether from genetic, proteomic or transcriptomic data. Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015, but also Liu, Lafferty et Wasserman, 2009 ; Liu et al., 2012, are good resources where the Gaussian Copula graphical model is found able to reproduce with accuracy graph patterns in the above settings ; Dobra et Lenkoski, 2011 ; Mohammadi et al., 2015 in particular, aside from slight approximations, share the same modelling framework with the one presented in this work.

In the following we will compare the method with Filippi et Holmes, 2015 on a few 2-dimensional synthetic datasets, to show where the Gaussian Copula graphical model can be applied to detect dependence between two, or more, random variables.

A different area of application, found for example in Kunihama et Dunson, 2014, is dimension reduction in regression problems. In the Machine Learning literature this problem has already been tackled with RKHS-type of estimators for cross-covariance operators (Fukumizu et al., 2007 ; Fukumizu, Bach et Jordan, 2009 ; Zhang et al., 2012). These estimators however often share the same difficulty of being forced to evaluate multiple pair-wise relations, which in dimension reduction can result in a quite computationally expensive algorithm if applied as a criterion for best subset selection.

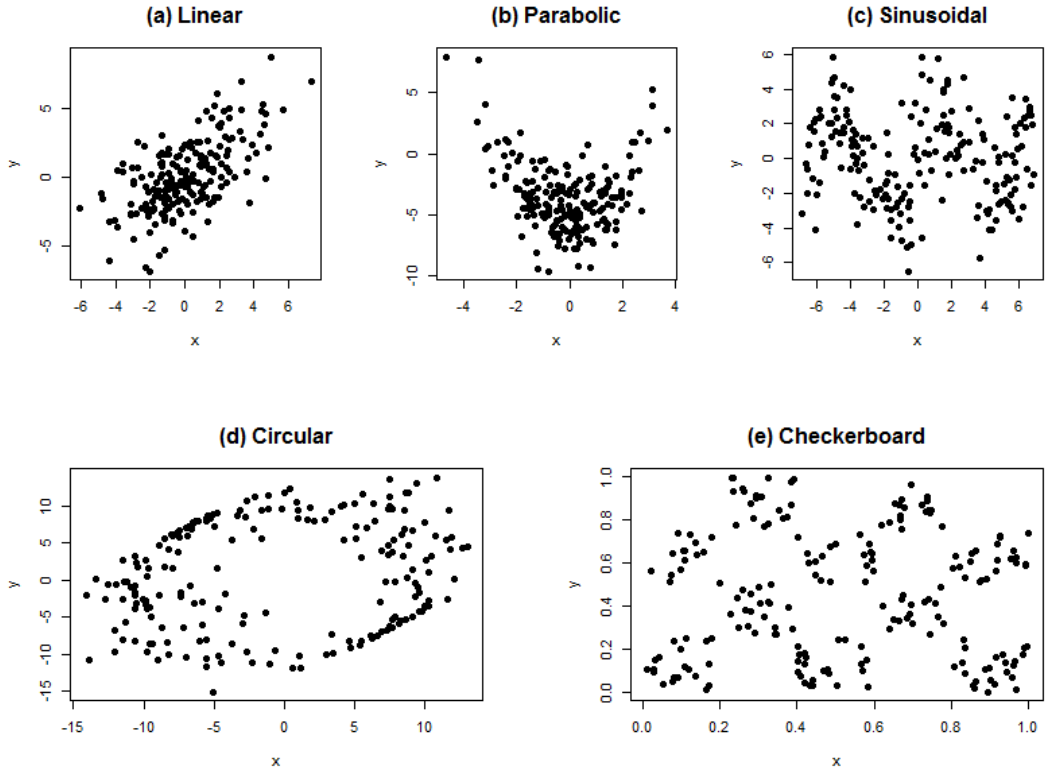


FIGURE 4.1: Synthetic datasets generated for $x, \varepsilon \sim \mathcal{N}(0,1), \theta \sim \mathcal{U}(0,1)$, from (a) a linear model ($y = 2x/3 + \varepsilon$), (b) a parabolic model ($y = 2x^2/3 + \varepsilon$), (c) a sinusoidal ($y = \sin(x) + \varepsilon$), (d) a circular model ($x = 10\cos(\theta) + \varepsilon, y = 10\sin(\theta) + \varepsilon$) and (e) a checkerboard model (KINNEY et ATWAL, 2014).

In the second subsection we will thus test Approximate Bayesian Computation (ABC,MARIN et al., 2012) performances on a coalescent model proposed in JOYCE et MARJORAM, 2008; BLUM et al., 2013; NUNES et BALDING, 2010 after having reduced the dimension of the problem thanks to the Gaussian Copula graphical model.

4.4.1 Challenging synthetic bivariate data

Following FILIPPI et HOLMES, 2015 we test our procedure for its capability to detect dependence in a few challenging synthetic datasets generated following the guidelines given in KINNEY et ATWAL, 2014; an example for $n = 200$ and normal noise with variance $\sigma^2 = 2$ is shown in Figure 4.1.

We let either sample size n vary from a few units to a 1000 samples keeping the noise level $\sigma^2 = 2$, or fix $n = 200$ and move σ^2 from 0.1 to 3. In principle we do not need replications to compute the variability of our estimated inclusion probability for a given sample X, but in order to remove the effect of a particular dataset on the results, and to directly compare with FILIPPI et HOLMES, 2015, we averaged all above runs 100 times.

The results for $N = 1000$ iterations of the Markov Chains are presented in Figures 4.2, 4.4 for the model in (4.7) and in Figures 4.3,4.5 for the semiparametric modelling procedure explained in 4.3.2. While being quite similar, the semiparametric version seems to outperform slightly the full model, in particular for models (a)

and (b), probably because of the information spent in estimating the marginals. The first method seems to perform better against the circular model.

It appears however that the model is not able to correctly guess the dependence in the last model, the Checkerboard. The Gaussian Copula, although flexible in its formulation, is eventually a parametric model and being able to find counter-examples is not surprising ; Copula modelling being based on the marginal distributions, which result to be uniform in this case, is another complication for our formalization. We are however positively surprised that example (d), the circular distribution, is detected so well, again supporting the robustness of the Gaussian Copula family of distribution ; see for example Liu et al., 2012.

As remarked by Figure 4.6, for both the full model and the extended rank likelihood model, the results for the synthetic model (e) should raise a warning sign anyway given that even for high sample-size the estimated inclusion probability, our proxy for dependence, does not converge to zero but rather floats around 0.5, which is the prior probability β of the edge presence in this experiment. Figure 4.6 shows that for an independent model the Gaussian Copula graphical model correctly estimate that the edge should have a low probability of presence.

We can argue nonetheless that for most of non-crafted, real-world application the Gaussian Copula graphical model is an adaptable enough tools for density and especially dependency estimation, as highlighted by the literature which successfully uses it in a wide range of applied works.

Note moreover that all these presented datasets are bivariate, even if the Gaussian Copula graphical model can go beyond two dimensions. We generated hence some additional dimensions (up till $d = 10$, meaning eight extra) by permutations from the initial two. We noticed no relevant difference between the presented results for the estimation of the inclusion probability of the edge between the original data if not for very small sample sizes where the posterior distributions were slightly flatter, reflecting the additional uncertainty introduced by the extra dimensions.

4.4.2 ABC Sufficient Dimension Reduction via Gaussian Copula graphical model

In the ABC context is typical to have a set s_{obs} given by experts of the field due to the too large or complex structure of the data y_{obs} (e.g. in population genetics or epidemiology) that prevent comparison directly between raw data. In this situation redundant information in the form of collinearity or, more generally, dependence between elements in s is expected. Finding the minimal subset $u \in s$ that contains the whole information is hence critical in order to reduce the dimension of s_{obs} and improve the performance of ABC in all his variants (Marin et al., 2012 ; Blum et al., 2013). Additional informations are in the Supplementary material, Section 4.6.1.

