Probabilistic models for image processing: applications in vascular surgery

Hugo Gangloff

Université de Strasbourg

École doctorale MSII

Laboratoire ICube - UMR 7357

THÈSE présentée par :

Hugo GANGLOFF

soutenue le :

15 décembre 2020

pour l'obtention du grade de : **docteur de l'Université de Strasbourg**

Discipline / Spécialité : signal, image, automatique, robotique

Probabilistic models for image processing: applications in vascular surgery

THÈSE dirigée par :

Christophe Collet	Professeur, Université de Strasbourg	Directeur de thèse
Nabil Chakfé	Professeur, Université de Strasbourg	Co-directeur de thèse

RAPPORTEURS :

Isabelle Bloch	Professeur, Télécom Paris
Cédric Richard	Professeur, Université Nice Sophia Antipolis

AUTRES MEMBRES DU JURY :

Wojciech Pieczynski	Professeur, Télécom SudParis
Emmanuel Monfrini	Maître de Conférences, Télécom SudParis

In the morning all the birds are chirping and everything is going on. I grab a coffee and get to work. Then when I get tired, I take off. If I run out of ideas, I go for a run, and things fall into place.

— Bernd Heinrich, interviewed by *Runner's World*.

Acknowledgements

Je te tiens tout d'abord à remercier Isabelle Bloch et Cédric Richard d'avoir accepté d'être les rapporteurs de mon travail. Je remercie également Wojciech Pieczynski d'avoir présidé mon jury de thèse.

Merci à Mauro Gargiulo (Professeur à l'Université de Bologne) d'avoir honoré le jury de sa présence, en qualité de membre invité.

Un grand merci à mes directeurs de thèse, Christophe Collet et Nabil Chakfé, et à mon encadrant de thèse, Emmanuel Monfrini. Depuis le début, ils m'ont fait confiance, m'ont encouragé et soutenu. Merci de m'avoir ouvert les portes des laboratoires ICube et Geprovas pour ces trois intenses années de thèse qui viennent de s'écouler...

Merci à Jean-Baptiste Courbot, qui m'a transmis une partie de ses connaissances et de son expérience d'enseignant-chercheur, en donnant beaucoup de son temps et de son énergie.

Merci à ma famille. En particulier, la présence sans faille de mes deux mamies et de mes parents, depuis le tout début, a été essentielle à la réussite de cette thèse. Merci à ma sœur, qui démontre que le succès peut prendre une infinité de formes.

Finalement, je veux dire merci à toutes les personnes que j'ai croisées pendant ces dernières années, je suis certain d'avoir appris quelque chose de chacune de nos rencontres. C'était un plaisir de passer du temps à vos côtés et d'avancer, dans la thèse ou dans toutes les autres aventures qu'offre la vie. Donner une liste de noms me semble risqué et grandement insatisfaisant ; j'ai confiance dans le fait que vous vous reconnaîtrez...

Contents

List of Acronyms

Probabilistic Graphical Models

CN	Correlated Noise.
CRF	Conditional Random Field.
DAG	Directed Acyclic Graph.
DGM	Directed Graphical Model.
fcCRF	fully-connected Conditional Random Field.
GMM	Gaussian Mixture Model.
GMRF	Gaussian Markov Random Field.
GMRF	Gaussian Random Field.
GPMF	Gaussian Pairwise Markov Field.
HMC	Hidden Markov Chain.
HMF	Hidden Markov Field.
HMM	Hidden Markov Model.
HMT	Hidden Markov Tree.
IN	Independent Noise.
MC	Markov Chain.
MoE	Mixture of exponentials.
MoG	Mixture of Gaussians.
MRF	Markov Random Field.
MT	Markov Tree.
P-IN	Potts-Independent Noise.
PMF	Pairwise Markov Field.
SBN	Spatial Bayes Network.
STMT	Spatial Triplet Markov Tree.
UGM	Undirected Graphical Model.
UN	Uncorrelated Noise.

Others

AUC	Area Under Curve.
Ba	Background.
BHC	Beam Hardening Correction.
BIC	Bayesian Information Criterion.
BM3D	Block-Matching 3D.

CNN	Convolutional Neural Network.
CT	Computed Tomography.
CVD	Cardiovascular Disease.
DL	Deep Learning.
EM	Expectation Maximization.
FB	Forward Backward.
FN	False Negative.
FP	False Positive.
FT	Fatty Tissue.
GC	Graph-Cut.
IPET	Iterative Parameter Estimation for Trees.
KL	Kullback Leibler.
LLS	Linear Least Square.
MAP	Maximum A Posteriori.
MAR	Metal Artifact Reduction.
MCMC	Markov Chain Monte Carlo.
mCT	micro-Computed Tomography (also **microCT**).
MF	Mean Field.
ML	Maximum Likelihood.
MPM	Maximum Posterior Mode.
MRI	Magnetic Resonance Imaging.
NC	Nodular Calcification.
PET	Positron Emission Tomography.
PR	Precision Recall.
pyIS	pyImSegm.
RNN	Recurrent Neural Network.
ROC	Receiver Operating Characteristic.
RORPO	Ranking the Orientation Responses of Path Operator.
SA	Simulated Annealing.
SC	Sheet Calcification.
SEM	Stochastic Expectation Maximization.
SH	Specimen Holder.
SPD	Symmetric Positive Definite.
SPE	Stochastic Parameter Estimation.
ST	Soft Tissue.
T-Gibbs	Tempered-Gibbs.
UD	Upward Downward.
VI	Variational Inference.

List of Notations

General

a	Light lower case letters represent scalars or functions.
$\boldsymbol{a}$	Bold lower case letters represent vectors.
A	Capital letters represent matrices.
$\propto$	Equality up to a constant factor.
δ	Kronecker function.
$\mathbb{1}$	Indicator function.
$\lvert \cdot \rvert$	Cardinal of a set or absolute value of scalar.
$\lVert . \rVert$	A distance function (Euclidean, $\ell 1$, ...).
$\det(\cdot)$	Determinant of a matrix.
$.^T$	Transpose operator on vectors or matrices.
$\times$	Product (explicit notation) or Cartesian product.
$\bullet$	Element-wise power.
$\odot$	Element-wise multiplication.
DFT2$(\cdot)$	Bidimensional Discrete Fourier Transform.
IDFT2$(\cdot)$	Bidimensional Inverse Fourier Transform.
$\Re(\cdot)$	Real part of complex number.
$\Im(\cdot)$	Imaginary part of complex number.
$\cup$	Set union operator.
$\cap$	Set intersection operator.
$\wedge$	Logical *and* operator.
$\vee$	Logical *or* operator.

Graph Theory

$\mathcal{G}$	A graph.
$\mathcal{E}$	Set of edges of a graph.
$\mathcal{S}$	Set of vertices of a graph.
$\mathcal{N}_s$	Neighborhood of site s.
$\mathcal{P}(s)$	Set of parents of s.
s^-	A parent vertice of s.
$\bar{\mathcal{S}}$	Set of vertices with at least a parent vertice.
$\boldsymbol{s}^{++}$	The set of all descendants of s.
$\boldsymbol{s}^{--}$	The set of all ascendants of s.

Probability Theory

X	Capital letters represent random variables.
x	The realization of the random variable X.
$\boldsymbol{X}$	Bold capital letters represent random vectors.
$\mathbb{E}[X]$	Expectation of random variable X.
$\mathbb{KL}(q\|\|p)$	Kullback Leibler divergence between distributions q and p.
P_X	The law of random variable X.
$p(x)$	-Discrete case: the probability $p(\{X = x\})$. -Continuous case: the probability density of random variable X.
$\mathcal{N}(\cdot, \cdot)$	Normal (Gaussian) density function. (first argument is the mean and second the variance).
$\sim$	Indicates the law followed by a random variable.

Résumé étendu

Introduction

Les maladies cardiovasculaires sont la première cause de mortalité dans le monde. Les coûts de ces maladies pour les sociétés sont humain et économique, ils sont déjà importants (chiffrés en plusieurs centaines de milliards de dollars) et ne cessent de croître[1] [2]. Le vieillissement et l'augmentation de la population mondiale sont des facteurs contribuant à l'augmentation de ces pathologies. Pour y faire face, la recherche est poussée à mieux connaître ces maladies et développer de nouveaux traitements plus efficaces.

Dans ce contexte, les approches modernes en chirurgie vasculaires sont basées sur des opérations mini-invasives guidées par l'image et sur l'implantation de biomatériaux (dont l'élément le plus connu est le stent). On parle de chirurgie endovasculaire. Ces nouveaux types d'intervention offrent une alternative à la chirurgie ouverte classique. Cependant, la chirurgie endovasculaire, déjà largement pratiquée, soulève des questions auxquelles la recherche n'a actuellement que des réponses trop partielles. Ces questions concernent aussi bien le comportement mécanique des biomatériaux dans le corps du patient et leur implantation par le chirurgien, que la correction d'images scanner illisibles et l'utilisation per-opératoire de ces images.

Les problématiques présentées ci-dessus relèvent d'une recherche fortement pluri-disciplinaire. Dans cette thèse, nous présentons des contributions aux traitement des images médicales. Ces contributions ont pour but de permettre la création de nouveaux outils qui doivent servir à d'autres chercheurs (biomécaniciens, chirurgiens, spécialistes du textile, radiologues, histopathologistes, ...) pour, *in fine*, mieux connaître les maladies cardiovasculaires, leur traitement et leur prévention. Nos contributions sont liées à une principale application : celle de la segmentation des images médicales.

Les images que nous traitons dans la thèse sont de type *Computed Tomography* (CT) ou *micro-Computed Tomography* (mCT). Les principaux défis dans la segmentation des images médicales en chirurgie vasculaire concernent la complexité et les dégradations subies par images à rayons-X lorsque celles-ci contiennent un biomatériau métallique, la disponibilité limitée des données et la taille des données à traiter. Dans la section suivante, nous décrivons plus en détails ces problèmes, en mettant en avant les méthodes graphiques probabilistes utilisées pendant la thèse pour les traiter. La Figure 0.2 illustre un cas typique

[1] https://healthmetrics.heart.org/wp-content/uploads/2017/10/Cardiovascular-Disease-A-Costly-Burden.pdf

[2] https://www.bhf.org.uk/what-we-do/our-research/heart-statistics/heart-statistics-publications/cardiovascular-disease-statistics-2017

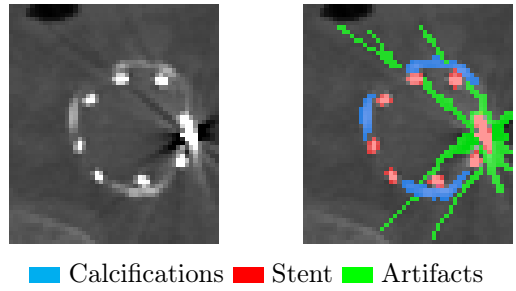


Figure 0.2.: Coupe de CT scan typique des données à traiter. Le stent se trouve dans un environnement complexe, entouré de calcifications et d'artéfacts. Notons qu'un bruit corrélé est une modélisation pertinente de ce phénomène.

des données de chirurgie vasculaire que nous souhaitons pouvoir segmenter.

Méthodes

Dans les modèles graphiques probabilites appliqués à la segmentation des images, des sommets du graphe sont associés à des variables aléatoires qui représentent l'image observée, et d'autres sommets sont associés à des variables aléatoires qui correspondent aux classes dans l'image segmentée. Les arêtes du graphe représentent une relation de dépendance entre les deux variables reliées. Les arêtes peuvent être orientées ou non-orientées ce qui modifie la formulation probabiliste du modèle (Murphy, 2012).

Les modèles graphiques probabilistes les plus connus pour la segmentation sont ceux de la famille des modèles de Markov cachés (Baum and Petrie, 1966). Parmi les modèles de Markov cachés les plus populaires, nous notons les chaînes et arbres de Markov (Baum and Petrie, 1966) (Laferté et al., 2000) (modèles graphiques orientés) et les champs de Markov (Kato and Zerubia, 2012) (modèle graphique non-orienté). Nous étudions largement ces modèles et leurs extensions : les modèles de Markov cachés couples et triplets (Pieczynski and Tebbache, 2000) (Lanchantin et al., 2011) (Gorynin, Gangloff, et al., 2018). Les modèles probabilistes markoviens offrent un moyen simple et intuitif d'introduire de la dépendance entre les pixels de l'image, ils sont une réponse pertinente aux problématiques posées par les images étudiées pendant la thèse.

Nous passons maintenant en revue les principales notions théoriques sur les modèles probabilistes graphiques, à la lumière des problématiques propres à notre application.

Modèles probabilistes et segmentation d'images bruitées

Une problématique particulière à laquelle nous devons faire face est celle des artéfacts. Lors de l'acquisition des images à rayons-X, les biomatériaux mé-

talliques présents dans le corps du patient vont intéragir avec les rayons-X ce qui va créer, à la sortie de l'algorithme de reconstruction CT, de forts artéfacts. Ces derniers empêchent de discerner facilement l'anatomie environnante. D'un point du vue du traitement du signal les artéfacts peuvent être modélisés comme du bruit spatialement corrélé. Ce type bruit peut être naturellement modélisé dans certains modèles graphiques probabilistes comme les nouveaux modèles de Markov couples et triplets (Gorynin, Gangloff, et al., 2018) qui sont présentés dans la thèse. En parallèle, les corrélations spatiales sont classiquement étudiées dans les modèles de type Gaussian Markov Random Fields (Rue and Held, 2005) (GMRF) qui sont également étudiés dans notre travail.

Modèles probabilistes et disponibilité des données

Les biomatériaux sont récents et leur étude par l'imagerie après explantation n'a débutée que récemment. Ainsi, la quantité de données est encore relativement restreinte et les données sont, pour la plupart, non-annotées[3].

Dans de tels cas, les approches *non-supervisées* sont privilégiées, voire nécessaires. Les modèles probabilistes graphiques sont souvent utilisés car ils bénéficient d'algorithmes relativement efficaces dans les cas où les données sont rares ou manquantes. En contexte non-supervisé, les paramètres des distributions de probabilité des variables aléatoires du modèle graphique doivent être estimés à l'aide seulement de l'observation. Dans ce contexte, les approches de type Expectation-Maximization (Dempster et al., 1977) (EM) qui ont pour objectif de maximiser la vraisemblance des données sont les plus connues. Plusieurs versions de ces algorithmes existent et elles sont déclinables pour les modèles graphiques orientés et non-orientés (Celeux et al., 1995) (Tieleman, 2008). Nous nous intéressons, dans notre travail, à ce type d'approches non-supervisées et leur procédure d'estimation de paramètres.

Le cas où une base de données annotées est disponible (approche *supervisée*) est aussi étudié dans la thèse. La segmentation des images bidimensionnelles est faite par un réseau de neurones convolutionnel (Srinidhi et al., 2019). Ces approches qui relèvent de l'apprentissage profond sont introduites dans la thèse. En effet, elles donnent les meilleurs résultats dans un grand nombre de problèmes de segmentation supervisée et sont devenues incontournables dans le domaine. En revanche, l'étape ultérieure qui correspond à la reconstruction d'images segmentées tridimensionnelles est plus complexe. Elle nécessite des approches plus fines, notamment des associations de modèles d'apprentissage profond et de modèles probabilistes graphiques, que nous étudions également dans le manuscrit (Ben-Cohen et al., 2016) (Kamnitsas et al., 2017) (Novikov et al., 2018).

[3] Les travaux de cette thèse utilisent la base de données d'explants du laboratoire Geprovas (https://geprovas.org) qui a vu le jour grâce à son programme de collecte et d'analyse d'explants.

Modèles probabilistes et le coût de l'inférence

Les images de chirurgie vasculaire de type mCT (qui offrent une bien meilleure résolution spatiale que les images CT) sont de très grande taille. Les tâches d'inférence dans les modèles graphiques probabilistes associés à des images de grande taille peuvent être très coûteuses en temps de calcul, voire même infaisables.

De nombreuses approches sont alors fondées sur une approximation de l'étape d'inférence. Par exemple, la technique de l'*inférence variationnelle* approche la distribution de probabilité pour laquelle l'inférence est coûteuse par une distribution simplifiée. Les recherches sur cette approche sont aujourd'hui très actives (C. Zhang et al., 2018) et l'inférence variationnelle est étudiée en détail dans notre travail pour des modèles graphiques orientés et non-orientés.

Les champs de Markov cachés sont les modèles graphiques probabilistes les plus populaires pour la segmentation non-supervisée d'images médicales. Or l'inférence dans les modèles graphiques non-orientés (auxquels appartiennent les champs de Markov) ne peut se faire, sauf dans des cas particuliers, de manière directe. L'inférence dans ces modèles repose sur des calculs indirects (qui peuvent être approximants ou non) comme l'approche itérative de l'échantillonneur de Gibbs (S. Geman and D. Geman, 1984). Les chaînes de Markov cachées peuvent alors être une alternative judicieuse car leur structure intrinsèquement unidimensionelle et les possibilités d'inférence exacte offrent des coûts calculatoires faibles relativement aux approches classiques d'inférence dans les champs. Ces propriétés sont rencontrées dans (Bricq et al., 2008) (Courbot, Rust, et al., 2015) et vues dans cette thèse également.

Contributions

Après avoir introduit les principales notions sur les modèles graphiques probabilistes, nous donnons ici les principales contributions de la thèse.

Le modèle Gaussian Pairwise Markov Fields

Nous présentons un nouveau modèle de champs de Markov couples cachés, nommé *Gaussian Pairwise Markov Fields* (GPMF), qui permet la segmentation non-supervisée d'images corrompues par du bruit spatialement corrélé modélisé par un GMRF (Rue and Held, 2005). Nous étudions, théoriquement et expérimentalement, dans quelle mesure ce modèle est une généralisation du modèle de champ de Markov caché classique. Nous proposons également un algorithme stochastique itératif pour l'estimation non-supervisée des paramètres. Le modèle est utilisé pour segmenter la matière organique dans des images mCT touchées par des artéfacts de stents. La Figure 0.3 illustre cet axe de recherche.

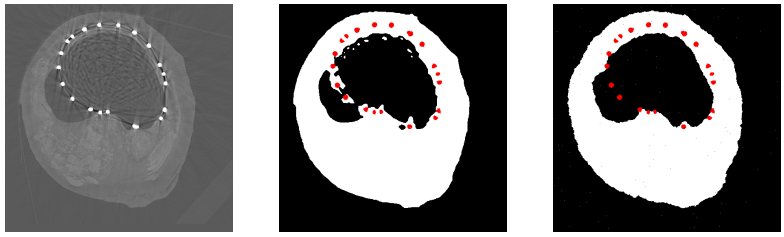


Figure 0.3.: Segmentation non-supervisée de la matière organique. De gauche à droite : le mCT observé, la segmentation par champ de Markov caché classique, la segmentation par le nouveau modèle GPMF. La résolution des artéfacts par le modèle GPMF est bien meilleure, elle réduit le nombre de faux-positifs et faux-négatifs.

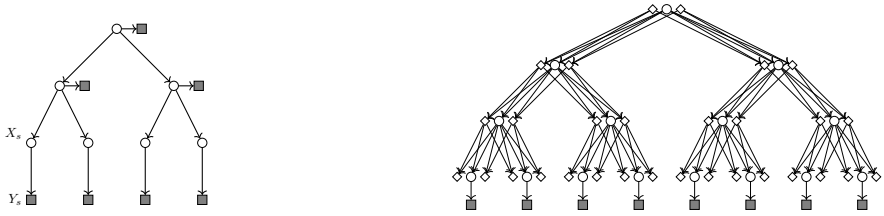
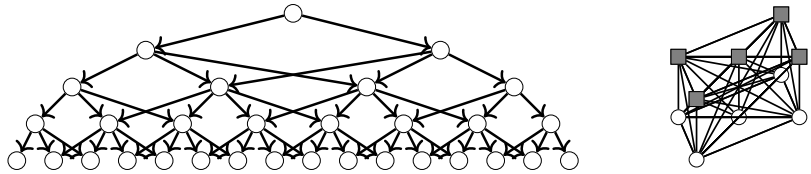


Figure 0.4.: Arbre de Markov caché classique (*gauche*) et STMT (*droite*). Les arbres sont de taille 4. Les variables cachées sont en rond, les variables visibles en carré grisé et les variables auxiliaires en losange. Nous notons les corrélations directes bien plus riches dans le modèle STMT où l'inférence reste cependant exacte.

Le modèle Spatial Triplet Markov Tree

Dans cette thèse, nous développons également un nouveau modèle d'arbre de Markov triplet nommé Spatial Triplet Markov Tree (STMT) basé sur (Courbot, Monfrini, et al., 2018). Ce modèle intègre des variables aléatoires auxiliaires afin d'enrichir les possibilités de modélisations. Le modèle STMT est une généralisation des arbres de Markov caché classique.

De plus nous étudions ses liens avec les champs de Markov cachés classiques. Le nouveau modèle semble en effet exhiber des corrélations semblables aux champs mais offre des possibilités d'inférence exacte grâce à la structure d'arbre. Une telle propriété a pour principal avantage d'éviter le recours à des procédures itératives souvent approximantes, plus longues et induisant des pertes de précision. Nous étudions les corrélations à l'intérieur du nouveau modèle STMT avec la technique de l'inférence variationnelle à variables aléatoires auxiliaires. Le modèle STMT est aussi utilisé dans le cadre de la segmentation de calcifications sur des images mCT. La Figure 0.4 illustre les modèles d'arbres de Markov.



(a) (b)

Figure 0.5.: Exemples de modèles probabilistes graphiques plus généraux. *Gauche :* Spatial Bayes Network. *Droite :* fully-connected Conditional Random Field. Les ronds blancs représentent une variable cachée et les carrés gris représentent une variable observée.

Inférence variationnelle dans des modèles plus complexes

Nous étudions des modèles plus généraux dans le sens où plus de dépendances directes sont modélisées entre les variables aléatoires.

Pour les modèles graphiques orientés, alors que tous les modèles de Markov cachés classiques sont associés à des graphes ne présentant ni cycles ni semi-cycles, nous proposons un nouveau modèle qui contient des semi-cycles appelé Spatial Bayes Network. La représentation graphique de ce modèle est donnée Figure 3.9. L'introduction de semi-cycles rend l'inférence beaucoup plus complexe et nous résolvons ce problème en ayant recours à l'inférence variationnelle.

Pour les modèles graphiques non-orientés, il est possible de complexifier le modèle en ajoutant des dépendances conditionnelles entre les variables qui agrandissent les voisinages. Le modèle des fully-connected Conditional Random Fields est un modèle probabiliste populaire en segmentation d'images qui est associé à un graphe totalement connecté (voir Figure 0.5b). Nous étudions à nouveau plusieurs techniques d'inférence variationnelle afin de pouvoir mener l'inférence dans ce modèle.

Segmentation de stents

Dans les images CT, les stents peuvent apparaître fortement déformés à cause des artéfacts qui les entourent. Nous proposons une modélisation basée sur les chaînes de Markov cachées et un modèle de bruit particulier qui permet de segmenter finement le stent dans un environnement complexe (tel que celui de la Figure 0.2). Notre approche opère de manière non-supervisée et est entièrement automatique grâce à l'algorithme d'estimation des paramètres que nous développons également. Une illustrations des résultats de cet algorithme est donnée dans la Figure 0.6.

Segmentation histologique tridimensionnelle d'artères

Dans une dernière partie, nous mettons au point un protocole pour la création de la première base de données d'images annotées d'artères explantées

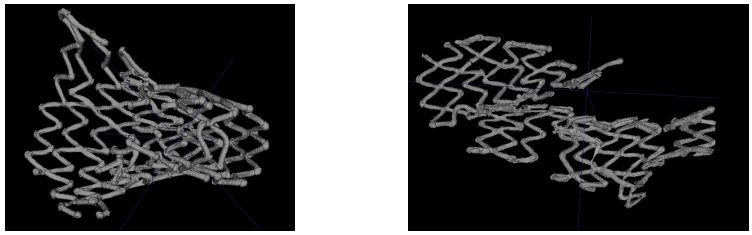


Figure 0.6.: Segmentation d'un stent explanté par notre approche.

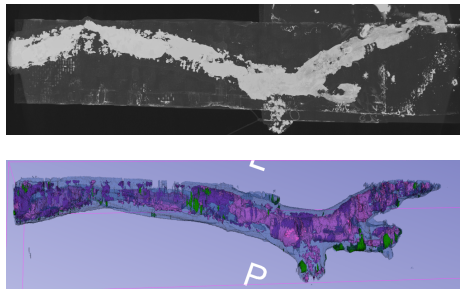


Figure 0.7.: Segmentation tridimensionnelle d'une artère pathologique. *Haut :* Image rayons-X originale. *Bas :* Notre segmentation histologique.

touchées par l'athérosclérose. Ensuite, nous développons un réseau de neurones convolutionnel de type U-Net (Ronneberger et al., 2015) afin de procéder à la segmentation bidimensionnelle histologique des images. L'objectif de ce projet unique est de permettre une première analyse histologique de l'artère uniquement à l'aide du scanner.

Une reconstruction tridimensionnelle propre est obtenue grâce à un post-traitement basé sur une inférence variationnelle dans un modèle de champs aléatoires conditionnels (Krähenbühl and Koltun, 2011). Nous étudions également une amélioration de cette dernière procédure d'inférence variationnelle utilisant les chaînes de Markov. Un exemple de segmentation tridimensionnelle d'une artère est visible dans la Figure 0.7.

Conclusion

Les modèles développés au cours de la thèse apportent des contributions sur des enjeux cruciaux de la modélisation des images par modèles graphiques probabilistes tout en répondant à des problématiques modernes de la chirurgie vasculaire.

Introduction

Advances in vascular surgery and current issues

Cardiovascular Diseases

Cardiovascular Diseases (CVDs) are a major cause of mortality in the world, and particularly in developed countries, with an increasing human and monetary cost for societies. In 2016, approximately 17.9 million people died worldwide from CVDs, which represents 31% of all deaths[4]. In 2015, 49 million people (9.6% of the population) in the European Union where living with a CVD, the total costs are estimated to €210 billion with €111 billion of direct health care costs and €99 billion indirect costs (productivity loss and informal care of people)[5]. A major cause of the rise of CVDs is also the increasing lifetime: this causes new diseases, related to old age, to be seen more widely and needing to be treated. But CVDs are also often linked with life hygiene: smoking, sedentarity and nutrition are habits with direct effects on health in general and more particularly on the cardiovascular system[6].

CVDs include a number of heart and blood vessel conditions. Among the most prevalent conditions, one can note ischemic heart disease, stroke and high blood pressure. These conditions often relate to another condition called *atherosclerosis*. Atherosclerosis is a disease of the arterial wall (not restricted to coronary arteries). It consists in the formation of plaques, called atheromatous plaques, which mostly contain lipids, macrophage cells, connective tissues and calcium (Rafieian-Kopaei et al., 2014). The deposition of calcium contributes to the plaque hardening and the sclerosis of the artery, but it is also what makes the plaque clearly visible on X-ray images. Therefore, one commonly refers to the atheromatous plaques as *calcifications* since the latter are the most easily identifiable elements of the plaques. Those plaques may grow and/or break which disturbs or, in the most severe cases, stops the blood flow (total occlusion of the artery, rupture of the artery, *etc.*). These situations are medical emergencies (Rosenfeld, 2000). However, in some cases, atherosclerotic plaques can remain stable and/or regress (Dave et al., 2013). Much is yet to discover about this phenomenon which is summarized in Figure 0.8.

In this thesis, our research will be applied to pathologies affecting the arteries of the lower limbs, as well as their surgeries and their treatments, which are described in the next section. More precisely, we will focus on the superficial

[4] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[5] http://www.ehnheart.org/cvd-statistics.html
[6] https://healthmetrics.heart.org/wp-content/uploads/2017/10/Cardiovascular-Disease-A-Costly-Burden.pdf

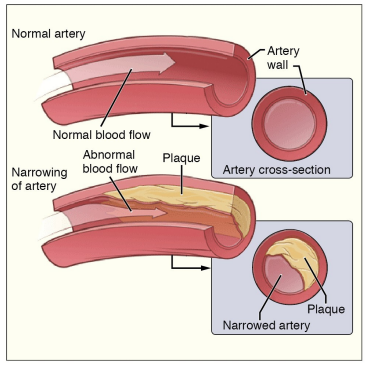


Figure 0.8.: Summary of the atherosclerotic process. From (CNX, 2016).

femoral artery and the popliteal artery which are located in the human body in Figure 0.9. The choice of the femoropopliteal arterial segment was made because it is still the most challenging area for treatment, it is also the most often affected segment in the arterial tree.

Remark: Strictly speaking, *cardiovascular* refers to the heart and its main arteries and veins. Hence, we will use the adjective *vascular* to refer to the arteries of the lower limb. However, the problematics and pathologies that affect the vessels are identical or very similar.

Remark: Veins will not be mentioned in this thesis since they do no develop atherosclerosis (G. Saul and Gerard, 1991).

Vascular surgery: open and endovascular approaches

> *Vascular surgery began with the first attempt to control a bleeding vessel.* (S. G. Friedman, 2005)

The traditional approach in vascular surgery, often called *open surgery*, relies in opening the patient body up to the affected vessel. The vessel is then repaired or replaced. This approach can lead to high-risk interventions and might not be tolerated by already weak patients such as old patients or patients with a chronic illness. In such cases, the latter have increased chances of complications, including death, following the intervention.

A more recent approach is *endovascular* surgery, in which new interventions and protocols are regularly created since the last two decades of the 20^{th} century. In this interventional technique, the vessel is reached and treated from the inside. With an incision in a peripheral artery, the surgical tools and treatments are brought to the desired location, by the inside of the blood vessel, following the patient's vascular tree. The treatment often includes a *biomaterial*. Biomaterials are introduced in the next section.

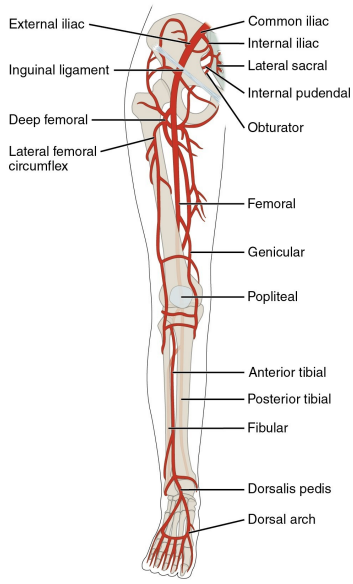


Figure 0.9.: Arteries of the lower limbs (anterior view). From (CNX, 2016).

Endovascular interventions rely heavily on imaging to see inside the patient which is not opened: the terms *image-guided surgery* and *mini-invasive* are then often associated with endovascular surgery. This surgical technique is often a solution proposed to patients who cannot undergo open surgery. The risk of medical complications and the patient's traumatism following an endovascular procedure is reduced (Nichols and Wei, 2011). While endovascular techniques have classically been linked with a higher monetary cost than open surgery procedures, the gradually decreasing cost of the treatment and of the patient's length of stay becomes an additional asset for endovascular procedures (Sternbergh III and Money, 2000) (Brinster et al., 2020).