If we define the minimal-dimension maximal-informative set $u \in s$ as the subset such that

$$p(\theta|u) \perp\!\!\!\perp p(s) \text{ which implies } p(\theta|u) = p(\theta|s) \tag{4.15}$$

it is clear to see that if s is not sufficient neither will be u, but given a set s there is no better set to use for ABC than u in our view. A similar perspective is adopted in a regression context for example in Fukumizu, Bach et Jordan, 2009.

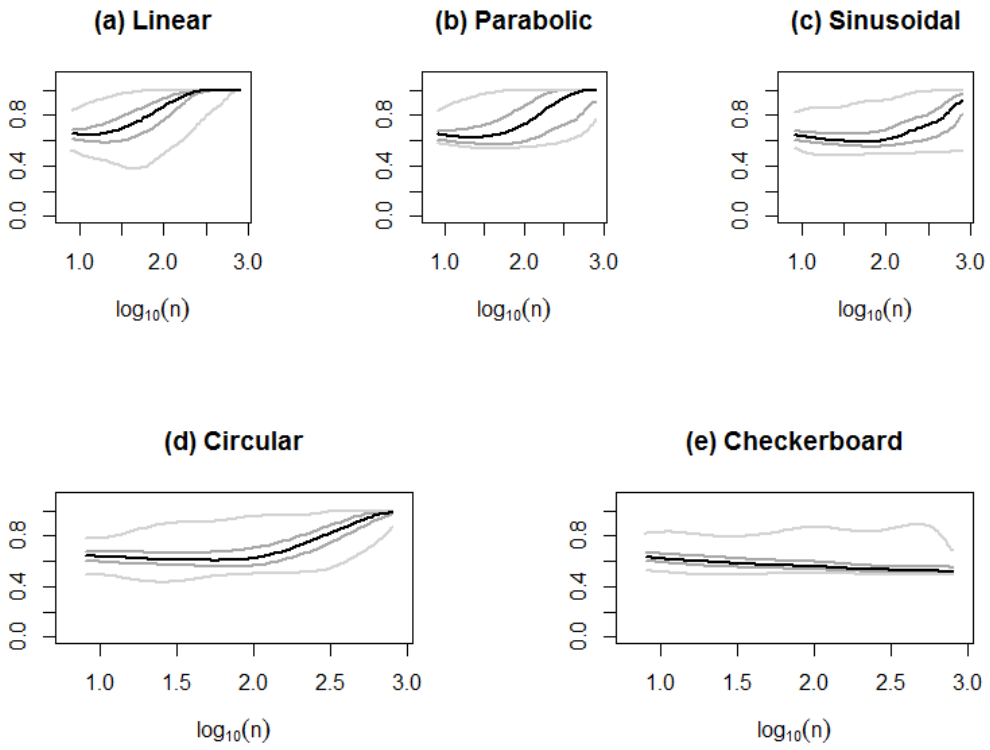


FIGURE 4.2: $\sigma^2 = 2$, each vertical slice represent the observed distribution over 100 runs of the full Gaussian Copula graphical model, with exhaustive estimation of the marginals, of the probability of inclusion of the edge (v_1, v_2) for a given sample size n. $(v_1, v_2) \in E$ expressing dependence in the considered model. Represented in black the mean, in dark gray the first and third quartiles and in light grey the minimum and maximum observed values.

This conditional independence structure can easily be translated into a graph G and hence this example seems like a prime application for our proposed method.

We proceed to measure ABC performances in a few cases : standard ABC rejection algorithm (MARIN et al., 2012), semi automatic ABC (FEARNHEAD et PRANGLE, 2012), best subset selection with minimum Entropy criterion (NUNES et BALDING, 2010) and finally summary selection via Gaussian Copula graphical model. The considered measure of performances is the averaged (across independent repetition on different data) root sum of squared errors, named $\overline{RSSE}$, rescaled to intuitively give a performance gain with respect to the standard ABC.

Following BLUM et al., 2013 we test the procedure on a dataset simulated from a coalescent model, first considered by JOYCE et MARJORAM, 2008 ; NUNES et BALDING, 2010, with a two-dimensional parameter (ψ, ρ), six *real* summary statistics and five independent random variables as confounding factors. The reader is referred to the Supplementary material for additional information on the model and on the procedure ; see Section 4.6.1.

Results are presented in Table 4.1.

The results on the Entropy method seem concerning, given that the method barely perform better than standard ABC, in clear discordance with the results presented in BLUM et al., 2013. As shown in Figure 4.7 the reason is most likely that

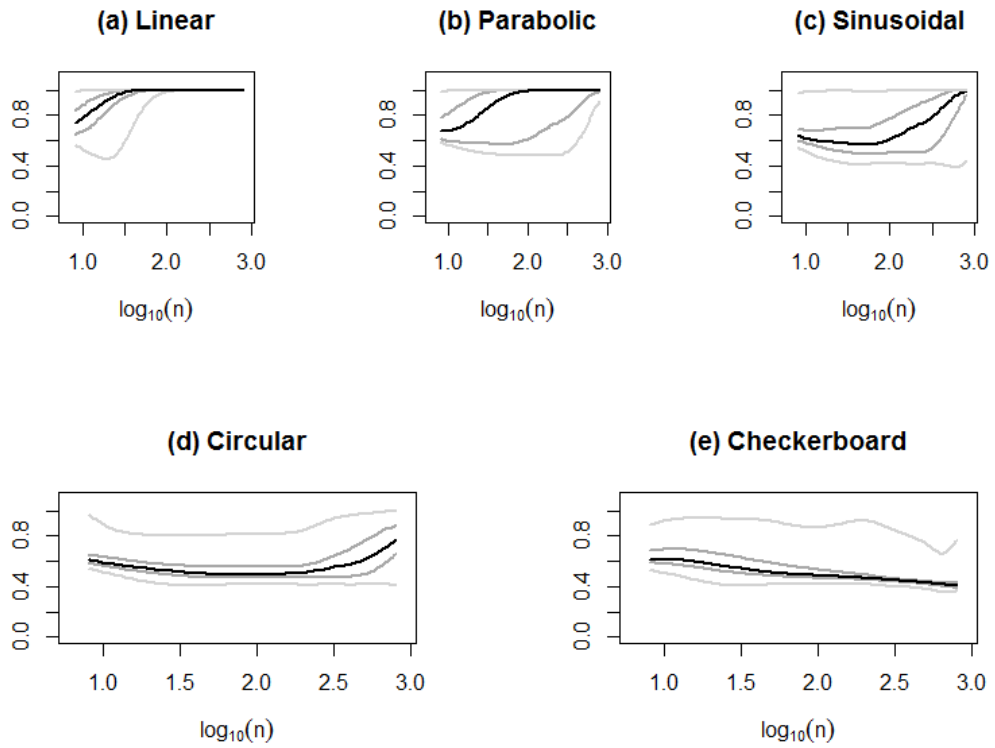


FIGURE 4.3: $\sigma^2 = 2$, each vertical slice represent the observed distribution over 100 runs of the semiparametric Gaussian Copula graphical model of the probability of inclusion of the edge (v_1, v_2) for a given sample size n. $(v_1, v_2) \in E$ expressing dependence in the considered model. Represented in black the mean, in dark gray the first and third quartiles and in light grey the minimum and maximum observed values.

the two-stages procedure needs in principle to perform the whole ABC procedure for each subset of summaries. With only $B = 500$ points to train on, the corresponding ABC estimates are highly inaccurate and hence in all but a few repetitions the method chooses to keep the whole set, reverting thus to standard ABC. Performances could surely be improved by working on larger training samples but the associated gain is quickly overshadowed by the extra computational cost.