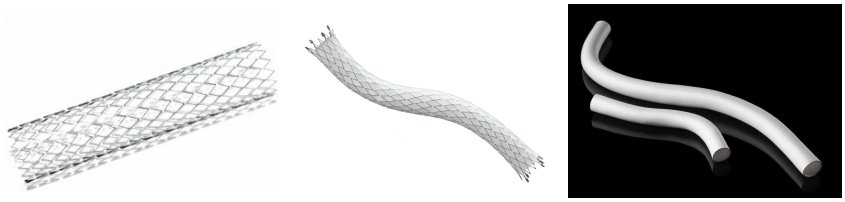
Note that interventional techniques, either from endovascular or open surgery, are constantly improved and updated. A complete reference to endovascular history, techniques and biomaterials is available in (Jing et al., 2018).

Endovascular surgery is a modern approach which raises new questions on biomaterials and new challenges on the per-operative use of imagery during surgical interventions and at all the other stages of the patient's treatment. Those questions are the motivations behind this thesis.

Biomaterials and related problematics

Biomaterials

> *[A biomaterial is] a systemically and pharmacologically inert substance designed for implantation within or incorporation with living systems.* (J. Park and Lakes, 2007)



(a) (b) (c)

Figure 0.10.: Common biomaterials in vascular surgery: (a) a stent manufactured by Cook[8], (b) a stentgraft manufactured by Bard[9], (c) a vascular prosthesis manufactured by Gore[10].

Biomaterials are found in many medical fields: knee or hip prosthesis in orthopedic surgery, intraocular lens in ocular surgery, heart valve in cardiac surgery, *etc.* We focus here in biomaterials in vascular surgery. Following (Jing et al., 2018), we list the most common biomaterials:

- Stents are metallic biomaterials widely used to solve problems such as stenosis or dissection of the artery. Once they have been brought to the location of the lesion via the intravascular path, they are released. They are either self-expandable or expanded with a surgical balloon. A stent is illustrated in Figure 0.10a.

- Stentgrafts are distinguished from the (bare) stents described above by the covering (most of the time in polytetrafluoroethylene or polyethylene terephthalate[7]) added over the stent. Stentgrafts are used mainly in aneurysm repairs or aortic dissections for their capacity to seal the vessel wall. The Figure 0.10b depicts a stentgraft.

- In the most severe cases of lesions, such as complete stenosis or occlusion, the vessel can be replaced with vascular protheses. They may originate from an autogenous vein (the ideal case for biocompatibility) or from an artificial prosthesis which can again be made of PTFE or PET. Note that, as opposed to stents and endografts, an implantation of a vascular prosthesis involves an open-surgery procedure. However they offer better performance and stability than stentgrafts in terms of biocompatibility, fracture resistance and ability to seal the vessel wall. A vascular prosthesis can be seen in Figure 0.10c.

To treat a given lesion, the choice of the model and size of the biomaterial depends mainly on the patient's imaging analysis and the surgeon's experience.

[7] Polytetrafluoroethylene (PTFE) and Polyethylene Terephthalate (PET) are thermoplastics. Note that their biocompatibility with the human body is debated (Nabil Chakfé et al., 2020).

Figure 0.11.: Explanted stent from a superficial femoral artery. The stent itself can be seen, it is encapsulated in a biological material (artery).

Remark: In this thesis, we will mainly work with medical images of *explants*. In vascular surgery, explanting is the act of removing an implanted biomaterial or a part of a blood vessel. Note that when a biomaterial is explanted, it is almost always encapsulated in some biological material. An explant can be seen in Figure 0.11.

Problematics

Biomaterials in vascular surgery are recent treatments. While these devices benefit from constant improvements and are increasingly implanted, many questions remain unanswered:

- The mechanical behavior of the implanted stents are almost unknown. Simulations, or *in vitro* experiments, only partially reflect the reality of the constraints applied on a biomaterial implanted in a living body. *In vivo* mechanical data are still very much ignored and biomaterial manufacturers would greatly take advantage of such data. For example, the optimal stent design is still much debated (Raffort et al., 2020).
- The choice of the endovascular procedure is still very dependent on the experience of the surgeon. Objective mathematical tools and databases to guide clinicians are still lacking (Ohana et al., 2014).
- In many cases, the cause(s) of failure of the vascular biomaterials are ignored or unclear. Why did the stentgraft covering tear? Why did the stent fracture? This highlights the need of new analysis tools for explanted biomaterials to investigate biomaterial failures (Chakfé and Heim, 2017) (Lejay et al., 2018).
- Although the whole branch of endovascular surgery relies heavily on images, those **medical images remain widely unused**. Indeed we could benefit from deeper, large-scale and automated analyses, carried by new tools that researchers in image processing could provide (Raffort et al., 2020). The perspectives of image processing for vascular surgery are the

[8] https://www.cookmedical.com/products/224e3666-308f-4244-8695-6fd23bbd671c/
[9] https://www.crbard.com/Peripheral-Vascular/en-US/Products/FLUENCY-Plus-Endovascular-Stent-Graft
[10] https://www.goremedical.com/me/products/vgstretch

central motivation for this thesis and are discussed in depth later in this introduction.

We have only listed a small sample of questions arising today in the field of vascular surgery. Yet, addressing these questions already requires joint work from research teams from many fields (biomechanics, image processing, vascular surgery, textile sciences, artificial intelligence, histolopathologists, *etc.*). However, this translational research is also dependent on the availability of data (explanted biomaterials, explanted arteries, patient's clinical images, patient's clinical data, antecedents and outcomes, *etc.*). Since this research involves human subjects on a new topic, the data is scarce and sometimes unavailable due to legal issues. Availability of the data is a crucial point in the field of image processing as we will see later.

The Geprovas laboratory

The Geprovas[11] laboratory (Groupe Européen de Recherche sur les Prothèses liées à la Chirurgie Vasculaire), located in Strasbourg, France, has been founded in 1993 by Professor Nabil Chakfé. It has become a worldwide actor in the field of biomaterials, with an original and unique expertise for devices from vascular surgery. The Geprovas is organized around four activities: the collect and analysis of explants, the innovative research on biomaterials, the medical education program and the clinical analysis program. One of the main goals of Geprovas is to remove barriers and motivate translational research to help answer questions in medical research such as the questions we listed in the previous section.

The work presented in this thesis is initiated and supported by the explant analysis program of the Geprovas. In this context, our work benefited from the wide knowledge of the Geprovas experts as well as from its unique database of explants from vascular surgery.

Medical image processing and new challenges

In the previous section, we mentioned medical imaging and its key role in the modern approaches in vascular surgery. In this second part of the introducion, we present medical imaging in more details. We start with very general definitions and gradually shift our focus towards the mathematical methods linked with image analysis studied in this thesis. The goal is to develop approaches which are automated. Automated analysis is indeed the modern way to process images relying on the power of computers to support medical practice and research.

[11] https://geprovas.org

Medical image processing

Main concepts and vocabulary

A bidimensional (respectively tridimensional) image is a set of values arranged on a rectangular (respectively cubic) grid[12]. For numerical images, the grid is discretized. If the values are scalars, the image is called *grayscale*, if the values are vectors, the image is a color image in some color space (Fieguth, 2010).

The most common medical image processing operations are image *segmentation*, which consists in dividing an image into non-overlapping regions with homogeneous properties, image *registration*, which deals with associating the objects/features in one image with those in one or more other images, and image *denoising*, which aims at estimating a noise-reduced image given an image corrupted by noise (Fieguth, 2010). Image segmentation is the task studied in this thesis and all the methods we develop can be linked with the final goal of segmentation.

A numerical image has to be constructed with some devices, in medical image processing, it is related to the notion of *image modality*.

Medical imaging modalities

The observed image that we have to process can be of different types in medical research. Those are called the imaging modalities: they are acquired with different devices relying on different physical phenomena to form the image. Thus, the imaging of the same physical object will result in different images with different properties according to the modality. Following (Suetens, 2017), the main medical imaging modalities for diagnosis and treatment[13] are now listed:

- X-ray Computed Tomography (CT) is an imaging technique based on X-rays. X-rays are electromagnetic waves with wavelengths around 10^{-10}m which are attenued differently according to the matter it interacts with. The computation of the attenuations undergone by X-rays at certain locations of the space are at the foundation of X-ray images. Computed Tomography (CT) refers to the technique of the computations of the attenuations. X-rays have been used in medical imaging starting from the end of the 19^{th} century. Figure 0.12a shows a typical CT scan image from vascular surgery.

- Magnetic Resonance Imaging (MRI) is an imaging modality introduced in the medical field in the 1970s. In MRI, the reconstructed image illustrates the magnetic properties of the object of interest. In normal operating conditions, MRI is safer for patients than CT imaging or PET imaging (see next item) because MRI does not rely on ionizing radiations.

[12] The definition straightforwardly extends to higher dimensions.

[13] The use of these modalities for medical diagnosis and treatment defines the medical field of *radiology*.

- Positron Emission Tomography (PET) is a technique from the field of nuclear medical imaging. It uses a tracer molecule that will be involved in a metabolic process of the body. The tracer molecule is injected to the patient and can be localized because some identified radioactive atoms of the molecule emit γ-rays (wavelength below 10^{-11}m). This way, an abnormal metabolism can be detected. PET is used clinically since the 2000s.

Note that some modalities describe the anatomy, this is called *structural imaging* and this is the case for most CT and MRI performed. Others depict a function, this is called *functional imaging*; PET is a functional imaging modality. In more specific contexts, a modality can be ambivalent and relate to both structural and functional imaging.

In this thesis, we will only develop approaches for X-ray images. Indeed, X-ray images are more common than any other modalities in vascular surgery notably because current MRI devices can be unsafe for patients with a metallic implant. Indeed, MRI might induce stent dislodgement, heating and important image artifacts (Jabehdar Maralani et al., 2020). X-ray images have also the advantage of being a totally *non-invasive* technique[14], the body of the patients does not need to be touched.

Our work will be more particularly focused on the potential assets of micro-computed tomography (mCT or microCT) X-ray images used for the research in vascular diseases. The formation of mCT images follows the same principles as CT X-ray images described above but they are suited for the imaging of small objects (animals [15], stones, wood, *etc.*). Notably, mCT modality deals particularly well with the imaging of explanted arteries. Those images offer a much greater spatial resolution than classic CT scans since pixel sizes are on the scale of the micrometer (Flannery et al., 1987). The first report of a microtomographic image dates back to 1982 in (J. C. Elliott and Dover, 1982). Figure 0.12b shows a mCT image of an arterial cross-section.

Remark: We will also mention *optical microscopy*, which is an image modality that uses light to image a very thin section of an object of interest, which often cannot be analyzed or seen by the naked eyes (Davidson and Abramowitz, 2002). Figure 0.12c describes an arterial cross-section seen in optical microscopy. The work in this thesis will not deal with a direct processing such images, however microscopic images will be essential in the project described in Chapter 5.

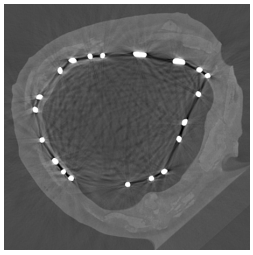
Main challenges

In this section, we describe the difficulties linked with medical image processing that motivate the approaches we developed during the thesis.

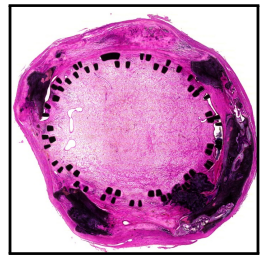
[14] Provided we tolerate the ionizing radiations which are problematic for some patients (Pearce et al., 2012) but also clinicians (Brun et al., 2018).

[15] Some illustrations of the possibilities offered by mCTs in the imaging of small animals are depicted in (Roque-Torres, 2020).

(a) CT: slice of a CT scan which depicts an abdominal cross-section of a human body. A stent implanted in the aorta is visible at the center of the image. Artifacts are also visible, see Figure 0.13 for more details.

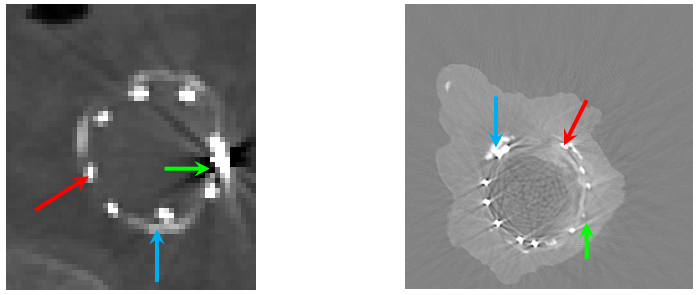


(b) mCT: cross-section of a stented and explanted superficial femoral artery.



(c) Microscopy: histological cross-section of an explanted, stented and totally occluded superficial femoral artery.

Figure 0.12.: Illustrations of the CT, mCT and microscopy medical imaging modalities with examples from vascular surgery.



(a) CT scan (b) mCT scan

Figure 0.13.: Metallic artifacts in X-ray scans hiding the underlying anatomy. Elements around the stent parts are hardly discernible. Green, red and blue arrows respectively indicate stent metallic artifacts, stent components and calcification components.

Metallic biomaterials and artifacts

We have discussed above that metallic stents (and metallic biomaterials in general) are an issue for MRI scanning but some problems also arise in CT scans. Due to high variations of attenuations in metallic parts according to the X-ray energy level, the CT reconstruction algorithms introduce artifacts in the reconstructed images. The artifacts are problematic since they hide the underlying anatomy. Figure 0.13 illustrates stent metallic artifacts in vascular X-ray scans. This issue is becoming increasingly important because of the development of biomaterials (orthopedic prostheses, stents, pacemakers, *etc.*). To answer this problem, many approaches to Metal Artifact Reduction (MAR) have been developed in the literature. See (H. S. Park et al., 2015) for an introduction on stent metallic artifacts and MAR.

Availability of the data

In some medical applications data are scarce while in other fields data abound. This variable then plays a role on the processing approach to follow and on the problem complexity. In vascular surgery, biomaterials were usually destroyed after explantation or patient's death and relatively little research could be carried. As stated earlier, the goal of the explant analysis program of the Geprovas is precisely to collect such data, and make it available in a database that can be used in subsequent studies.

An image analysis task in which no images can be used to learn the algorithm parameters, *i.e.*, where the algorithm has to work with the observed image alone, is called *unsupervised*. In the opposite case, it is called a *supervised* task. Both approaches will be seen in the thesis. Note that a supervised problem requires a human interaction at some stage, for example, to add annotations to the data.

Remark: Collections of biological samples used for scientific investigation, or *biobanks*, are initiatives that are slowly gaining popularity (PW Scholtes et al., 2011).

Computational costs

Image processing can rapidly become a computationally intensive task, especially for medical images. We encounter this problem in our work dealing with mCT images which are big, high resolution images. In other problematics, when a database with hundreds of elements must be processed, the treatments must be wisely chosen. In this thesis, some of the works study the problem of computational complexity.

Remark: Our developments are however not constrained by speed considerations. Our primary data are images of explanted materials, thus the time constraint is not comparable at all to developments linked with videos or real-time analyses such as (Nwoye et al., 2019).

Image processing and probabilistic modeling

In this thesis, the main contributions will be focused on probabilistic graphical models applied to image segmentation, particularly Hidden Markov Models (HMM) (Baum and Petrie, 1966) and their extensions to Pairwise and Triplet Markov Models (Pieczynski and Tebbache, 2000) (Lanchantin et al., 2011) (Gorynin, Gangloff, et al., 2018). Markovian models offer a simple and intuitive way to introduce dependencies between the pixels of the images and thus, they can be a relevant answer to the problematics previously presented.

For example, it will be seen that stent metallic artifacts (illustrated in Figure 0.13) can be modeled by spatially correlated noise: a pixel value at a location is highly dependent on the neighboring pixel values. Such a noise is naturally taken into account in Pairwise and Triplet Markov Models. Such correlations are also studied in models based on Gaussian Markov Random Field (Rue and Held, 2005) (GMRF). In this thesis, these models are then studied as possible improvements to MAR algorithms and as a way to perform segmentation in images degraded with spatially correlated noise.

A second asset to probabilistic models is that they are relatively efficient in unsupervised problems. Following the popular Expectation Maximization (Dempster et al., 1977) (EM) algorithm, there is a rich literature of learning algorithms in unsupervised context for graphical models (Celeux et al., 1995) (Tieleman, 2008). They are then competitive methods for problems with scarce and/or missing data.

Probabilistic models also benefit from a very active research community focusing on optimization methods. Efficient approaches exist to approximate computations that cannot be done exactly. See for example (C. Zhang et al., 2018) which deals with Variational Inference (VI). They also offer much modelization power with possibly auxiliary random variables (Lanchantin et al.,

2011) and can then fit many complex real life problems (Courbot, Monfrini, et al., 2018).

Probabilistic graphical models will be introduced in depth in Chapter 1.

Outline of the thesis

This thesis is composed of five chapters and their appendices.

Chapter 1 consists in a literature review of the main concepts and developments around probabilistic modeling with graphical models with a special focus on image segmentation. This chapter sets up all the definitions and notations that will be used more in depth in subsequent chapters. Chapters 2, 3 and 4 present the core research of this thesis dealing with the improvement and development of new probabilistic models for image segmentation. Chapter 2 proposes a new model of Markov Fields, called Gaussian Pairwise Markov Fields (GPMF), for the unsupervised segmentation of regions that are corrupted with long-spatially correlated noise. The model is tested on both artificial and real datasets. Chapter 3 describes a new model of Markov Trees, called Spatial Triplet Markov Tree (STMT), aiming at being a deterministic counterpart to Markov Fields, which are also used in the context of unsupervised image segmentation. Theoretical results and applications on real data are provided. STMTs make use of auxiliary random variables, the enhanced correlations they introduce are largely studied. Chapter 4 proposes a review on Deep Learning for medical imaging and describes the development of a Convolutional Neural Network for semantic segmentation. This segmentation is subsequently improved by a new approximate inference approach for a model of Conditional Random Fields that is also presented in Chapter 4. Chapter 5 illustrates the applications to vascular surgery of the probabilistic graphical models developed along the thesis.

The appendix of each chapter gives the full details of calculations and additional observations.

List of publications

During this thesis, several journal and conferences articles have been written. They are listed below according to the part of the thesis they deal with. For completeness, the other publications from the author of the thesis are also listed even if their topic is not directly related to the content of this thesis.

First we list the publications primarily directed to the signal processing community:

- Chapter 2:
  - **Under review**: Unsupervised Segmentation with Gaussian Pairwise Markov Fields, H. Gangloff, J.-B. Courbot, E. Monfrini, C. Collet, *Computational Statistics & Data Analysis*, 2020.

  - (Gangloff, Courbot, et al., 2019): Segmentation non-supervisée dans les champs de Markov couples gaussiens, H. Gangloff, J.-B. Courbot, E. Monfrini, C. Collet, *Colloque GRETSI*, 2019.

- Chapter 3:
  - **Under review**: Unsupervised Image Segmentation with Spatial Triplet Markov Trees, H. Gangloff, J.-B. Courbot, E. Monfrini, C. Collet, *International Conference on Acoustics, Speech, and Signal Processing*, 2021.
  - (Gangloff, Courbot, et al., 2020): Spatial Triplet Markov Trees for auxiliary variational inference in Spatial Bayes Networks, H. Gangloff, J.-B. Courbot, E. Monfrini, C. Collet, *Stochastic Modeling Techniques and Data Analysis*, 2020.

- Chapter 4:
  - **Under review**: Markov Chain Variational Inference in Fully-Connected Conditional Random Fields, H. Gangloff, E. Monfrini, C. Collet, *International Conference on Acoustics, Speech, and Signal Processing*, 2021.

- Chapter 5:
  - (Gangloff, Monfrini, Collet, et al., 2020): Unsupervised segmentation of stents corrupted by artifacts in medical X-ray images, H. Gangloff, E. Monfrini, C. Collet, N. Chakfé, *International Conference on Image Processing Theory, Tools and Applications*, 2020.
  - (Gangloff, Monfrini, Collet, et al., 2019): Segmentation de stents dans des données médicales à rayons-X corrompues par les artéfacts, H. Gangloff, E. Monfrini, C. Collet, N. Chakfe, *Colloque GRETSI*, 2019.

- Not in the thesis:
  - (Gorynin, Crelier, et al., 2016): Performance comparison across hidden, pairwise and triplet Markov models' estimators, I. Gorynin, L. Crelier, H. Gangloff, E. Monfrini, W. Pieczynski, *International Conference on Applied and Computational Mathematics*, 2016.
  - (Gorynin, Gangloff, et al., 2018): Assessing the segmentation performance of pairwise and triplet Markov models, I. Gorynin, H. Gangloff, E. Monfrini, W. Pieczynski, *Signal Processing*, 2018.
  - (Gangloff, Monfrini, Ghariani, et al., 2020): Improved Centerline Tracking for new descriptors of atherosclerotic aortas, H. Gangloff, E. Monfrini, M.Z. Ghariani, M. Ohana, C. Collet, N. Chakfé, *International Conference on Image Processing Theory, Tools and Applications*, 2020.

We then present the publications primarily directed to the vascular surgery community:

- Chapter 5:
  - (Kuntz et al., 2020): Co-registration of peripheral atherosclerotic plaques assessed by conventional CT-angiography, micro-CT and histology in CLTI patients, S. Kuntz, H. Jinnouchi, M. Kutyna, S. Torii, A. Cornelissen, Y. Sato, M. E. Romero, F. Kolodgie, A. V. Finn, A. Schwein, M. Ohana, H. Gangloff, A. Lejay, N. Chakfé, R. Virmani, *European Journal of Vascular & Endovascular Surgery*, 2020.
  - **Under review**: Automated histological segmentation on micro-computed tomography images of atherosclerotic arteries, S. Kuntz*, H. Gangloff*, H. Naamoune, E. Monfrini, C. Collet, A. Lejay, M. Kutyna, R. Virmani, N. Chakfé, *European Journal of Vascular & Endovascular Surgery*, 2020.

Chapter 1.

Main concepts in probabilistic modeling

Contents

1.1. Introduction

This chapter introduces probabilistic modeling with a focus on Hidden Markov Models and on applications to probabilistic image segmentation. More in-depth introductions can be found in popular books on the topic such as (C. M. Bishop, 2006) (Goodfellow et al., 2016) (Koller and N. Friedman, 2009) (Murphy, 2012) (Wainwright and Michael I Jordan, 2008).

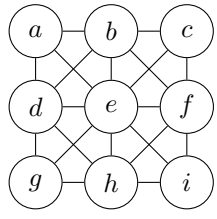
1.2. Graphical modeling

1.2.1. Main definitions

Elements of graph theory

A *graph* $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ is formed by a set of *vertices* $\mathcal{S}$, that we also refer to as *sites* or *nodes*, and by a set of *edges* $\mathcal{E}$. We have $\mathcal{E} \subset \mathcal{S} \times \mathcal{S}$, since, in general, not all vertices are connected by an edge. The set of vertices connected to a vertice $s \in \mathcal{G}$ is called the *neighborhood* of s and is denoted $\mathcal{N}_s$. Note that we have $s \notin \mathcal{N}_s$. A *clique* is a subset of $\mathcal{S}$ with the property that every two elements are connected by an edge. A clique is then a *fully-connected* subset. A *maximal clique* of $\mathcal{G}$ is clique of $\mathcal{G}$ which cannot accept any more vertice from $\mathcal{G}$ without breaking the clique property. A *tree* is a connected graph without any cycle. In a graph $\mathcal{G}$, the edges can either be directed, undirected or a mix of both. This leads to different families of probabilistic models which we discuss in the next sections. In the graphical representation of the probabilistic graphical models we will use arrows to represent directed edges. The absence of arrow represents an undirected edge. These first definitions are illustrated in Figure 1.1.

In a probabilistic graphical model, each vertice s of $\mathcal{G}$ is associated with a random variable whose name then contains the name of the vertice as subscript, for example, X_s.



An undirected graph

$\mathcal{S} = \{a, b, c, d, e, f, g, h, i\}$

$c = \{a, b, d, e\}$ is a clique
The subset c is fully connected

$\mathcal{N}_e = \{a, b, c, d, f, g, h, i\}$
$\mathcal{N}_a = \{b, d, e\}$

Figure 1.1.: Examples and notations of graph theory I.

Elements of probability theory

Let X be a *random variable* from a probability to space, equipped with a probability measure p, to a measurable space $(E, \mathcal{E})$. If E is a finite or countable set, then X is said to be a *discrete* random variable. If E is an uncountable

set, then X is said to be a *continuous* random variable. In both cases, we are interested in measuring the probability of X taking certain values, *i.e.*, we want to evaluate the quantities of the type $P_X(A) = p(X \in A), \forall A \in \mathcal{E}$. Such computations involve P_X another probability measure. P_X is called the *law* of X which is more precisely called *distribution* in the discrete case and *density* in the continuous case. The definitions are similar for random vectors. When considering the law of random vectors made of both continuous and discrete random variables, we may use indistinctly the term distribution or density. If P_X is the law of a random variable X, we may write $X \sim P_X$.

If X is a discrete random variable, we denote by x its realization, $\forall x \in \mathcal{E}$. When A is a singleton, *i.e.*, $A = x, \forall A \in \mathcal{E}$, we have $p(X \in A) = p(\{X = x\})$. When there is no ambiguity we write $p(\{X = x\}) = p(X = x)$ or even $p(\{X = x\}) = p(x)$[1] to denote the probability of a realization x of X. Similarly, if X is a continuous random variable, in case of no ambiguity, we might use $p(x)$ to refer to the density of X. If $X \sim P_X$, we denote the expectation of X by $\mathbb{E}[X]$, which can also be written, with a slight notation abuse, $\mathbb{E}_{x \sim p(x)}[x]$.

The conditioning of a probability law on the realization of some other random variable(s) will be written classically with a vertical bar $|$. For example, using the notation shortcuts mentioned above, the law of X given the realization y of random variable Y will be written $p(x|y)$.

All our definitions extend to the multivariate case: random variables then become *random vectors*. Some of the random vectors we will work with are more precisely *stochastic processes*. However, stochastic processes and their properties are out of the scope of this thesis and little will be said about them.

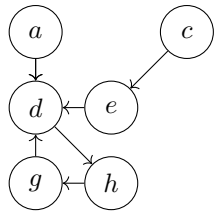
1.2.2. Directed Graphical Models

Directed Graphical Models (DGMs), also known as Beliefs Networks, have *directed* edges represented by an arrow between two vertices. Let $(s, s') \in \mathcal{S}^2$, if there is a directed edge from s to s' then s is called the *father* node of s', conversely, s' is called the child node of s. A node s may have several fathers, which are refered to as the set of *parents* of s, denoted $\mathcal{P}(s)$. A node for which the parent set is empty is refered to as a *root* node. Furthermore we denote $\bar{\mathcal{S}}$ the set of vertices which have at least one father. Figure 1.2 illustrates the new notions we have just introduced.

The notion of Directed Acyclic Graph (DAG), also known as Bayesian Networks, refers to graphs where all the edges are directed and where the graphs do not contain directed cycles. The directed edges of a DGM leads to a partial ordering of its vertices (Wainwright and Michael I Jordan, 2008). A *directed tree* is a DAG which does not contain any directed cycles and not any *semi cycles*. A semi cycle is the DAG that has been obtained after reversing some arrow directions of a directed cycle.

In this thesis we will focus on several kinds of *directed rooted trees*, called *arborescences*. In such graphs the root vertice is unique, it is then the only

[1] Note that this last notation shortcut then confuses p with P_X but no ambiguity is raised in the work of this thesis.



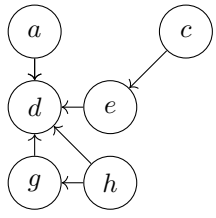
A directed graph

$\mathcal{S} = \{a, c, d, e, g, h\}$
$\bar{\mathcal{S}} = \{d, e, g, h\}$
a and c are root nodes

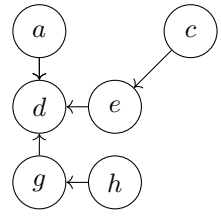
a is the father of d
d is the son of a
$\mathcal{P}(d) = \{a, e, g\}$

$\{d, g, h\}$ forms a directed cycle

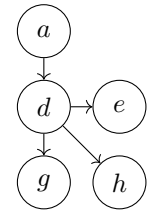
Figure 1.2.: Examples and notations of graph theory II.



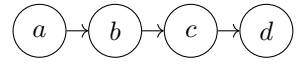
A DAG
$\{d, g, h\}$ is a semi cycle



A directed tree



An arborescence



A directed chain

Figure 1.3.: Examples and notations of graph theory III.

vertice with no edge pointing towards it, and all the other edges are directed away from it. Note that in arborescences, the fact that there is only one root node implies that all the other non-root nodes have exactly one father. When all the vertices of an arborescence have one son (except for the last vertice), we obtain a *directed chain*. Figure 1.3 shows examples of the newly defined graphs.

DGMs are specified by *local conditional probabilities* which are given for every random variables associated to every site: $\forall s \in \mathcal{S}$ the associated conditional probability is $p(x_s|\boldsymbol{x}_{\mathcal{P}(s)})$. If r is a root node, the associated conditional probability is $p(x_r)$. The *joint probability distribution* defining a DGM is obtained, using the definition of conditional probabilities, by multiplying the local conditional probabilities:

$$p(x_{s_1}, x_{s_2}, \ldots, x_{s_N}) = \prod_{r \in \mathcal{S} \setminus \bar{\mathcal{S}}} p(x_r) \prod_{s \in \mathcal{S}} p(x_s|\boldsymbol{x}_{\mathcal{P}(s)}), \tag{1.1}$$

where $s_1, \ldots, s_N$ are the N vertices of $\mathcal{S}$.

Sampling in a DGM is performed by *ancestral sampling*: starting from the root nodes we follow the partial order over the nodes, and we sample each time the associated random variable using the local conditional probability (Goodfellow et al., 2016).

Conditional independence of random variables given other random variables (or *Markov* property) is an important property when dealing with graphical models. The procedure of *d-separation* can determine whether there is conditional independence (Murphy, 2012). A particular result of this procedure that will be underlying many of our studies is for arborescences: let A, B and C be subsets of $\mathcal{S}$. $\forall a \in A$, $\forall c \in C$, the random variables $\{X_a\}_{a\in A}$ and $\{X_c\}_{c\in C}$ are said to be *conditionally independent* given B if every path (without considering the direction of the edges) between a and c goes through a vertice from B.

1.2.3. Undirected Graphical Models

Undirected Graphical Models (UGMs), also known as Markov Random Fields (MRFs), have undirected edges. They are defined differently from DGMs. We call *potential functions*, or *factors* or *unnormalized probability distributions*, a real non-negative function $\phi_C(\{x_s\}_{s\in C}) = \phi_C(\boldsymbol{x}_C)$ associated with a clique C. Potential functions describe the interactions between random variables. We then have that an UGM is defined by a set of random variables $\boldsymbol{X}$ which admits for joint distribution:

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\phi_C(\boldsymbol{x}_C), \tag{1.2}$$

where $\mathcal{C}$ is the set of all cliques of $\mathcal{G}$ and Z is the *partition function* or *normalization constant*, defined as:

$$Z = \sum_{\boldsymbol{x}}\prod_{C\in\mathcal{C}}\phi_C(\boldsymbol{x}_C). \tag{1.3}$$

The partition function ensures that the distribution sums to 1. As a special parametrization, a probability distribution which takes the form of Equation 1.4 is known as a *Gibbs distribution*:

$$p(\boldsymbol{x}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{x})/T\right) = \frac{1}{Z}\exp\left(-\frac{1}{T}\sum_{C\in\mathcal{C}}\psi_C(\boldsymbol{x}_C)\right) \tag{1.4}$$

where E is called the *energy function*, ψ_C are some potential functions and T is a parameter called the *temperature* (with $T > 0$). All the MRFs seen in this thesis will follow a Gibbs distribution.