In conclusion we have shown that our method, even for small training samples, outperform the competitor best-subset-selection methods and even the semi-automatic ABC of FEARNHEAD et PRANGLE, 2012. For reference, in Figure 4.8 we reported the graph representing in red the relevant associations between parameters and summaries chosen by the Gaussian Copula graphical model.

TABLE 4.1: Relative $\overline{RSSE}$ on a coalescent model

GCgm	**std ABD**	**F&P**	**Entropy**
0.732	1.00	0.816	0.98

As a final remark, though we will not pursue this direction in the present work, note that when we are interested in dimension reduction we could take advantage of the fact that we are primarily interested in the dependence structure between the response variable and the covariates. We could devise a prior probability, and a corresponding proposal, on the graph space $\mathcal{G}$ such that only these edges are allowed to vary, while the others are forced to be present. This will weaken the modelling within the covariates, and possibly within the response variables, but will help in exploring faster $\mathcal{G}$ in the Markov Chain, leading to a faster mixing.

4.5 Conclusions

In this work we introduce an exact algorithm to perform Bayesian inference on the Gaussian Copula graphical model. We also extend on the literature by providing an exact Markov Chain Monte Carlo procedure to estimate together the copula and the marginals, making use of the Dirichlet process mixture model.

This flexible model, aimed primarily at detecting dependence between random variables, has a wide array of possible usages both in applied and in methodological works and we showed its performances on two very different simulated examples to reflect that.

Future work should be focused on improving the current proposal on the graph space, limited for the moment to neighbour graphs, as we conjecture that global moves would dramatically improve the performances of the already efficient algorithm. The method could finally also be extended via vine factorizations.

Acknowledgements

Christian P. Robert research is partly financed by Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) on the 2012–2015 ANR-11-BS01-0010 grant "Calibration" and by a 2010–2015 senior chair grant of Institut Universitaire de France. Marco Banterle PhD is funded by Université Paris Dauphine.

4.6 Supplementary material to : Bayesian inference on dependency structure

4.6.1 ABC Sufficient Dimension Reduction via Gaussian Copula graphical model

Dimension Reduction in ABC

Approximate Bayesian Computation (ABC) is a family of algorithms which aim is to perform Bayesian inference, when the likelihood term is intractable either for mathematical or computational reasons (Marin et al., 2012).

ABC algorithms are usually built on top of two approximations :

— A summary s_{obs} of the data y_{obs} is used, typically to allow a reasonable comparison between simulation and observation,

$$p(\theta|s_{obs}) \approx p(\theta|y_{obs})$$

— The ABC posterior is finally obtained by marginalizing a sample from

$$p(\theta, s|s_{obs}) \sim p(\theta) \times p(s|\theta) \times K_\varepsilon(||s - s_{obs}||)$$

generated by sampling from $p(\theta) \times p(s|\theta)$ and weighting with a kernel function $K_\varepsilon(\cdot)$.

While it is impossible to quantify the loss in taking a non-sufficient statistics, a low dimensional quasi-sufficient statistics helps in reducing the obvious trade off arising between the two approximations because of the so-called *curse of dimensionality.*

Coalescent experiment

The data consist in a two dimensional parameter $\theta = (\psi, \rho)$, mutation and recombination rate respectively. For each parameter 5001 base pair DNA sequences for 50 individuals are generated from the coalescent model and finally 6 summary statistics are computed. We add another 5 independent random variables with different distributions as confounding factors.

We will select the subset u as in (4.15) by estimating the underlying graph G via the proposed model and then perform ABC conditioned solely on u via a standard rejection sampling and later applying non-linear regression adjustment Marin et al., 2012 ; Blum et al., 2013 ; Blum, 2010. The same correction will be applied to every model compared. Note finally that, as some of the summaries are discrete, we will use the procedure of Section 4.3.2.

To compare our proposition with the ones proposed in Blum et al., 2013, we reproduce part of the methods used thanks to the **R** package ABCtools (Nunes et Prangle, 2015) ; namely we tested the semi automatic procedure of Fearnhead et Prangle, 2012, the two-stages minimum Entropy of Nunes et Balding, 2010 and standard ABC.

The quantity we will use to quantify performances is the mean root sum of square errors ($\overline{RSSE}$,) as in Blum et al., 2013. At each repetition we select one random couple (θ_i, S_i) in the reference table of simulated values and consider it as (θ_{true}, S_{obs}) ; we then divide the rest of the dataset into $B = 500$ points for the

training set, which we use to estimate the subset of summaries u, and the rest will be used to perform the ABC rejection algorithm.

We repeat the procedure 500 times, each time obtaining N points θ_j from the approximate posterior by keeping the first 1% closest to S_{obs} (with respect to the Euclidian distance) and compute the RSSE as in

$$RSSE = \sum_{j=1}^{N} \left(||\theta_j, \theta_{true}||^2 \right)^{1/2}.$$

We finally average the repetitions to obtain our measure of performance as

$$\overline{RSSE} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{N} \left(||\theta_j, \theta_i||^2 \right)^{1/2} \tag{4.16}$$

and rescale it with respect to the standard ABC performance.

4.6.2 Figures

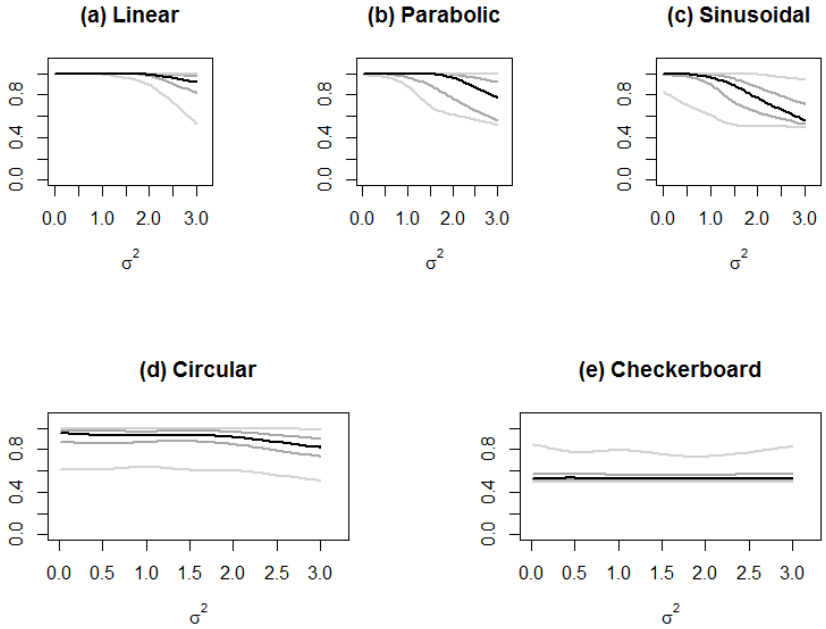


FIGURE 4.4: $n = 200$, each vertical slice represent the observed distribution over 100 runs of the semiparametric Gaussian Copula graphical model of the probability of inclusion of the edge (v_1, v_2) for a given noise level σ^2. $(v_1, v_2) \in E$ expressing dependence in the considered model. Represented in black the mean, in dark gray the first and third quartiles and in light grey the minimum and maximum observed values.