An equivalent definition for UGMs is possible due to the Hammersley-Clifford theorem (Hammersley and Peter, 1971). A set of random variables $\boldsymbol{X}$, defined such that:

$$\begin{cases} p(\boldsymbol{x}) > 0, \text{ for all set of realizations } \boldsymbol{X} = \boldsymbol{x}, \\ p(x_s|\boldsymbol{x}_{\mathcal{S}\setminus\{s\}}) = p(x_s|x_{\mathcal{N}_s}), \forall s \in \mathcal{S}, \end{cases} \tag{1.5}$$

forms an UGM or MRF, with respect to the neighborhood $\mathcal{N}$. The MRF definition of Equation 1.5 is called the *full conditional* definition. The Hammersley-Clifford theorem proves that the definitions of Equations 1.4 and 1.5 are equivalent. It then appears that the conditional independence property (Markov property) for UGMs is simpler than for DGMs. A random variable at site s,

given the realizations of the random variables located in its neighborhood $\mathcal{N}_s$, is conditionally independent from all the remaining random variables (*i.e.*, from all the vertices in $\mathcal{S} \setminus \mathcal{N}_s$).

Sampling realizations from UGMs is not as straightforward as for DGMs. Except from few exceptions where sampling can be done without approximations (Stoehr, 2017), Markov Chain Monte Carlo (MCMC) approaches are the most popular way for this task. The *Gibbs sampler* is a widely used algorithm to sample from UGMs. It belongs to the MCMC approaches and is a special case of the Metropolis-Hastings algorithm (S. Geman and D. Geman, 1984). The Gibbs sampler relies on the full conditional probabilities and generates a Markov chain of samples which is proved to converge to the joint distribution. The Gibbs sampler is interesting when samples from the joint distribution cannot be directly drawn, as it is the case in general. The algorithm is given in Algorithm A.1.

Extensions of the Gibbs sampler algorithm are an active research topic, notable examples are the chromatic Gibbs sampler (Gonzalez et al., 2011) or the blocked Gibbs sampler (Brown et al., 2019). Another extension for the Gibbs sampler consists in improving the exploration of the modes of the joint distribution using *tempered* transitions (Salakhutdinov, 2009). Lastly, in the particular case of Gaussian Markov Random Fields studied in Chapter 2, sampling can be done in a single iteration of a *fully-blocked* Gibbs sampler or a single sampling from a standard normal distribution followed by Fourier-based transformations (Rue and Held, 2005).

Remark: The notions of trees and chains also exist for UGMs, they will not be mentioned in this chapter for brevity.

Remark: In practice, many models, such as those studied in this thesis, have both directed and undirected edges. In such cases, the associated probability distribution, the parameter estimation process and the inference process for these mixed models will straightforwardly follow either from the DGM theory or UGM theory.

Remark: Note that Factor Graphs (Frey, 2003) are a popular kind of graphical models that aims at unifying UGMs and DGMs. Factor Graphs are however out of the scope of this thesis.

1.3. Inference in probabilistic models

Inference in a probabilistic model can refer to several closely related computational tasks:

- Computing a marginal distribution $p(\boldsymbol{x}_A), A \subset \mathcal{S}$.
- Computing a conditional distribution $p(\boldsymbol{x}_A|\boldsymbol{x}_B), A \subset \mathcal{S}, B \subset \mathcal{S}$ and $A \neq B$.

- Computing a function of a probability distribution. Two important inference problems are:
  - Marginal inference[2]: $\forall s \in \mathcal{S}, \hat{x}_s = \text{argmax}_{x_s} p(x_s)$.
  - Maximum A Posteriori (MAP) inference: $\hat{\boldsymbol{x}} = \text{argmax}_{\boldsymbol{x}} p(\boldsymbol{x})$.

  In Section 1.5.2, these two inference problems will be studied in a Bayesian context.

Inference rapidly becomes a computationally intractable task for general graphs, hence the need for approximate computations in order to relieve the computational burden. Exact computations are however possible in DAGs and is still the subject of ongoing research for more general graphs. We use the term *exact*, or *direct*, for an inference process which benefits from exact computations, and *approximate*, or *indirect*, for an inference process which requires approximate computations. In the next sections, we first review key concepts for exact inference, then for approximate inference.

1.3.1. Exact inference

One of the basics of exact inference is the Variable Eliminitation algorithm which enables computing isolated marginals in arbitrary DGMs or UGMs (N. L. Zhang and Poole, 1994) (Michael I Jordan, 2004). However in most of the cases we are interested in several marginals and running many times the VE algorithm becomes computationally prohibitive and a waste of resources (most of the operations are the same throughout the instances of the algorithm).

In order to compute several marginals, the principle of the Variable Elimination algorithm is reformulated into a message-passing algorithm called Sum Product algorithm (or Belief Propagation algorithm with summations) (Pearl, 1982) (Gormley and Eisner, 2015). The Sum Product algorithm is the basis for exact marginal inference in undirected acyclic graphs. In the case of MAP inference in undirected acyclic graphs, the Max Product algorithm (or Belief Propagation with maximizations) (Loeliger, 2004) is the basic approach.

In the case of marginal inference applied to arborescences, the Sum Product algorithm becomes equivalent to the popular Forward Backward (FB) algorithm (Baum, Petrie, et al., 1970), generalized in the Upward Downward (UD) algorithm (J.-B. Durand and Gonçalves, 2001) (Laferté et al., 2000). The FB algorithm is the foundation to many extensions following its principles. The FB and the UD algorithms are frequently used for inference in particular arborescences: the Markov Chains and the Markov Trees that will be presented in Section 1.6 and studied in this thesis. The FB and UD algorithms are presented respectively in Algorithm A.2 and in Algorithm 3.1. In the case of MAP inference in arborescences, the Viterbi algorithm (Viterbi, 1967) is a popular approach (a version of the Max Product algorithm dedicated to arborescences). The Viterbi algorithm has been generalized to Markov Trees in (Laferté et al., 2000).

[2] Also called Maximum Posterior Mode (MPM).

1.3.2. Approximate inference

Approximate inference is the general approach for all the cases but those mentioned in the previous section. The first approach we refer to is the direct extension of the BP algorithm to arbitrary graphs: the Loopy Belief Propagation algorithm (Gormley and Eisner, 2015). Despite being a popular approach to approximate inference, the convergence of the Loopy Belief Propagation algorithm is not guaranteed: the message massing procedure can never terminate on arbitrary graph structures.

Variational Inference has been introduced more recently. It recasts approximate inference into a deterministic optimization problem (Michael I Jordan et al., 1999) (Lauritzen and Spiegelhalter, 1988) (L. K. Saul et al., 1996). In this context, a probability distribution p in which inference is complex is approximated by a *variational distribution* q in which inference is easy (often exact using the methods from Section 1.3.1). The optimization problem consists in finding q which minimizes the reverse Kullback-Leibler (KL) divergence between p and q. More on Variational Inference will be seen in Chapters 3 and 4. We can note that it is a very active research topic (C. Zhang et al., 2018).

Let us now mention popular approaches for approximate inference in UGMs defined by a Gibbs distribution. Well-known approaches for the computation of the mode of a density are the simulated annealing approaches (Kirkpatrick et al., 1983). Serial Gibbs Simulated Annealing (SA) (S. Geman and D. Geman, 1984) is a probabilistic approach based on samplings from a Gibbs distribution (Equation 1.4) with a varying temperature to ease the exploration of the modes of the distribution. The algorithm is theoretically guaranteed to converge towards the mode of the distribution, but putting SA into practice can be cumbersome and one can end up with an approximation of the optimum we want to find. SA is described in Algorithm A.3. The literature on simulated annealing approaches is large. (Delahaye et al., 2019) offers a good introduction on the topic. The Gibbs sampler presented in Section 1.2.3, as well as other sampling approaches, are then often part of an inference process. Additional comments on inference will be made in Section 1.5.2.

Remark: There are other popular approaches from different subfields to answer inference problems, in particular for MAP inference which seems more studied for image segmentation. For example, two other main approaches are based on linear programming (Komodakis et al., 2010) or graph-cut (Kappes et al., 2016). However such approaches are out of scope of our work.

1.4. Parameter estimation

Either local conditional probabilities of DGMs or potential functions of UGMs include some parameters that need to be estimated. This section is devoted to the estimation of those parameters.

Remark: This section on parameter estimation does not cover the topic

of learning the structure of the graphical models (Drton and Maathuis, 2017) which is a closely related topic. We also do not mention how parameters can be estimated in a fully-Bayesian statistics context (see Section 1.5.2).

1.4.1. Maximum Likelihood estimation

In a graph $\mathcal{G}$, if the realizations of all the random variables are available, or *observed*, then we say that we possess the *complete* data. If the realizations of some random variables are not available, or *hidden*, then we say that we have *incomplete* data.

Let $\boldsymbol{X}$ the vector of the hidden variables of the model and $\boldsymbol{Y}$ the vector of the observed variables. $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ is then the vector of the *completed* data. In this section, the parameters of the model are stacked into a vector $\boldsymbol{\theta}$, the dependence of the distributions on the parameters will be made explicit by using the notation $p(\boldsymbol{y}; \boldsymbol{\theta})$. The *likelihood* of a probabilistic model is a function of the parameters $\boldsymbol{\theta}$ defined by:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = p(\boldsymbol{y}; \boldsymbol{\theta}). \tag{1.6}$$

The *Maximum Likelihood* (ML) estimator seeks the vector of parameters $\boldsymbol{\theta}^*$ which maximizes the likelihood function. It is equivalent and often more easy to maximize the log likelihood function, then:

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \operatorname{argmax}_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \boldsymbol{y}). \tag{1.7}$$

Using the ML parameters follows the intuition that the parameters which maximizes the probability of the observed variables are the best. The rest of this section focuses on the ML estimation of the parameters for both DGMs and UGMs.

1.4.2. Estimation for DGMs

Supervised parameter estimation

We talk about *supervised parameter estimation* when we possess the complete data. Then, let $\boldsymbol{y}$ be the realizations of the random vector associated to a DGM where all the variables are observed (complete data). The Equation 1.7 can be used by plugging in the joint distribution of the DGM (Equation 1.1). The ML estimation then becomes:

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{r \in \mathcal{S} \setminus \bar{\mathcal{S}}} p(y_r) \sum_{s \in \mathcal{S}} p(y_s; \boldsymbol{y}_{\mathcal{P}(s)}). \tag{1.8}$$

We then maximize Equation 1.8 as a function of $\boldsymbol{\theta}$. In some cases, it is possible to derive a closed form solution to the ML estimation. If no closed form solution exists one must use iterative methods such as a gradient descent or the BFGS[3] algorithm to solve this numerical optimization problem (Nocedal and Wright, 2006).

[3] Broyden-Fletcher-Goldfarb-Shanno, the names of the main contributors to the algorithm.

Unsupervised parameter estimation

We talk about *unsupervised parameter estimation* when we do not possess the complete data. Some random variables are hidden and the ML estimation of the parameters is more complex. Indeed, in order to plug the joint distribution of a DGM in Equation 1.7, one must sum over all the hidden variables:

$$\begin{aligned} \boldsymbol{\theta}^* &= \text{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}), \\ &= \text{argmax}_{\boldsymbol{\theta}} \log \sum_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}). \end{aligned} \tag{1.9}$$

Because of the logarithm that ties together the parameters, maximizing Equation 1.9 is much harder than maximizing Equation 1.8. The Expectation Maximization (EM) algorithm (Dempster et al., 1977) is an iterative approach to perform such a maximization in the case of hidden data. It alternates between an E-step in charge of inferring the missing values and a M-step which performs the actual maximization with the completed data. The EM algorithm is guaranteed to converge to a point of null gradient (local maximum or saddle point). Note that the E-step is a step of probabilistic inference, and the methods seen in Section 1.3 need to be used. The EM algorithm is given in Algorithm A.4.

1.4.3. Parameter estimation for UGMs

In general, the parameter estimation techniques for DGMs cannot be applied to UGMs for mainly one reason: the partition function is intractable for UGMs and computations must always be approximate. The algorithms we review here are stochastic approximations of the previous algorithms or modified versions of the ML estimation.

Remark: In some exceptions the partition function is tractable (see for example (Sutton and McCallum, 2012)), then one can apply the techniques of Section 1.4.2 to UGMs.

Stochastic versions of gradient descent

Let us first consider fully-observed models. In order to apply a gradient descent based method to estimate the parameters $\boldsymbol{\theta}$, let us consider the gradient of the log likelihood with respect to $\boldsymbol{\theta}$ in the case of a UGM (Equation 1.2):

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{y}_C; \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}). \tag{1.10}$$

It can be shown that:

$$\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y} \sim p(\boldsymbol{y}; \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \log \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{y}_C; \boldsymbol{\theta}). \tag{1.11}$$

This identity is used in several algorithms using MCMC approximations to compute the expectation in the gradient expression. A popular example is the

Stochastic Maximum Likelihood (Younes, 1999), also discovered as Persistent Contrastive Divergence (Tieleman, 2008).

Identity 1.11 can also be used to approximate the untractable summation which appears in the case of ML estimation with hidden variables (see Equation 1.9). This then extends the Stochastic Maximum Likelihood and Persistent Contrastive Divergence algorithms to the case of incomplete data.

Pseudolikelihood

One way to avoid the problem of the partition function is to change the objective function so that there is no more partition function to compute. This concept is known as the *pseudolikelihood* approach.

There exists many different pseudolikelihoods. A popular choice introduced in (J. Besag, 1975) is the pseudolikelihood where the joint distribution is a simple product of the factors:

$$\mathcal{PL}(\boldsymbol{\theta};\boldsymbol{y}) = \log \prod_{c\in\mathcal{C}} \phi_C(\boldsymbol{y}_C). \tag{1.12}$$

The difficulty involved by the partition function is now discarded and one may continue the optimization with the approaches from Section 1.4.2.

Alternative forms for the EM algorithm

In the EM algorithm introduced in Section 1.4.2. Both the inference (E) and maximization (M) step can be complex for UGMs for all the reasons we have given so far. Moreover the other well-known limitations of the EM algorithm, such as its dependence to initial parameters and slow convergence, has motivated the development of stochastic versions of the EM algorithm (Celeux et al., 1995). Among them, the Stochastic Expectation Maximization (SEM) (Celeux, 1985) algorithm will be studied in this thesis. Its principle is to replace the E-step by sampled realizations of the hidden variables to complete the data. The algorithm is not proved to converge, however it empirically behaves well in practice. Algorithm A.5 describes the SEM algorithm.

When the inference step (E) is intractable, another common idea is to approximate this step with variational approximations of the EM algorithm (Michael I Jordan et al., 1999).

Remark: Stochastic versions of the EM algorithm can also be applied to DGMs.

1.5. Probabilistic models for image segmentation

1.5.1. Definitions and context

In this section we give the general setting in which probabilistic graphical models can be used for image segmentation.

For each pixel of the image, the associated graph will contain two vertices, both at site s. An observed variable Y_s is associated with the first vertice and its realization is taken equal to the pixel value at site s. A hidden variable X_s is associated with the second vertice, X_s has value in the set of classes. The total number of sites is also the number of pixels and is taken equal to N. Then, more formally, $\boldsymbol{X}$ is a hidden process with value in $\boldsymbol{\Omega}^N$ the set of classes of the image, $\boldsymbol{X}$ forms the *segmented image*. $\boldsymbol{Y}$ is the observed process with value in $\mathbb{R}^N$ or $\mathbb{Z}^N$ depending on the pixel values.