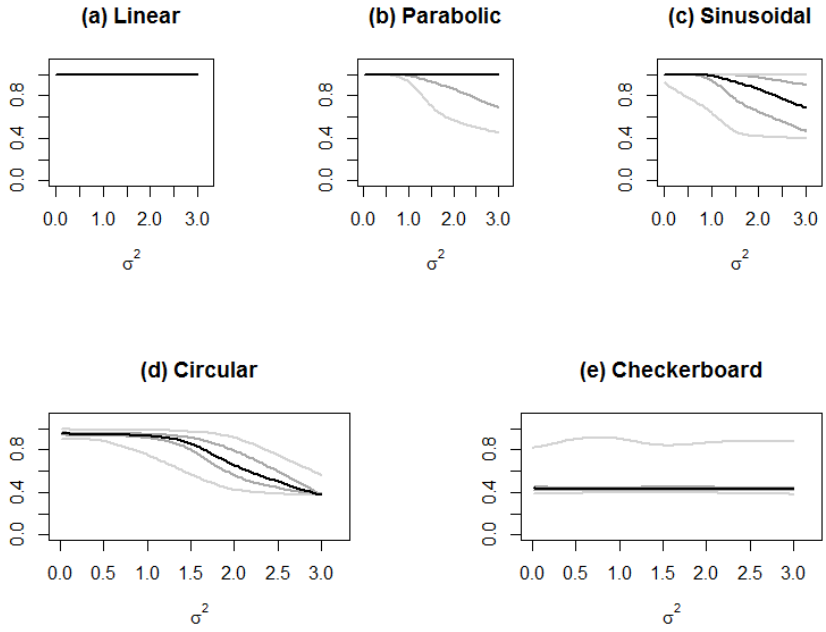


FIGURE 4.5: $n = 200$, each vertical slice represent the observed distribution over 100 runs of the full Gaussian Copula graphical model, with exhaustive estimation of the marginals, of the probability of inclusion of the edge (v_1, v_2) for a given noise level σ^2. $(v_1, v_2) \in E$ expressing dependence in the considered model. Represented in black the mean, in dark gray the first and third quartiles and in light grey the minimum and maximum observed values.

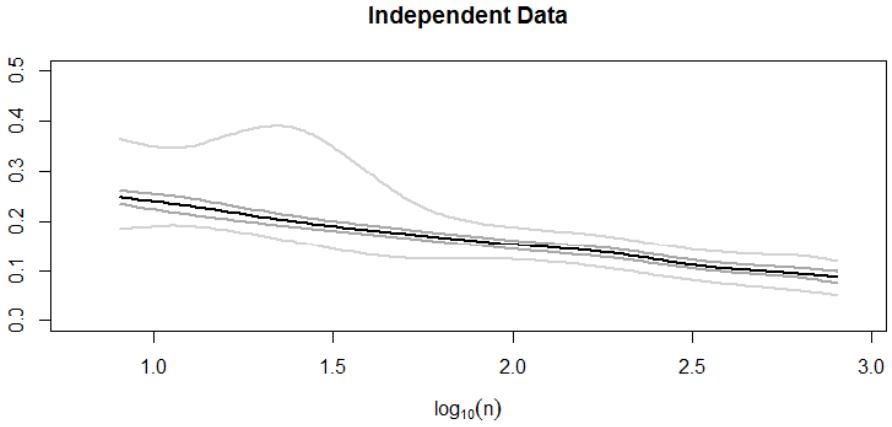


FIGURE 4.6: Probability of edge inclusion –i.e. dependence– in the case of independent bivariate normal data for the Gaussian Copula graphical model

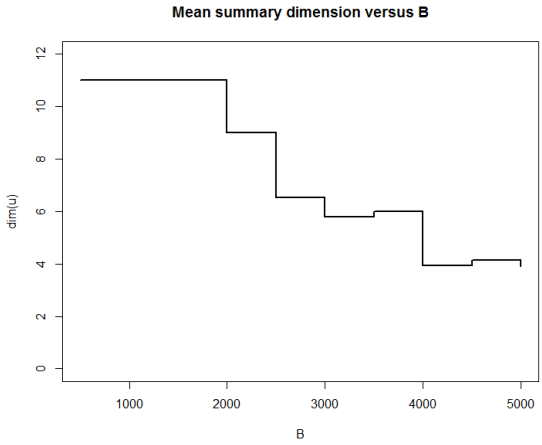


FIGURE 4.7: Path representing the mean dimension of the selected summary statistics for the two-stages minimum Entropy criterion versus ten different sizes of training set B (averaged over 20 runs per sample size, over different randomized samples). Only from around $B = 2000$ the procedure is consistently able to exclude some of the independent variable from the set, even though it quickly converges to a more reasonable estimate from there.

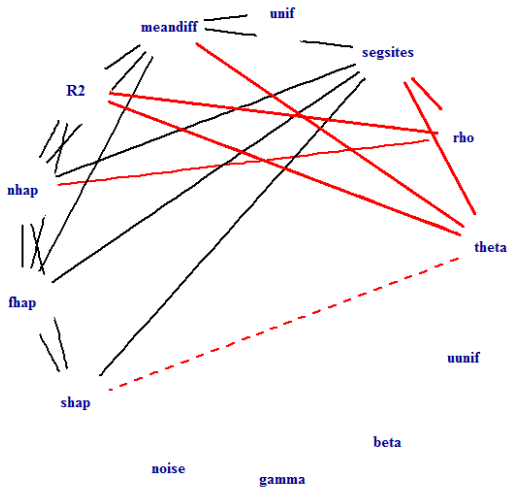


FIGURE 4.8: In red the relevant edges consistently selected to be used in the ABC procedure by the Gaussian Copula graphical model. Dashed in red one edge selected approximatively 60% of the times.

References

Armstrong, Helen et al. (2009). "Bayesian covariance matrix estimation using a mixture of decomposable graphical models". In : *Statistics and Computing* 19.3, p. 303–316.

Banterle, M. et al. (2015). "Accelerating Metropolis-Hastings algorithms by Delayed Acceptance". In : *ArXiv e-prints*. arXiv : `1503.00996 [stat.CO]`.

Blum, M. (2010). "Approximate Bayesian Computation: a non-parametric perspective". In : 105.491, p. 1178–1187.

Blum, M. G. B. et al. (2013). "A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation". In : *Statistical Science* 28.2, p. 189–208.

Bornn, Luke, François Caron et al. (2011). "Bayesian clustering in decomposable graphs". In : *Bayesian Analysis* 6.4, p. 829–846.

Dempster, Arthur P (1972). "Covariance selection". In : *Biometrics*, p. 157–175.

Dobra, Adrian et Alex Lenkoski (2011). "Copula Gaussian graphical models and their application to modeling functional disability data". In : *Ann. Appl. Stat.* 5.2A, p. 969–993.

Dobra, Adrian et al. (2004). "Sparse graphical models for exploring gene expression data". In : *Journal of Multivariate Analysis* 90.1, p. 196–212.

Eddelbuettel, Dirk et Romain François (2011). "Rcpp: Seamless R and C++ Integration". In : *Journal of Statistical Software* 40.8, p. 1–18.

Fearnhead, Paul et Dennis Prangle (2012). "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, p. 419–474.

Filippi, S. et C. Holmes (2015). "A Bayesian nonparametric approach to quantifying dependence between random variables". In : *ArXiv e-prints*. arXiv : `1506.00829 [stat.ME]`.

Fukumizu, Kenji, Francis R Bach et Michael I Jordan (2009). "Kernel dimension reduction in regression". In : *The Annals of Statistics*, p. 1871–1905.

Fukumizu, Kenji et al. (2007). "Kernel Measures of Conditional Dependence." In : *NIPS*. T. 20, p. 489–496.

Ge, Hong et al. (2015). "Distributed Inference for Dirichlet Process Mixture Models". In : *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, p. 2276–2284.

Hoff, Peter D. (2007). "Extending the rank likelihood for semiparametric copula estimation". In : *Ann. Appl. Stat.* 1.1, p. 265–283.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. isbn : 9781466583221.