The setting we have just described is the foundation for many works in image segmentation. In particular, Hidden Markov Models described in Section 1.6 are popular for image segmentation.

1.5.2. Bayesian image segmentation

In this section we focus our development on Bayesian inference applied to the context of image segmentation. We are interested in the *posterior* distribution of $\boldsymbol{X}$ given the observed process $\boldsymbol{Y}$. Using Bayes theorem:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}), \tag{1.13}$$

where $p(\boldsymbol{x})$ is called the *prior* distribution and $p(\boldsymbol{y}|\boldsymbol{x})$ is the *conditional likelihood* distribution.

Once the posterior distribution is computed, a decision needs to be taken to produce the actual segmentation. This is done in the context of Bayes decision theory. We first need to define a *loss* function L which measures the compatibility of an estimation with the hidden truth. In the case of image segmentation, L measures the compatibility between the estimated segmentation $\hat{\boldsymbol{x}}$ and the hidden class image $\boldsymbol{x}$. The optimal estimation, having observed $\boldsymbol{Y} = \boldsymbol{y}$, is defined to be that minimizing the *posterior expected loss*, $\tilde{L}(\hat{\boldsymbol{x}}|\boldsymbol{y})$:

$$\tilde{L}(\hat{\boldsymbol{x}}|\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{y})}[L(\hat{\boldsymbol{x}}, \boldsymbol{x})] = \sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})L(\hat{\boldsymbol{x}}, \boldsymbol{x}). \tag{1.14}$$

One common loss function, not restricted to image processing, is the *zero-one loss* function defined by:

$$L^1(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \begin{cases} 0, & \text{if } \hat{\boldsymbol{x}} = \boldsymbol{x}, \\ 1, & \text{otherwise.} \end{cases} \tag{1.15}$$

It can be shown that the estimation which minimizes $\tilde{L}(\hat{\boldsymbol{x}}|\boldsymbol{y})$ when $L = L^1$, is the Maximum A Posteriori (MAP) (S. Geman and D. Geman, 1984) estimate, *i.e.*:

$$\hat{\boldsymbol{x}}^{MAP} = \text{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}). \tag{1.16}$$

Another loss function, less widely known, aims at minimizing the number of misclassifications:

$$L^2(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \sum_{s\in\mathcal{S}} (1 - \delta_{x_s}^{\hat{x}_s}), \tag{1.17}$$

where $\delta_{x_s}^{\hat{x}_s}$ the Kronecker function. This loss function leads to the Maximum Posterior Mode (MPM) (Marroquin et al., 1987) estimator, also known as the Marginal MAP estimator (Liu and Ihler, 2013):

$$\forall s \in \mathcal{S}, \hat{x}_s^{MPM} = \text{argmax}_{x_s} p(x_s|\boldsymbol{y}). \tag{1.18}$$

Algorithm A.6 from (Marroquin et al., 1987) offers an approach for the MPM computation.

Remark: The Bayesian framework of estimation (centered around the use of *prior* distributions) still holds in a more general setting than image segmentation. In particular, prior distributions can be used over the model parameters themselves in the case of fully Bayesian parameter estimation, which is not covered in this thesis. See (Zoubin Ghahramani, 2002) for an introduction and (Morris et al., 1997) (Dobigeon et al., 2009) (Vacar and Giovannelli, 2019) for study examples.

1.5.3. Discriminative and generative models

We conclude this section by noting the difference between discriminative and generative modeling (Minka, 2005) (Sutton and McCallum, 2012).

In the case of image segmentation, computing the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ is the end goal in order to effectively segment the image. There are two ways of doing so. On the one hand, one can directly model the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$, this is called *discriminative* modeling. Chapter 4 of the thesis deals with such a modeling approach. On the other hand, one can model the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ first, which is, by Bayes rule, the product of a prior $p(\boldsymbol{x})$ and a conditional likelihood $p(\boldsymbol{y}|\boldsymbol{x})$. The posterior is then obtained with Equation 1.13. This is called *generative* modeling and Chapters 2 and 3 study generative models.

While it is hard to predict which type of modeling will perform best on a given dataset, it is commonly admitted that generative modeling is more suited to unlabeled datasets or small datasets where modeling a prior can be beneficial and have a regularizing effect. However, on larger annotated datasets, discriminative models tend to perform more accurately (Ng and Michael I Jordan, 2002).

1.6. Hidden Markov Models

Hidden Markov Models (HMMs) (Baum and Petrie, 1966) are the most popular type of probabilistic graphical models. HMMs are generative models which relate to the Bayesian formulation with a prior and a conditional likelihood presented in Section 1.5.2. The image processing context is generalized here. HMMs are a central element of this thesis, and this section is dedicated to these models and some of their extensions.

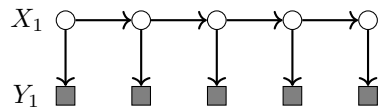


Figure 1.4.: Graphical model corresponding to a HMC with Independent Noise (HMC-IN) for $N = 5$. X_1 is the root node. The white circled vertices are associated with the hidden process $\boldsymbol{X}$. The gray squared vertices correspond to the variables from the observed process $\boldsymbol{Y}$.

1.6.1. Main families of HMMs

Hidden Markov Chains

A Hidden Markov Chain (HMC) (Baum and Petrie, 1966) is a probabilistic model particularly suited for sequence modeling. The associated graph $\mathcal{G}$ is here a directed chain, called, in this context, Markov Chain (MC). Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ be a random vector or process. More precisely we have that $\boldsymbol{X}$ is a MC if:

$$p(x_s|x_{s-1}, \ldots, x_1) = p(x_s|x_{s-1}), \forall s \in \{2, \ldots, N\}. \tag{1.19}$$

Equation 1.19 means that the realization of the random variable at site s, given the preceding realizations in the chain, depends only on the realization of the random variable at site $s-1$[4]. Using the conditioning property, one can show that the joint distribution of the MC $\boldsymbol{X}$ is then:

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)\ldots p(x_N|x_{N-1}). \tag{1.20}$$

Note that in our work we will consider MCs that are *homogeneous*: $p(x_s|x_{s-1})$ is the same $\forall s \in \{2, \ldots, N\}$, see (Brémaud, 2017) for more details. A MC is then defined by a root distribution and transition distribution.

In the HMC model, $\boldsymbol{X}$ then is a MC whose variables are *hidden* in which we want to perform an inference task ($p(\boldsymbol{x})$ is the prior). It is associated to an *observed* process $\boldsymbol{Y}$ whose variables are taken, in the most common case, independent given $\boldsymbol{x}$ ($p(\boldsymbol{y}|\boldsymbol{x})$ is the conditional likelihood) and depending only of the hidden realization at the same site. This is the hypothesis of the Independent Noise (IN). The joint distribution $(\boldsymbol{X}, \boldsymbol{Y})$ of such HMC-IN is then:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x_1) \prod_{s\in\{2,\ldots,N\}} p(x_s|x_{s-1}) \prod_{s\in\{1,\ldots,N\}} p(y_s|x_s). \tag{1.21}$$

In HMCs, inference is directly computed with the Forward Backward (FB) algorithm. In an unsupervised context, using the EM algorithm with FB at the E-step is known as the Baum Welch algorithm (Jelinek et al., 1975). Figure 1.4 illustrates a HMC with Independent Noise.

The HMC has been used in a wide variety of contexts, famous ones are gene prediction (Stanke et al., 2006), speech processing (Toda, 2011) or stock option

[4] $s-1$ is the father node of s in the associated gaphical representation, see Figure 1.4.

forecasting (Gupta and Dhingra, 2012). In Chapter 5 of this thesis, we develop a model of HMC for image processing.

Hidden Markov Trees

We focus on two particular arborescence graphs which are known as Markov Trees (MTs) (Laferté et al., 2000):

- *Dyadic* MT: all the vertices of the arborescence have two sons (except on the last layer).
- *Quadtree* MT: all the vertices of the arborescence have four sons (except on the last layer).

Let $\mathcal{S} = \{\mathcal{S}^1, \dots, \mathcal{S}^L\}$ be the layer-wise subdivision of the set of vertices. $\mathcal{S}^1$ is the first layer containing only one vertice, the *root* r. Dyadic or Quadtree MTs then have similar factorizations to MCs. Let $\boldsymbol{X}$ be a MT, then, $\forall s \in \bar{\mathcal{S}}$:

$$p(x_s|\boldsymbol{x}_{s' \in \mathcal{S}, s' \neq s}) = p(x_s|x_{\mathcal{P}(s)}). \tag{1.22}$$

Equation 1.22 is the direct extension of Equation 1.19 to general arborescences. Recall that in arborescences, the set of parent of a vertice is either empty or contains only one vertice denoted s^-. The joint distribution of a MT $\boldsymbol{X}$ is then:

$$p(\boldsymbol{x}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s|x_{s^-}). \tag{1.23}$$

Hidden Markov Trees (HMTs) are similarly defined as HMCs by a root distribution and a transition distribution. In a HMT, the hidden process $\boldsymbol{X}$ has MT joint distribution. The conditional likelihood distribution implying the observed process $\boldsymbol{Y}$ can be freely formed, but in the most popular case, it is taken as independent given $\boldsymbol{X}$ realizations and depending only on realizations of $\boldsymbol{X}$ at the same site (case of Independent Noise as in Section 1.6.1). The joint distribution $(\boldsymbol{X}, \boldsymbol{Y})$ for such a HMT-IN is then:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s|x_{s^-}) \prod_{s \in \mathcal{S}} p(y_s|x_s). \tag{1.24}$$

Figure 1.5 shows the graphical model of a dyadic HMT-IN.

HMTs and derived models, known as *Latent Tree Models*, are popular in multi-resolution analysis (Kinebuchi et al., 2001) (J.-B. Durand and Gonçalves, 2001) and in phylogenetic analysis (Mourad et al., 2013) (Zwiernik, 2015). To a lesser extent, HMTs have also been used for image processing (Laferté et al., 2000) (Hanzouli et al., 2013). Models based on HMTs are presented in Chapter 3 of this thesis.

Hidden Markov Fields

Hidden Markov Fields (HMFs) (S. Z. Li, 2009) are UGMs but they are constructed as the other previously described HMMs. In HMFs, the hidden process

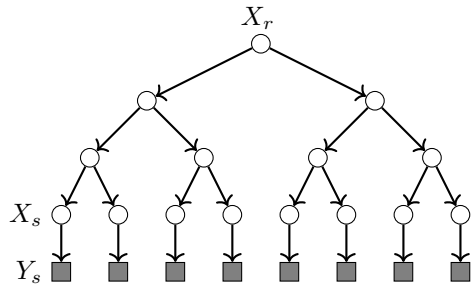


Figure 1.5.: Graphical model corresponding to a dyadic HMT with Independent Noise (HMT-IN) for $L = 4$. X_r is the root node. The white circled vertices are associated with the hidden process $\boldsymbol{X}$. The gray squared vertices correspond to the variables from the observed process $\boldsymbol{Y}$.

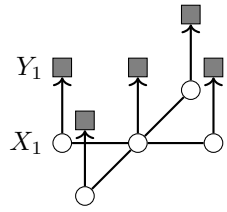


Figure 1.6.: Graphical model corresponding to a HMF with Independent Noise (HMF-IN) for $N = 5$. The white circled vertices are associated with the hidden process $\boldsymbol{X}$. The gray squared vertices correspond to the variables from the observed process $\boldsymbol{Y}$.

$\boldsymbol{X}$ has the joint distribution of a Markov Field (Section 1.2.3). As an example, using the same conditional likelihood as before, *i.e*, an Independent Noise, we get the classical joint distribution $(\boldsymbol{X}, \boldsymbol{Y})$ of a HMF with Potts prior (J. Besag, 1986)[5] and independent Gaussian noise:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\left(\sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s \cup \{s\}}} \beta \delta_{x_s}^{x_{s'}} + \sum_{s \in \mathcal{S}} \left[\log\left(\sqrt{2\pi}\sigma_{x_s}\right) - \frac{(y_s - \mu_{x_s})^2}{2\sigma_{x_s}^2} \right] \right). \tag{1.25}$$

In the last equation, δ is the Kronecker function and β, μ and σ are the model parameters (the parameters will be seen in depth in Chapter 2). The graphical model of a HMF with Independent Noise can be seen in Figure 1.6.

From the pioneering works of (S. Geman and D. Geman, 1984) and (Marroquin et al., 1987), HMFs have been used a lot in image processing fields such

[5] The Potts and Ising (J. E. Besag, 1972) models are equivalent when the hidden stochastic process is binary valued.

as image segmentation (Schmitt et al., 1996) (Mignotte et al., 1999) (Kato and Zerubia, 2012) or computer vision (Wang et al., 2013).

1.6.2. Pairwise and Triplet extensions

Pairwise HMMs make the hypothesis that the joint distribution $(\boldsymbol{X}, \boldsymbol{Y})$ of the HMM is a Markov Chain/Markov Tree/Markov Field.

For example, considering chains, making the Pairwise Markov Chain assumption leads us to the general joint distribution of a Pairwise Markov Chain:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(x_1, y_1) \prod_{s \in \{2,\dots,N\}} p(x_s, y_s | x_{s-1}, y_{s-1}) \tag{1.26}$$

It appears that this equation is a strict generalization of the classic HMC (Equation 1.21). This means that more correlations between random variables can be modeled. In particular, $\boldsymbol{X}$ is not restricted to be a Markov process anymore. Similar observations can be made in the case of Pairwise Markov Trees and Pairwise Markov Fields.

Triplet HMMs go one step further by introducing a third *auxiliary* random process $\boldsymbol{V}$ with values in Λ^N. We then make the assumption that the triplet $(\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{Y})$ is a Markov Chain/Markov Tree/Markov Field. Again, in the case of the chains, a Triplet Markov Chain has for joint distribution:

$$p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{y}) = p(x_1, v_1, y_1) \prod_{s \in \{2,\dots,N\}} p(x_s, v_s, y_s | x_{s-1}, v_{s-1}, y_{s-1}) \tag{1.27}$$

Again, the greater generality of TMCs over HMCs (Equation 1.21) but also over PMCs (Equation 1.27) is to be noted. In particular, neither $\boldsymbol{X}$, $\boldsymbol{V}$, $(\boldsymbol{Y}, \boldsymbol{V})$, $(\boldsymbol{X}, \boldsymbol{Y})$ nor $(\boldsymbol{X}, \boldsymbol{V})$ is necessarily a Markov process anymore.

New pairwise models are developed in Chapter 2 of the thesis, while developments around triplet models can be found in Chapter 3. In the literature, applications of pairwise and triplet HMMs are much less numerous, but they exhibit each time better performances than the classical HMMs (Pieczynski, Hulard, et al., 2003) (Pieczynski and Tebbache, 2000) (Courbot, Mazet, et al., 2019) (Hanzouli-Ben Salah et al., 2017) (Gorynin, Monfrini, et al., 2017) (Courbot, Monfrini, et al., 2018) (Lanchantin et al., 2011) (Gorynin, Gangloff, et al., 2018).

Remark: Auxiliary random variables are central to *deep probabilistic models*. Triplet Markov Trees and Spatial Bayes Networks developed in Chapter 3 can be seen as belonging to this family of models. Other deep probabilistic models will not be studied in this thesis and we refer the reader to (Goodfellow et al., 2016) for an in-depth overview.

1.7. Conclusion

Probabilistic graphical models are powerful approaches to express direct dependencies within a set of random variables capable of modeling a complex

phenomenon. Using graphical models is a way to control the expressiveness of the direct dependencies between random variables and the tractability of the models thanks to the conditional independencies. A successful model is based on two important tasks: the estimation of the parameters of the probability distribution and the inference task. Such tasks have been introduced in this first chapter. As we have seen, it is little to say that there is a wide variety of approaches to solve these challenges and put into practice probabilistic models.

Our presentation also focused on probabilistic graphical models which are very propular in the signal processing field: the Hidden Markov Models. Indeed, the latter offer a good compromise between the dependencies they introduce and their computational performances. HMMs are a family of models which is central in this thesis.

Many of our contributions presented in the next chapters are new probabilistic models that enrich the dependencies between the random variables. Our goal is to model more complex phenomena to address problems from real world applications.

Chapter 2.

Gaussian Pairwise Markov Fields

Contents

2.1. Introduction

Probabilistic models for spatially correlated random variables

This chapter studies the task of unsupervised image segmentation, in the context of Bayesian image segmentation presented in Section 1.5.2, when strong spatially correlated noise corrupts the image. In this case, classical approaches reach their limits and new dedicated models need to be considered to improve the accuracy of the segmentation. We will focus on Pairwise Markov Fields (PMFs) presented in Section 1.6.2. Such models relax the Markovian assumption on the hidden process, which enables the modeling of more complex correlations, while keeping Bayesian inference easily available. Figure 2.1 depicts the undirected graphical models of the classical HMF model and of several PMF models.

We introduce Gaussian Random Fields (GRFs) and Gaussian Markov Random Fields (GMRFs) (Rue and Held, 2005) which are powerful probabilistic models capable of dealing with a large variety of correlated random variables and especially with long-range spatial correlations between pixels in images. GRFs and GMRFs are Undirected Graphical Models (UGMs) defined with respect to a graph $\mathcal{G}$. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, $n < \infty$, $\boldsymbol{X}$ is a GRF with mean $\boldsymbol{\mu}$ (a $(n \times 1)$ vector) and Symmetric Positive-Definite (SPD) covariance matrix Σ (a $(n \times n)$ matrix), if, and only if, its density is a multivariate Gaussian function:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.1}$$

While the graphical representation of a GRF is a fully connected graph, a conditional independence assumption (Markov property) is made to define a GMRF. The latter is defined as a GRF with respect to $\mathcal{G}$ but in which a vertice s is only connected to the vertices of a subset $\mathcal{N}_s \subset \mathcal{S}$ called the neighborhood of s. The associated random variable, X_s, is then independent of $X_{s'}, \forall s' \in \mathcal{S} \backslash \mathcal{N}_s$, given the realizations of $X_{s''}, \forall s'' \in \mathcal{N}_s$. In such case, $Q_{s',s''} = 0$ with $Q = \Sigma^{-1}$, where Q is called the *precision* matrix.

Then formally, the density of a GMRF with mean $\boldsymbol{\mu}$ and precision matrix Q is given by:

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} \det\left(Q^{-1}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})Q(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.2}$$

The conditional independence assumption is a way to alleviate the computational cost of the probabilistic model. Indeed, smaller neighborhoods (few edges in $\mathcal{G}$) lead to a sparse Q which makes computations based on Equation 2.2 much easier. Appendix B recalls the main definitions and properties of GRFs and GMRFs.

Researches have been carried for regression and classification using *conditional* GMRF, *i.e.*, these studies consider a discriminative probabilistic model (Radosavljevic et al., 2010) (Vemulapalli et al., 2016) (Petrović et al., 2019).

However, to the best of our knowledge, there exists no literature on generative probabilistic models using GMRF for image segmentation. This chapter aims at proposing such a new probabilistic modelization. Indeed, generative models might be advantageous when no training dataset is available, for example, in the case of unsupervised image segmentation.

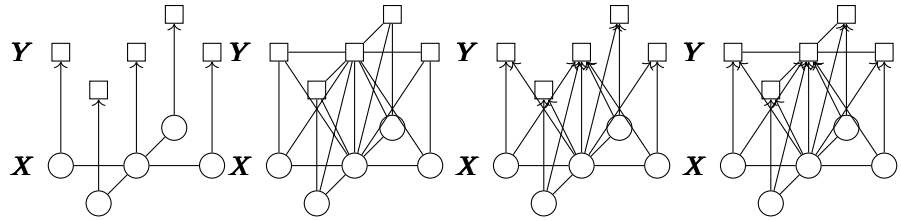
We introduce a new model, called Gaussian Pairwise Markov Fields (GPMF), which belongs to the PMF family. It combines a generalization of the classical HMF model and the ability to model strongly correlated variables as a GMRF, while preserving tractability. First, we show that the PMF hypothesis is a natural way to answer the problem of modeling correlated noise and introducing long correlations by using the GMRF model as conditional likelihood. Secondly, the new model introduces depencies between the observed and latent processes which generalize the usual Markovianity hypothesis of the hidden process. Besides, we handle the unsupervised image segmentation problem which requires a crucial step of parameter estimation. To this end, we propose a stochastic parameter estimation algorithm for the PMF models. In the case of unsupervised image segmentation, the GPMF model performs better than other classical unsupervised approaches.

Remark: Pairwise Markov Field models have been defined independently in (Dimitrova and Kocarev, 2018) and (Y. Park et al., 2017). However, those works mostly focus on structure learning without latent variables (Drton and Maathuis, 2017), which significantly modifies the work hypothesis. To the best of our knowledge, there is no work on structure learning in Pairwise Markov Fields with latent variables. Thus, despite similar model names, the content of our chapter deals with a different problem since latent variables are central to our work.

Link with image processing in vascular surgery

In the context of mCT X-rays scans of human arteries that contain a metallic stent. These images are very noisy because of the strong artifacts caused by the interactions between the X-rays and the metallic stent. To improve the design of biomaterials and their implantation in the human body, it is crucial to analyze, *in situ*, the biomaterial when it fails and when it needs to be explanted from the patient body. In such analyses, we need to precisely segment, the stent, the organic material and the background despite the strong noise in the image. The scarcity of such data requires the use of an unsupervised approach as developed in this chapter. In this image processing problem, the stent artifacts are modeled as a correlated noise. An automatization of the segmentation process could help processing more data and create enhanced inputs for biomechanical studies (such as 3D meshes) carried to increase the knowledge about the vascular diseases.

In Section 2.5.4, we present the application of the new probabilitic model dedicated to handle correlated noise in the task of unsupervised segmentation of mCT images in presence of metallic artifacts.



(a) HMF-IN (b) PMF (c) PMF-UN (d) HMF-CN

Figure 2.1.: Graphical Models of the classical Hidden Markov Field with Independent Noise (HMF-IN) (a), the Pairwise Markov Field (PMF) with full direct dependencies (b), the Pairwise Markov Field-Uncorrelated Noise (PMF-UN) (c), and the Hidden Markov Field-Correlated Noise (HMF-CN) (d). Within the PMF models, $(\boldsymbol{X}, \boldsymbol{Y})$ is the Markov process. GPMFs belongs to the PMF family, as well as PMFs-UN and HMFs-CN which are intermediate cases between the HMF and PMF (with full direct dependencies) models in terms of direct dependencies. PMFs-UN and HMFs-CN will be introduced and studied in Section 2.3. We note the numerous correlations that can be introduced between random variables when the pairwise assumption holds.

2.2. Gaussian Pairwise Markov Fields

2.2.1. Model definition

$\boldsymbol{X} = (X_1, \ldots, X_N)$ is a discrete-valued random vector with values in Ω^N, with $\Omega = (\omega_0, \ldots, \omega_{K-1})$. $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ is a real-valued random vector with values in $\mathbb{R}^N$. $(\boldsymbol{X}, \boldsymbol{Y})$ is a stationary Markov process on a graph, whose vertices are indexed by $\mathcal{S}$ such that $|\mathcal{S}| = N$. The Markov process is defined with respect to the neighborhood $\mathcal{N}$ and we have $\mathcal{N} = (\mathcal{N}^X \cup \mathcal{N}^Y) \subset \mathcal{S}$. Indeed, in general, neighborhoods are different if we consider the hidden or observed variables, thus, a neighborhood can be decomposed on a part linked to the hidden variables ($\mathcal{N}^X$) and another to the observed variables ($\mathcal{N}^Y$).

We use the notation $\tilde{p}$ to refer to factors (or unnormalized probability distributions, see Section 1.2.3). The classical Hidden Markov Field model with Independent Noise (HMF-IN) is defined by the joint distribution of Equation 1.25 that we rewrite as:

$$p(\boldsymbol{x}, \boldsymbol{y}) \propto \prod_{s \in \mathcal{S}} \tilde{p}(x_s | \boldsymbol{x}_{\mathcal{N}_s^X}) \tilde{p}(y_s | x_s). \tag{2.3}$$

The undirected graphical model of a HMF-IN is given in Figure 2.1a.

The PMF family of models generalizes the HMF-IN model as mentioned in Section 1.6.2. Recall that in the PMF family, the assumption of $(\boldsymbol{X}, \boldsymbol{Y})$ being a Markov field is made. Note that this implies that $\boldsymbol{X}$ given some realizations

$\boldsymbol{y}$ of $\boldsymbol{Y}$ is a Markov field. A PMF is defined by the distribution:

$$p(\boldsymbol{x}, \boldsymbol{y}) \propto \prod_{s \in \mathcal{S}} \tilde{p}(x_s, y_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}), \tag{2.4}$$

The greater generality of the pairwise fields hypothesis enables the modeling of more complex correlations between variables. Indeed, neither $\boldsymbol{X}$ nor $\boldsymbol{Y}$ is necessarily Markovian.

Following the notations of Section 1.5, the $\boldsymbol{X}$ are the hidden variables, and the $\boldsymbol{Y}$ are the observed variables. The new GPMF model then offers new modelization properties, notably, it is capable of handling Gaussian spatially correlated noise. Moreover, in the context of Bayesian image segmentation, an estimation of the hidden truth will be done under the MAP criterion (Equation 1.16) and the MPM criterion (Equation 1.18). The MAP estimator is here classicaly estimated with the Simulated Annealing (Algorithm A.3). The MPM will be approximated by the Marroquin algorithm (Algorithm A.6) (Marroquin et al., 1987) using the local expression of the posterior Markov field:

$$p(x_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}) = \frac{\tilde{p}(x_s, y_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{\sum_{x'_s} \tilde{p}(x'_s, y_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}. \tag{2.5}$$

2.2.2. Description of the GPMF distribution

In our new Gaussian Pairwise Markov Field (GPMF) model we define the factor of Equation 2.4 by, $\forall s \in \mathcal{S}$:

$$\begin{aligned} \tilde{p}(x_s, y_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) = \exp\Bigg(- \Big(2 \sum_{s' \in \mathcal{N}_s^X} V(x_s, y_s) + \frac{1}{2} Q_{s,s} \bar{y}_s^2 \\ + \sum_{s' \in \mathcal{N}_s^Y} Q_{s,s'} \bar{y}_s \bar{y}_{s'} \Big) \Bigg), \end{aligned} \tag{2.6}$$

where we denote $\bar{y}_s = y_s - \mu_{x_s}$ and the potential function V is given by:

$$V(x_s, y_s) = -\delta_{x_s}^{x_{s'}} \beta \left(1 - \frac{1}{2} (\bar{y}_s - \bar{y}_{s'})^2 \right), \tag{2.7}$$

with $\delta_{x_s}^{x_{s'}}$ the Kronecker delta function. Equation 2.6 is parametrized by a *granularity* parameter β, spatially varying *means* μ_{x_s}, and a *precision* matrix Q. These model parameters are detailed in Section 2.2.4.

Remark: The potential function of Equation 2.7 can be seen as an enhanced Potts potential (S. Z. Li, 2009) (Pereyra et al., 2013) where the granularity coefficient β is adjusted to the observations available in the neighborhood of each site.

2.2.3. The GPMF conditional likelihood

The local expression of the Pairwise Markov Field of Equation 2.4 derives from the following joint Gibbs distribution (see Appendix C.1 for a proof):

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \exp\big(-E(\boldsymbol{x}, \boldsymbol{y})\big), \tag{2.8}$$

where $E(\boldsymbol{x}, \boldsymbol{y})$ is the joint energy and Z is the normalization constant. It can be shown that:

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^X \cup \{s\}}} \left[-\delta_{x_s}^{x_{s'}} \beta \left(1 - \frac{1}{2} (\bar{y}_s - \bar{y}_{s'})^2 \right) \right] \\ &\quad + \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^Y \cup \{s\}}} \left[\frac{1}{2} Q_{s,s'} \bar{y}_s \bar{y}_{s'} \right] - \beta \left(1 - \frac{1}{2} (\bar{y}_s - \bar{y}_s)^2 \right), \\ &= \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^X \cup \{s\}}} -\delta_{x_s}^{x_{s'}} \beta + \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^X \cup \{s\}}} \left[\frac{1}{2} \delta_{x_s}^{x_{s'}} \beta (\bar{y}_s - \bar{y}_{s'})^2 \right] \\ &\quad + \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^Y \cup \{s\}}} \left[\frac{1}{2} Q_{s,s'} \bar{y}_s \bar{y}_{s'} \right] - \beta. \end{aligned} \tag{2.9}$$

Using standard calculus rules, one can show that:

$$\sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^X \cup \{s\}}} \frac{1}{2} \delta_{x_s}^{x_{s'}} \beta (\bar{y}_s - \bar{y}_{s'})^2 = \frac{1}{2} \bar{\boldsymbol{y}}^T P \bar{\boldsymbol{y}}, \tag{2.10}$$

where P is a matrix with elements:

$$P_{s,s'} = \begin{cases} 2 \sum_{s' \in \mathcal{N}_s^X} \delta_{x_s}^{x_{s'}} \beta, \text{ if } s = s', \\ -2 \delta_{x_s}^{x_{s'}} \beta, \text{ if } s' \in \mathcal{N}_s^X, \\ 0 \text{ otherwise.} \end{cases} \tag{2.11}$$

We have $l = |\mathcal{N}^X| + 2$. Note that, P approximates a diagonally banded matrix for two reasons. First, beyond a certain range (the $\mathcal{N}^X$ neighborhood), all the terms in P are null, starting from the diagonal. Second, among the remaining terms, some are null because $x_s \neq x_{s'}$. Hence, with the restriction $\beta \in \mathbb{R}_+^*$, P has the property of diagonal dominance which makes P a SPD matrix. If Q is also a SPD matrix, $R = P + Q$ is a SPD matrix[1]. The energy can then be

[1] The conditions on β and Q for the SPDness of R are respected, see Section 2.2.4.

written:

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{(s,s')\in\mathcal{S}^2} \left[\frac{1}{2} P_{s,s'} \bar{y}_s \bar{y}_{s'}\right] + \sum_{(s,s')\in\mathcal{S}^2} \left[\frac{1}{2} Q_{s,s'} \bar{y}_s \bar{y}_{s'}\right] \\ &\quad - \sum_{s\in\mathcal{S}} \sum_{\substack{s'\in \\ \mathcal{N}_s^X \cup \{s\}}} \delta_{x_s}^{x_{s'}} \beta, \\ &= \sum_{(s,s')\in\mathcal{S}^2} \left[\frac{1}{2} R_{s,s'} \bar{y} \bar{y}_{s'}\right] - \sum_{s\in\mathcal{S}} \sum_{\substack{s'\in \\ \mathcal{N}_s^X \cup \{s\}}} \delta_{x_s}^{x_{s'}} \beta. \end{aligned} \tag{2.12}$$

The last term of the energy gets canceled between the denominator and the numerator in Equation (3). Using the result of the integral of the multivariate Gaussian, we finally have:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N \det(R^{-1})}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{y}}^T R \bar{\boldsymbol{y}}\right), \tag{2.13}$$

which is the density of a Gaussian Markov Random Field (GMRF) with non-stationary mean $\boldsymbol{\mu}_{\boldsymbol{x}}$ and precision matrix $R = P + Q$ (Cressie and Verzelen, 2008) (Rue and Held, 2005). R corresponds to the precision matrix Q which is perturbed by P.