Jones, Beatrix et al. (2005). "Experiments in stochastic computation for high-dimensional graphical models". In : *Statistical Science*, p. 388–400.

Joyce, Paul et Paul Marjoram (2008). "Approximately sufficient statistics and Bayesian computation". In : *Statistical applications in genetics and molecular biology* 7.1.

Kalli, Maria, Jim E Griffin et Stephen G Walker (2011). "Slice sampling mixture models". In : *Statistics and computing* 21.1, p. 93–105.

Kinney, Justin B. et Gurinder S. Atwal (2014). "Equitability, mutual information, and the maximal information coefficient". In : *Proceedings of the National Academy of Sciences* 111.9, p. 3354–3359. eprint : `http://www.pnas.org/content/111/9/3354.full.pdf`.

Kunihama, T. et D. B. Dunson (2014). "Nonparametric Bayes inference on conditional independence". In : *ArXiv e-prints*. arXiv : `1404.1429 [stat.ME]`.

Lenkoski, Alex (2013). "A direct sampler for G-Wishart variates". In : *Stat* 2.1, p. 119–128.

Liu, Han, John Lafferty et Larry Wasserman (2009). "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs". In : *J. Mach. Learn. Res.* 10, p. 2295–2328.

Liu, Han et al. (2012). "High-dimensional semiparametric Gaussian copula graphical models". In : *Ann. Statist.* 40.4, p. 2293–2326.

Marin, Jean-Michel et al. (2012). "Approximate Bayesian computational methods". English. In : *Statistics and Computing* 22.6, p. 1167–1180.

Mohammadi, A. et al. (2015). "Bayesian Gaussian Copula Graphical Modeling for Dupuytren Disease". In : *ArXiv e-prints*. arXiv : `1501.04849 [stat.AP]`.

Mohammadi, Abdolreza et Ernst Wit (2015). *BDgraph: Bayesian Graph Selection Based on Birth-Death MCMC Approach*. R package version 2.22. url : `http://CRAN.R-project.org/package=BDgraph`.

Murray, Iain, Zoubin Ghahramani et David MacKay (2012). "MCMC for doubly-intractable distributions". In : *arXiv preprint arXiv:1206.6848*.

Nunes, Matthew A et David J Balding (2010). "On optimal selection of summary statistics for approximate Bayesian computation". In : *Statistical applications in genetics and molecular biology* 9.1.

Nunes, Matthew A. et Dennis Prangle (2015). "abctools: An R package for tuning Approximate Bayesian Computation analyses". Forthcoming.

Pitt, Michael, David Chan et Robert Kohn (2006). "Efficient Bayesian Inference for Gaussian Copula Regression Models". English. In : *Biometrika* 93.3, pp. 537–554.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. url : `http://www.R-project.org/`.

Roverato, Alberto (2002). "Hyper Inverse Wishart Distribution for Non-Decomposable Graphs and Its Application to Bayesian Inference for Gaussian Graphical Models". English. In : *Scandinavian Journal of Statistics* 29.3, pp. 391–411.

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Rapp. tech. NICTA.

Scott, James G et James O Berger (2006). "An exploration of aspects of Bayesian multiple testing". In : *Journal of Statistical Planning and Inference* 136.7, p. 2144–2162.

Uhler, Caroline, Alex Lenkoski et Donald Richards (2014). "Exact formulas for the normalizing constants of Wishart distributions for graphical models". In : *arXiv preprint arXiv:1406.4901*.

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices". In : *Comm. Statist.* 36, p. 45–54.

Wang, Hao, Sophia Zhengzi Li et al. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions". In : *Electronic Journal of Statistics* 6, p. 168–198.

Wong, Nelson, Sheelagh Carpendale et Saul Greenberg (2003). "Edgelens: An interactive method for managing edge congestion in graphs". In : *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*. IEEE, p. 51–58.

Zhang, Kun et al. (2012). "Kernel-based conditional independence test and application in causal discovery". In : *arXiv preprint arXiv:1202.3775*.

Bibliographie

Andrieu, C. et G.O. Roberts (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In : 37.2, p. 697–725.

Andrieu, Christophe, Anthony Lee et Matti Vihola (2013). "Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers". In : *arXiv preprint arXiv:1312.6432*.

Andrieu, Christophe et Éric Moulines (2006). "On the ergodicity properties of some adaptive MCMC algorithms". In : *Ann. Appl. Probab.* 16.3, p. 1462–1505. issn : 1050-5164.

Angelino, E. et al. (2014). "Accelerating MCMC via Parallel Predictive Prefetching". In : *arXiv preprint arXiv:1403.7265*.

Armstrong, Helen et al. (2009). "Bayesian covariance matrix estimation using a mixture of decomposable graphical models". In : *Statistics and Computing* 19.3, p. 303–316.

Atchade, Y. F. et J. S. Rosenthal (2005). "On adaptive markov chain monte carlo algorithm". In : *Bernoulli* 11.5, p. 815–828.

Atchade, Yves et al. (2009). "Adaptive Markov chain Monte Carlo: theory and methods". In : *Bayesian Time Series Models*, p. 33–53.

Banerjee, Sudipto et al. (2008). "Gaussian predictive process models for large spatial data sets". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, p. 825–848.

Banterle, M. et al. (2015). "Accelerating Metropolis-Hastings algorithms by Delayed Acceptance". In : *ArXiv e-prints*. arXiv : 1503.00996 [stat.CO].

Bayes, T. (1763). "An Essay Toward Solving a Problem in the Doctrine of Chances". In : *Philosophical Transactions of the Royal Society of London* 53, p. 370–418.

Beal, Matthew James (2003). *Variational algorithms for approximate Bayesian inference*. University of London London.

Blum, M. (2010). "Approximate Bayesian Computation: a non-parametric perspective". In : 105.491, p. 1178–1187.

Blum, M. G. B. et al. (2013). "A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation". In : *Statistical Science* 28.2, p. 189–208.

Bornn, Luke, François Caron et al. (2011). "Bayesian clustering in decomposable graphs". In : *Bayesian Analysis* 6.4, p. 829–846.

Bornn, Luke, Gavin Shaddick et James V. Zidek (2012). "Modeling Nonstationary Processes Through Dimension Expansion". In : *Journal of the American Statistical Association* 107.497, p. 281–289. doi : 10.1080/01621459.2011.646919.

Breyer, L.A. et G.O. Roberts (2000). "From metropolis to diffusions: Gibbs states and optimal scaling". In : *Stochastic Processes and their Applications* 90.2, p. 181 –206. issn : 0304-4149. doi : http://dx.doi.org/10.1016/S0304-4149(00)00041-7. url : http://www.sciencedirect.com/science/article/pii/S0304414900000417.

Brockwell, A.E. (2006). "Parallel Markov chain Monte Carlo Simulation by Pre-Fetching". In : *J. Comput. Graphical Stat.* 15.1, p. 246–261.

Brown, Robert G, Dirk Eddelbuettel et David Bauer (2007). "dieharder: A Random Number Test Suite". In : *URL http://www. phy. duke. edu/~ rgb/General/dieharder. php. C program archive dieharder*.

Cappé, O., E. Moulines et T. Rydèn (2005). *Inference in Hidden Markov Models*. New York : Springer-Verlag.

Carpenter, Bob et al. (2015). "Stan: a probabilistic programming language". In : *Journal of Statistical Software*.

Casella, G. et C.P. Robert (1998). "Post-processing Accept–Reject samples: recycling and rescaling". In : *J. Comput. Graph. Statist.* 7.2, p. 139–157.

Chopin, N. (2002). "A sequential particle filter method for static models". In : *Biometrika* 89, p. 539–552.

Christen, J.A. et C. Fox (2005). "Markov chain Monte Carlo using an approximation". In : *Journal of Computational and Graphical Statistics* 14.4, p. 795–810.

Christensen, Ole F., Gareth O. Roberts et Jeffrey S. Rosenthal (2003). *Scaling Limits for the Transient Phase of Local Metropolis–Hastings Algorithms*.

Cressie, N. (1993). *Spatial Statistics*. New York.

Damian, Doris, Paul D Sampson et Peter Guttorp (2001). "Bayesian estimation of semi-parametric non-stationary spatial covariance structures". In : *Environmetrics* 12.2, p. 161–178.

Datta, A. et al. (2014). "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets". In : *ArXiv e-prints*. arXiv : 1406.7343 [stat.ME].

Del Moral, P. (2004). *Feynman-Kac formulae*. Probability and its Applications. New York : Springer-Verlag, p. xviii+555.

Del Moral, P., A. Doucet et A. Jasra (2006). "Sequential Monte Carlo samplers". In : 68.3, p. 411–436.

Dempster, Arthur P (1972). "Covariance selection". In : *Biometrics*, p. 157–175.

Devroye, Luc (1986). *Nonuniform random variate generation*. New York : Springer-Verlag, p. xvi+843. isbn : 0-387-96305-7.

Diebolt, J. et Christian P. Robert (1994). "Estimation of Finite Mixture Distributions by Bayesian Sampling". In : 56, p. 363–375.

Dobra, Adrian et Alex Lenkoski (2011). "Copula Gaussian graphical models and their application to modeling functional disability data". In : *Ann. Appl. Stat.* 5.2A, p. 969–993.

Dobra, Adrian et al. (2004). "Sparse graphical models for exploring gene expression data". In : *Journal of Multivariate Analysis* 90.1, p. 196–212.

Douc, Randal et Olivier Cappé (2005). "Comparison of resampling schemes for particle filtering". In : *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*. IEEE, p. 64–69.

Doucet, Arnaud et Adam M Johansen (2009). “A tutorial on particle filtering and smoothing: Fifteen years later”. In : *Handbook of Nonlinear Filtering* 12.656-704, p. 3.

Dryden, Ian L et Kanti V Mardia (1998). *Statistical shape analysis*. T. 4. J. Wiley Chichester.

Eaton, Daniel et Kevin Murphy (2012). “Bayesian structure learning using dynamic programming and MCMC”. In : *arXiv preprint arXiv:1206.5247*.

Eddelbuettel, Dirk et Romain François (2011). “Rcpp: Seamless R and C++ Integration”. In : *Journal of Statistical Software* 40.8, p. 1–18.

Eidsvik, Jo et al. (2012). “Approximate Bayesian inference for large spatial datasets using predictive process models”. In : *Computational Statistics & Data Analysis* 56.6, p. 1362–1380.

Fearnhead, P. et B. M. Taylor (2010). “An Adaptive Sequential Monte Carlo Sampler”. In : *ArXiv e-prints*. arXiv : 1005.1193 [stat.CO].

Fearnhead, Paul et Dennis Prangle (2012). “Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation”. In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, p. 419–474.

Filippi, S. et C. Holmes (2015). “A Bayesian nonparametric approach to quantifying dependence between random variables”. In : *ArXiv e-prints*. arXiv : 1506.00829 [stat.ME].

Finley, Andrew O et al. (2009). “Improving the performance of predictive process modeling for large datasets”. In : *Computational statistics & data analysis* 53.8, p. 2873–2884.

Fox, C. et G. Nicholls (1997). “Sampling conductivity images via MCMC”. In : *The Art and Science of Bayesian Image Analysis*, p. 91–100.

Fukumizu, Kenji, Francis R Bach et Michael I Jordan (2009). “Kernel dimension reduction in regression”. In : *The Annals of Statistics*, p. 1871–1905.

Fukumizu, Kenji et al. (2007). “Kernel Measures of Conditional Dependence.” In : *NIPS*. T. 20, p. 489–496.

Ge, Hong et al. (2015). “Distributed Inference for Dirichlet Process Mixture Models”. In : *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, p. 2276–2284.

Gelfand, A.E. et S.K. Sahu (1994). “On Markov chain Monte Carlo acceleration”. In : *J. Comput. Graph. Statist.* 3.3, p. 261–276.

Genton, Marc G, William Kleiber et al. (2015). “Cross-covariance functions for multivariate geostatistics”. In : *Statistical Science* 30.2, p. 147–163.

Girolami, M. et B. Calderhead (2011). “Riemann manifold Langevin and Hamiltonian Monte Carlo methods”. In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, p. 123–214.

Golightly, A., D. A. Henderson et C. Sherlock (2014). “Delayed acceptance particle MCMC for exact inference in stochastic kinetic models”. In : *ArXiv e-prints*. arXiv : 1401.4369 [stat.CO].

Gordon, N., D. Salmond et A. F. Smith (1993). “Novel approach to nonlinear/non-Gaussian Bayesian state estimation”. In : *IEE Proc. F, Radar Signal Process.* 140, p. 107–113.

Gramacy, R. B. et D. W. Apley (2013). "Local Gaussian process approximation for large computer experiments". In : *ArXiv e-prints*. arXiv : 1303.0383 [stat.ME].

Green, Peter J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In : *Biometrika* 82.4, p. 711–732. issn : 0006-3444.

Haario, Heikki, Eero Saksman et Johanna Tamminen (2001). "An adaptive metropolis algorithm". In : *Bernoulli* 7.2, p. 223–242. issn : 1350-7265.

Hastings, W.K. (1970). "Monte Carlo sampling using Markov Chains and their applications". In : *Biometrika* 57.1, p. 97–109.

Hay, John E (1983). "Solar energy system design: The impact of mesoscale variations in solar radiation". In : *Atmosphere-Ocean* 21.2, p. 138–157.

Higdon, Dave, J Swall et J Kern (1999). "Non-stationary spatial modeling". In : *Bayesian statistics* 6.1, p. 761–768.

Higdon, Dave et al. (2002). "Space and space-time modeling using process convolutions". In : *Quantitative methods for current environmental issues* 3754.

Higdon, David (1998). "A process-convolution approach to modelling temperatures in the North Atlantic Ocean". In : *Environmental and Ecological Statistics* 5.2, p. 173–190.

Hoff, Peter D. (2007). "Extending the rank likelihood for semiparametric copula estimation". In : *Ann. Appl. Stat.* 1.1, p. 265–283.

Hoffman, Matthew D et Andrew Gelman (2014). "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo". In : *The Journal of Machine Learning Research* 15.1, p. 1593–1623.

Jeffreys, H. (1939). *Theory of Probability*. Oxford, U.K. : Clarendon Press.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. isbn : 9781466583221.

Jones, Beatrix et al. (2005). "Experiments in stochastic computation for high-dimensional graphical models". In : *Statistical Science*, p. 388–400.

Joyce, Paul et Paul Marjoram (2008). "Approximately sufficient statistics and Bayesian computation". In : *Statistical applications in genetics and molecular biology* 7.1.

Kalli, Maria, Jim E Griffin et Stephen G Walker (2011). "Slice sampling mixture models". In : *Statistics and computing* 21.1, p. 93–105.

Kinney, Justin B. et Gurinder S. Atwal (2014). "Equitability, mutual information, and the maximal information coefficient". In : *Proceedings of the National Academy of Sciences* 111.9, p. 3354–3359. eprint : http://www.pnas.org/content/111/9/3354.full.pdf.

Korattikara, A., Y. Chen et M. Welling (2013). "Austerity in MCMC land: Cutting the Metropolis-Hastings budget". In : *arXiv preprint arXiv:1304.5299*.

Kunihama, T. et D. B. Dunson (2014). "Nonparametric Bayes inference on conditional independence". In : *ArXiv e-prints*. arXiv : 1404.1429 [stat.ME].

Lenkoski, Alex (2013). "A direct sampler for G-Wishart variates". In : *Stat* 2.1, p. 119–128.

Lin, Xiwu et al. (2000). "Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV". In : *Annals of Statistics*, p. 1570–1600.

Liu, Han, John Lafferty et Larry Wasserman (2009). “The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs”. In : *J. Mach. Learn. Res.* 10, p. 2295–2328.

Liu, Han et al. (2012). “High-dimensional semiparametric Gaussian copula graphical models”. In : *Ann. Statist.* 40.4, p. 2293–2326.

Livingstone, S. et al. (2016). “On the Geometric Ergodicity of Hamiltonian Monte Carlo”. In : *ArXiv e-prints*. arXiv : 1601.08057 [stat.CO].

Marin, Jean-Michel et al. (2012). “Approximate Bayesian computational methods”. English. In : *Statistics and Computing* 22.6, p. 1167–1180.

Matsumoto, Makoto et Takuji Nishimura (1998). “Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator”. In : *ACM Trans. Model. Comput. Simul.* 8.1, p. 3–30. issn : 1049-3301. doi : 10.1145/272991.272995. url : http://doi.acm.org/10.1145/272991.272995.

McLachlan, G. J. et D. Peel (2000). *Finite Mixture Models*. New York : J. Wiley.

Mengersen, K.L. et R.L. Tweedie (1996). “Rates of convergence of the Hastings and Metropolis algorithms”. In : 24, p. 101–121.

Metropolis, N. et al. (1953). “Equations of state calculations by fast computing machines”. In : *J. Chem. Phys.* 21.6, p. 1087–1092.

Meyn, S. P. et R. L. Tweedie (1993). *Markov chains and stochastic stability*. London : Springer-Verlag London Ltd., p. xvi+ 548. isbn : 3-540-19832-6.

Mohammadi, A. et al. (2015). “Bayesian Gaussian Copula Graphical Modeling for Dupuytren Disease”. In : *ArXiv e-prints*. arXiv : 1501.04849 [stat.AP].

Mohammadi, Abdolreza et Ernst Wit (2015). *BDgraph: Bayesian Graph Selection Based on Birth-Death MCMC Approach*. R package version 2.22. url : http://CRAN.R-project.org/package=BDgraph.

Murray, I., R. Prescott Adams et D. J. C. MacKay (2010). “Elliptical slice sampling”. In : *ArXiv e-prints*. arXiv : 1001.0175 [stat.CO].

Murray, Iain, Zoubin Ghahramani et David MacKay (2012). “MCMC for doubly-intractable distributions”. In : *arXiv preprint arXiv:1206.6848*.

Neal, Peter et Gareth Roberts (2011). “Optimal Scaling of Random Walk Metropolis Algorithms with Non–Gaussian Proposals”. English. In : *Methodology and Computing in Applied Probability* 13.3, p. 583–601. issn : 1387-5841. doi : 10.1007/s11009-010-9176-9. url : http://dx.doi.org/10.1007/s11009-010-9176-9.

Neal, R. (2001). “Annealed importance sampling”. In : *Statistics and Computing* 11, p. 125–139.

Neal, R.M. (1995). *Suppressing random walks in MCMC using ordered overrelaxation*. Rapp. tech. University of Toronto, Canada : Dept. of Statistics.

— (1997). *Markov chain Monte Carlo methods based on ‘slicing’ the density function*. Rapp. tech. University of Toronto.

— (2003). “Slice sampling (with discussion)”. In : 31, p. 705–767.

— (2012). “MCMC using Hamiltonian dynamics”. In : *arXiv preprint arXiv:1206.1901*.

Neiswanger, W., C. Wang et E. Xing (2013). “Asymptotically Exact, Embarrassingly Parallel MCMC”. In : *arXiv preprint arXiv:1311.4780*.

Nishihara, Robert, Iain Murray et Ryan P Adams (2012). “Generalizing elliptical slice sampling for parallel mcmc”. In : *Neural Information Processing Systems (NIPS), Big Learning Workshop on Algorithms, Systems, and Tools for Learning at Scale*. T. 3. 3, p. 4.

Nobile, A. (1998). "A hybrid Markov chain for the Bayesian analysis of the multinomial probit model". In : *Statistics and Computing* 8, p. 229–242.

Nunes, Matthew A et David J Balding (2010). "On optimal selection of summary statistics for approximate Bayesian computation". In : *Statistical applications in genetics and molecular biology* 9.1.

Nunes, Matthew A. et Dennis Prangle (2015). "abctools: An R package for tuning Approximate Bayesian Computation analyses". Forthcoming.

Paciorek, Christopher J (2007). "Computational techniques for spatial logistic regression with large data sets". In : *Computational statistics & data analysis* 51.8, p. 3631–3653.

Perrin, Olivier et Wendy Meiring (2003). "Nonstationarity in Rn is second-order stationarity in R2n". In : *Journal of applied probability*, p. 815–820.

Perrin, Olivier et Martin Schlather (2007). "Can any multivariate gaussian vector be interpreted as a sample from a stationary random process?" In : *Statistics & probability letters* 77.9, p. 881–884.

Peskun, P. H. (1973a). "Optimum Monte-Carlo sampling using Markov chains". In : *Biometrika* 60, p. 607–612.

Peskun, P.H. (1973b). "Optimum Monte Carlo sampling using Markov chains". In : 60, p. 607–612.

Pitt, Michael, David Chan et Robert Kohn (2006). "Efficient Bayesian Inference for Gaussian Copula Regression Models". English. In : *Biometrika* 93.3, pp. 537–554.

Plummer, Martyn et al. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC". In : *R News* 6.1, p. 7–11. url : `http://CRAN.R-project.org/doc/Rnews/`.

Quinonero-Candela, Joaquin et Carl Edward Rasmussen (2005). "A unifying view of sparse approximate Gaussian process regression". In : *The Journal of Machine Learning Research* 6, p. 1939–1959.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. url : `http://www.R-project.org/`.

Ripley, Brian D. (1987). *Stochastic simulation*. New York : John Wiley & Sons Inc., p. xiv+237. isbn : 0-471-81884-4.

Robert, C.P. (2007). *: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York : Springer.

Robert, C.P. et G. Casella (2004). *Monte Carlo Statistical Methods*. second. Springer-Verlag.

Robert, C.P. et M. Titterington (1998). "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation". In : *Statistics and Computing* 8.2, p. 145–158.

Roberts, G. O., A. Gelman et W. R. Gilks (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". In : *Ann. Appl. Probab.* 7.1, p. 110–120. issn : 1050-5164.

Roberts, G. O. et O. Stramer (2002). "Langevin Diffusions and Metropolis–Hastings Algorithms." In : *Methodology and Computing in Applied Probability* 4.4, p. 337–358. issn : 1387-5841. doi : `10.1023/A:1023562417138`.

Roberts, Gareth O. et Jeffrey S. Rosenthal (1997). "Geometric ergodicity and hybrid Markov chains". In : *Electron. Comm. Probab.* 2, no. 2, 13–25 (electronic). issn : 1083-589X.

— (2001). "Markov Chains and de-initializing processes". In : *Scandinavian Journal of Statistics* 28, p. 489–504.

Roberts, G.O. et J.S. Rosenthal (2005). "Coupling and Ergodicity of Adaptive MCMC". In : *J. Applied Proba.* 44, p. 458–475.

— (2009). "Examples of Adaptive MCMC". In : *J. Comp. Graph. Stat.* 18, p. 349–367.

Roberts, G.O. et R.L. Tweedie (1996). "Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithms". In : 83, p. 95–110.

Roeder, K. et L. Wasserman (1997). "Practical Bayesian density estimation using mixtures of Normals". In : 92, p. 894–902.

Rosenthal, J. S. et G. O. Roberts (2007). "Coupling and Ergodicity of adaptive MCMC". In : *Journal of Applied Probablity* 44, p. 458–475.

Roverato, Alberto (2002). "Hyper Inverse Wishart Distribution for Non-Decomposable Graphs and Its Application to Bayesian Inference for Gaussian Graphical Models". English. In : *Scandinavian Journal of Statistics* 29.3, pp. 391–411.

Rue, H., S. Martino et N. Chopin (2009). "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations". In : 71, p. 319–392.

Ruppert, David, Matt P Wand et Raymond J Carroll (2003). *Semiparametric regression*. 12. Cambridge university press.

Sampson, Paul D et Peter Guttorp (1992). "Nonparametric estimation of nonstationary spatial covariance structure". In : *Journal of the American Statistical Association* 87.417, p. 108–119.

Sanderson, Conrad (2010). *Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Rapp. tech. NICTA.

Schmidt, Alexandra M et Anthony O'Hagan (2003). "Bayesian inference for nonstationary spatial covariance structure via spatial deformations". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.3, p. 743–758.

Schwaighofer, Anton et Volker Tresp (2002). "Transductive and inductive methods for approximate Gaussian process regression". In : *Advances in Neural Information Processing Systems*, p. 953–960.

Scott, James G et James O Berger (2006). "An exploration of aspects of Bayesian multiple testing". In : *Journal of Statistical Planning and Inference* 136.7, p. 2144–2162.

Scott, S.L. et al. (2013). "Bayes and big data: The consensus Monte Carlo algorithm". In : *EFaBBayes 250 conference* 16.

Seeger, Matthias, Christopher Williams et Neil Lawrence (2003). "Fast forward selection to speed up sparse Gaussian process regression". In : *Artificial Intelligence and Statistics 9*. EPFL-CONF-161318.

Sherlock, C. et al. (2013). "On the efficiency of pseudo-marginal random walk Metropolis algorithms". In : *ArXiv e-prints*. arXiv : `1309.7209 [stat.CO]`.

Sherlock, Chris et Gareth Roberts (2009). "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". In : *Bernoulli* 15.3,

p. 774–798. DOI : 10.3150/08-BEJ176. URL : http://dx.doi.org/10.3150/08-BEJ176.

Shestopaloff, A. Y. et R. M. Neal (2013). "MCMC for non-linear state space models using ensembles of latent sequences". In : *ArXiv e-prints*. arXiv : 1305.0320 [stat.CO].

Simpson, D. P. et al. (2014). "Penalising model component complexity: A principled, practical approach to constructing priors". In : *ArXiv e-prints*. arXiv : 1403.4630 [stat.ME].

Smola, Alex J et Peter Bartlett (2001). "Sparse greedy Gaussian process regression". In : *Advances in Neural Information Processing Systems 13*. Citeseer.

Snelson, Edward et Zoubin Ghahramani (2005). "Sparse Gaussian processes using pseudo-inputs". In : *Advances in neural information processing systems*, p. 1257–1264.

Stein, Michael L, Zhiyi Chi et Leah J Welty (2004). "Approximating likelihoods for large spatial data sets". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.2, p. 275–296.

Stephens, M. (1997). "Bayesian Methods for Mixtures of Normal Distributions". Thèse de doct. Department of Statistics.

Strid, I. (2010). "Efficient parallelisation of Metropolis–Hastings algorithms using a prefetching approach". In : *Computational Statistics & Data Analysis* 54.11, p. 2814–2835.

Stroud, J. R., M. L. Stein et S. Lysen (2014). "Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice". In : *ArXiv e-prints*. arXiv : 1402.4281 [stat.CO].

Tierney, L. (1994). "Markov chains for exploring posterior distributions (with discussion)". In : 22, p. 1701–1786.

Tierney, L. et A. Mira (1998). "Some adaptive Monte Carlo methods for Bayesian inference". In : *Statistics in Medicine* 18, p. 2507–2515.

Tierney, Luke (1998). "A note on Metropolis-Hastings kernels for general state spaces". In : *Ann. Appl. Probab.* 8.1, p. 1–9. ISSN : 1050-5164.

Uhler, Caroline, Alex Lenkoski et Donald Richards (2014). "Exact formulas for the normalizing constants of Wishart distributions for graphical models". In : *arXiv preprint arXiv:1406.4901*.

Vecchia, Aldo V (1988). "Estimation and model identification for continuous spatial processes". In : *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 297–312.

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices". In : *Comm. Statist.* 36, p. 45–54.

Wang, Hao, Sophia Zhengzi Li et al. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions". In : *Electronic Journal of Statistics* 6, p. 168–198.

Wang, X. et D.B. Dunson (2013). "Parallel MCMC via Weierstrass Sampler". In : *arXiv preprint arXiv:1312.4605*.

Wong, Nelson, Sheelagh Carpendale et Saul Greenberg (2003). "Edgelens: An interactive method for managing edge congestion in graphs". In : *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*. IEEE, p. 51–58.

Xia, Gangqiang et Alan E Gelfand (2005). "Stationary process approximation for the analysis of large spatial datasets". In : *ISDS, Duke University*.

Zhang, Kun et al. (2012). “Kernel-based conditional independence test and application in causal discovery”. In : *arXiv preprint arXiv:1202.3775*.

Zhou, Y., A. M Johansen et J. A. Aston (2013). “Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach”. In : *ArXiv e-prints*. arXiv : 1303.3123 [stat.ME].

Résumé

Cette thèse présente des contributions à la littérature des méthodes de Monte Carlo utilisé dans l'analyse des modèles complexes en statistique Bayésienne; l'accent est mis à la fois sur la complexité des modèles et sur les difficultés de calcul.
Le premier chapitre élargit Delayed Acceptance, une variante computationellement efficace du Metropolis--Hastings, et agrandit son cadre théorique fournissant une justification adéquate pour la méthode, des limits pour sa variance asymptotique par rapport au Metropolis--Hastings et des idées pour le réglage optimal de sa distribution instrumentale.
Nous allons ensuite développer une méthode Bayésienne pour analyser les processus environnementaux non stationnaires, appelées Expansion Dimension, qui considèr le processus observé comme une projection depuis une dimension supérieure, où l'hypothèse de stationnarité pourrait etre acceptée.
Le dernier chapitre sera finalement consacrée à l'étude des structures de dépendances conditionnelles par une formulation entièrement Bayésienne du modèle de Copule Gaussien graphique.

Abstract

This thesis presents contributions to the Monte Carlo literature aimed toward the analysis of complex models in Bayesian Statistics; the focus is on both complexity related to complicate models and computational difficulties.
We will first expand Delayed Acceptance, a computationally efficient variant of Metropolis--Hastings, to a multi-step procedure and enlarge its theoretical background, providing proper justification for the method, asymptotic variance bounds relative to its parent MH kernel and optimal tuning for the scale of its proposal.
We will then develop a flexible Bayesian method to analyse nonlinear environmental processes, called Dimension Expansion, that essentially consider the observed process as a projection from a higher dimension, where the assumption of stationarity could hold.
The last chapter will finally be dedicated to the investigation of conditional (in)dependence structures via a fully Bayesian formulation of the Gaussian Copula graphical model.

Mots Clés

Methodès de Monte Carlo, Monte Carlo par chaînes de Markov, méthodes de Monte Carlo séquentielles, modélisation Bayésienne, processus Gaussiens, statistique environnementale, dépendances conditionnelles

Keywords

Monte Carlo methods, Markov Chain Monte Carlo, Sequential Monte Carlo, Bayesian modeling, Gaussian Processes, environmental statistics, conditional dependence