While we need to restrict $\beta \in \mathbb{R}_+^*$, typical values of β are much smaller than 1. Hence, in practice, the non-zero entries in P are much smaller than the entries of Q at the same site. Then, one assumes $R = P + Q \approx Q$. Since R is the precision matrix of a GMRF, it is sparse and so is Q. This approximation is the foundation for the parameter estimation procedure we propose, see Section 2.4.

2.2.4. The model parameters

We now detail the model parameters introduced in the previous section. The non-stationary mean vector $\boldsymbol{\mu}$, associated with the GMRF, is dependent on $\boldsymbol{x}$, the realizations of the field $\boldsymbol{X}$, such that $\forall s \in \mathcal{S}, \mu_s = \mu_{x_s} \in \mathbb{R}$.

$\beta \in \mathbb{R}_+^*$ is a coefficient whose role is similar to the granularity parameter in a classic Potts model (Pereyra et al., 2013).

Q is a SPD matrix which is approximated to be the precision matrix of the GMRF defined by the conditional likelihood, *i.e.*, $R \approx Q$, as explained in Section 2.2.3. This approximation leads to a much simpler parameter estimation procedure.

In order to use computationally efficient spectral methods for matrix manipulations, we also make a *periodic boundary assumption*. A detailed presentation of the periodic boundary assumption and its consequences is given in Appendix B.2. Then Q is a block-circulant matrix with circulant blocks whose inverse is the covariance matrix $\Sigma = Q^{-1}$. Σ is defined as a stationary covariance matrix with variance $\sigma \in \mathbb{R}_+^*$, and associated to an exponential correlation function with decay $r \in \mathbb{R}_+^*$ defined in Equation B.3. The Euclidean distance

on the torus (whose dimensions are the size of the image) is then used as defined in Equation B.11. The assumption of Markovianity for $(\boldsymbol{X}, \boldsymbol{Y})$, which leads to the Markovianity of $\boldsymbol{Y}$ given a realization of $\boldsymbol{X}$, implies that Q is a sparse matrix. This leads to efficient computations of the GPMF model equations.

Remark: The variance of the GMRF is stationary, as opposed to the mean of the GMRF. The simulation of the GMRF must be carried through its conditional equations (Rue and Held, 2005) (Brown et al., 2019) which are updated at each iteration since the neighbor values change. In this context it is known that introducing a non-stationary variance is very complex (Fuglstad et al., 2015) and is out of scope of this thesis.

2.3. Related models: the PMF model family

In this section we continue to consider $(\boldsymbol{X}, \boldsymbol{Y})$ as a Markov process and we present several models related to GPMF, and show that GPMF generalizes them. Using the conditioning formula we know that:

$$p(x_s, y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) = p(x_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) p(y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}, x_s). \tag{2.14}$$

Thus, the factor of Equation 2.4 can be split according to Equation 2.14 and, after normalization, a PMF has a distribution that can be written as:

$$p(\boldsymbol{x}, \boldsymbol{y}) \propto \prod_{s \in \mathcal{S}} \tilde{p}(x_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) \tilde{p}(y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}, x_s). \tag{2.15}$$

The last equation is a way to look at PMFs which highlights their greater generality and the more complex direct dependencies between variables than in HMFs (compare with Equation 2.3).

Then we propose new models based on Equation 2.14, by new factors with different conditional independence assumptions. These new intermediate models will be less general than the GPMF model but could also perform well at a smaller computational cost. We first present some new factors and then we use them to define models within the PMF model family with Equation 2.14.

Let us recall the Potts prior for Markov fields (Kato and Zerubia, 2012), it is based on the following local conditional probabilities:

$$\tilde{p}(x_s|\boldsymbol{x}_{\mathcal{N}_s^X}) = \exp\left(-2 \sum_{s' \in \mathcal{N}_s^X} V^P(x_s, x_{s'})\right), \tag{2.16}$$

where V^P is the Potts potential function (J. Besag, 1986) involving of a granularity coefficient β:

$$V^P(x_s, x_{s'}) = -\delta_{x_s}^{x_{s'}} \beta. \tag{2.17}$$

Let us also recall the independent Gaussian conditional likelihood (Kato and Zerubia, 2012) which leads to the local conditional probabilities:

$$\tilde{p}(y_s|x_s) = \exp\left(-\ln(\sqrt{2\pi}\sigma) - \frac{\bar{y}_s^2}{2\sigma^2}\right). \tag{2.18}$$

We also have the strict multivariate generalization of Equation 2.18 to the conditional GMRF likelihood in its local form (Rue and Held, 2005) (Brown et al., 2019):

$$\tilde{p}(y_s|x_s, \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) = \exp\left(-\ln\left(\sqrt{2\pi V_s}\right) - \frac{(y_s - M_s)^2}{2V_s}\right), \tag{2.19}$$

with

$$M_s = \mu_{x_s} - V_s \sum_{\substack{s' \in \mathcal{S} \\ s' \neq s}} Q_{s,s'} \bar{y}_{s'} \text{ and } V_s = Q_{s,s}^{-1}. \tag{2.20}$$

Let us introduce a generalized version of the Potts potential taking into account the spatial context:

$$\tilde{p}(x_s|y_s, \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) = \exp\left(-\left(2 \sum_{s' \in \mathcal{N}_s^X} -\delta_{x_s}^{x_{s'}} \beta \left(1 - \frac{1}{2}(\bar{y}_s - \bar{y}_{s'})^2\right)\right)\right). \tag{2.21}$$

Given the local conditional probabilities stated so far, three models belonging to the PMF family can be designed:

- *Potts-Independent Noise* (P-IN), with local distributions given by Equations 2.16 and 2.18, which is a HMF-IN model (Figure 2.1a). It corresponds to the most popular model of HMF-IN (Kato and Zerubia, 2012) (S. Z. Li, 2009).
- *Pairwise Markov Field with Uncorrelated Noise* (PMF-UN), with local distributions given by Equations 2.21 and 2.18 (Figure 2.1c). A similar model has already been studied in (Courbot, Mazet, et al., 2019).
- *Potts-Gaussian Markov Field* (P-GMRF)[2], with local distributions given by Equations 2.16 and 2.19, which belongs to the HMF-CN family (depicted in Figure 2.1d). We are not aware of similar models in the literature.

Table 2.1 summarizes the local distributions of these models. Note that GPMF, PMF-UN and P-GMRF all introduce more direct dependencies than the P-IN model. PMF-UN enhances the local distributions for the hidden variables and P-GMRF for the observed variables. Besides, note that all these models are submodels of the GPMF in the sense that they have direct dependencies that are ignored with respect to the GPMF direct dependencies. The submodels all ignore some direct dependencies that the pairwise factorization of Equation 2.4 integrates.

[2]Following Appendix C.1, it can be shown that the P-GMRF model arises from the joint energy E':

$$E'(\boldsymbol{x}, \boldsymbol{y}) = \sum_{s \in \mathcal{S}} \sum_{\substack{s' \in \\ \mathcal{N}_s^X}} -\delta_{x_s}^{x_{s'}} \beta + \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{N}_s^Y} \left[\frac{1}{2} Q_{s,s'} \bar{y}_s \bar{y}_{s'}\right].$$

Model	$p(x_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s} \boldsymbol{y}_{\mathcal{N}_s})$ factorizes using
P-IN	Eq. 2.16 and Eq. 2.18
PMF-UN	Eq. 2.21 and Eq. 2.18
P-GMRF	Eq. 2.16 and Eq. 2.19

Table 2.1.: Factorization of the related intermediate models.

2.4. Parameter estimation

2.4.1. Stochastic Parameter Estimation

In this section, we develop an algorithm for the unsupervised parameter estimation task in the GPMF model. Without loss of generality, in the following development, we consider $K = 2$, i.e. $\Omega = \{\omega_0, \omega_1\}$. Therefore the model is described with 5 parameters. Let $\boldsymbol{\theta} \in \Theta$ be the vector of parameters, then:

$$\begin{aligned} \boldsymbol{\theta} &= \{\mu_{\omega_0}, \mu_{\omega_1}, \beta, \sigma, r\} \triangleq \{\mu_0, \mu_1, \beta, \sigma, r\}, \\ \text{with } \Theta &= \mathbb{R}^2 \times (\mathbb{R}_+^*)^3. \end{aligned} \tag{2.22}$$

We develop a variation of the Stochastic Expectation Maximization algorithm (Celeux, 1985), which we call Stochastic Parameter Estimation (SPE). We first give the statistical estimators of the parameters given the *complete* data $(\boldsymbol{x}, \boldsymbol{y})$.

Generalization of the Linear Least Square estimator for GPMF

A generalization[3] of the approach of (Derin and H. Elliott, 1987) is established to retrieve the parameter β in the PMF family, using the Linear Least Square (LLS) estimator (J. Friedman et al., 2001) and a completed pair of realizations of $(\boldsymbol{x}, \boldsymbol{y})$. The derivation is done for the GPMF model and is similar for the other models.

First note that, $\forall s \in \mathcal{S}$:

$$\begin{aligned} \frac{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X} \vert \boldsymbol{y})}{p(\boldsymbol{x}_{\mathcal{N}_s^X} \vert \boldsymbol{y})} &= p(x_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}), \\ &= \frac{p(x_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{\sum_{x'_s \in \Omega} p(x'_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}, \\ &= \frac{\tilde{p}(x_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{\sum_{x'_s \in \Omega} \tilde{p}(x'_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}. \end{aligned} \tag{2.23}$$

Here the second equality has been seen in Equation 2.5. Then we have, $\forall s \in \mathcal{S}$:

$$\frac{\tilde{p}(x_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X} \vert \boldsymbol{y})} = \frac{\sum_{x'_s} \tilde{p}(x'_s, y_s \vert \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{p(\boldsymbol{x}_{\mathcal{N}_s^X} \vert \boldsymbol{y})}, \tag{2.24}$$

[3] This constitutes a generalization of the approach since we derive the estimators starting from the posterior distributions as opposed to the prior distribution as done in the original article.

where we can make the same key observation as in (Derin and H. Elliott, 1987): the right-hand side of the last equation is *independent of the realization* $x_s \in \Omega$. Then so is the left-hand side. Then, $\forall s \in \mathcal{S}$, $\forall (x_s, x'_s) \in \Omega^2$, we can write:

$$\begin{aligned} &\frac{\tilde{p}(x_s, y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})} = \frac{\tilde{p}(x'_s, y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{p(x'_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}, \\ \iff &\frac{\tilde{p}(x_s, y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})}{\tilde{p}(x'_s, y_s|\boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y})} = \frac{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}{p(x'_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}. \end{aligned} \tag{2.25}$$

Now, taking the exponential on each side, and using the expression of Equation 2.6, $\forall s \in \mathcal{S}$, $\forall (x_s, x'_s) \in \Omega^2$:

$$\begin{aligned} &\ln\left(\frac{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}{p(x'_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}\right) + \frac{1}{2}Q_{s,s}\bar{y}_s^2 + \\ &\sum_{t\in\mathcal{N}_s^Y} Q_{s,t}\bar{y}_s\bar{y}_t - \frac{1}{2}Q_{s,s}(\bar{y}'_s)^2 - \sum_{t\in\mathcal{N}_s^Y} Q_{t,s}\bar{y}'_s\bar{y}_t = \\ &2\sum_{t\in\mathcal{N}_s^X} V(x_s, y_s) - 2\sum_{t\in\mathcal{N}_s^X} V(x'_s, y_s), \end{aligned} \tag{2.26}$$

where we define $\bar{y}'_s = (y_s - \mu_{x'_s})$ which gives:

$$\begin{aligned} &\ln\left(\frac{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}{p(x'_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}\right) + \frac{1}{2}Q_{s,s}\bar{y}_s^2 + \\ &\sum_{t\in\mathcal{N}_s^Y} Q_{s,t}\bar{y}_s\bar{y}_t - \frac{1}{2}Q_{s,s}(\bar{y}'_s)^2 - \sum_{t\in\mathcal{N}_s^Y} Q_{t,s}\bar{y}'_s\bar{y}_t = \\ &\beta\Bigg(2\sum_{t\in\mathcal{N}_s^X} \delta_{x_s}^{x_t}\left(1 - \frac{1}{2}(\bar{y}_s - \bar{y}_t)^2\right) - \\ &\qquad 2\sum_{t\in\mathcal{N}_s^X} \delta_{x'_s}^{x_t}\left(1 - \frac{1}{2}(\bar{y}'_s - \bar{y}_t)^2\right)\Bigg). \end{aligned} \tag{2.27}$$

The last equation can be written for each site s, $\forall (x_s, x'_s) \in \Omega^2$. All these equations can be put in the form:

$$\boldsymbol{a} = \boldsymbol{b}\beta, \tag{2.28}$$

where $\beta \in \mathbb{R}_+^*$ and $\boldsymbol{a}, \boldsymbol{b}$ are real vectors with N elements with generic term, $\forall s \in \mathcal{S}$, $\forall (x_s, x'_s) \in \Omega^2$:

$$\begin{aligned} a_s = &\ln\left(\frac{p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}{p(x'_s, \boldsymbol{x}_{\mathcal{N}_s^X}|\boldsymbol{y})}\right) + \frac{1}{2}Q_{s,s}\bar{y}_s^2 + \\ &\sum_{t\in\mathcal{N}_s^Y} Q_{s,t}\bar{y}_s\bar{y}_t - \frac{1}{2}Q_{s,s}(\bar{y}'_s)^2 - \sum_{t\in\mathcal{N}_s^Y} Q_{t,s}\bar{y}'_s\bar{y}_t, \end{aligned} \tag{2.29}$$

and

$$b_s = \sum_{t \in \mathcal{N}_s^X} \left(2\delta_{x_s}^{x_t} - 2\delta_{x'_s}^{x_t} + (\mu_{x'_s} - \mu_{x_s})(-\bar{y}'_s - \bar{y}_s + 2\bar{y}_t) \right). \tag{2.30}$$

with $\bar{y}'_s = (y_s - \mu_{x'_s})$. The probabilities $p(x_s, \boldsymbol{x}_{\mathcal{N}_s^X} | \boldsymbol{y})$, $\forall x_s \in \Omega$, denote the probability of encountering a certain configuration of hidden variables at site s and its neighboring sites. These probabilities are independent of s and are estimated using the frequency estimator. The LLS estimator then gives that:

$$\hat{\beta} = (\boldsymbol{a}^T \boldsymbol{a})^{-1} \boldsymbol{a}^T \boldsymbol{b}. \tag{2.31}$$

Estimator for the other parameters

The Maximum Likelihood (ML) estimator is used to estimate μ_0, μ_1, σ. The expressions are, for $i \in \{0, 1\}$:

$$\hat{\mu}_i = \frac{1}{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{x_s = i\}}} \sum_{s \in \mathcal{S}} y_s \mathbb{1}_{\{x_s = i\}}, \tag{2.32}$$

and

$$\hat{\sigma} = \left(\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} (y_s - \hat{\mu}_{x_s})^2 \right)^{\frac{1}{2}}. \tag{2.33}$$

Let us now write an estimator of the range of the exponential correlation function, r, given the completed data. As given in (Cressie, 1992)[Section 2.4], the correlogram of the field can be estimated by:

$$\hat{C}(d) = \frac{1}{|\mathcal{D}(d)|} \sum_{(s,s') \in N(d)} \frac{1}{\hat{\sigma}^2} (y_s - \hat{\mu}_{x_s})(y_{s'} - \hat{\mu}_{x_{s'}}), \tag{2.34}$$

where $\mathcal{D}(d)$ is the set of pixel pairs whose Manhattan distance on the torus is $d \in \mathbb{N}^+$, *i.e.*:

$$\begin{aligned} \mathcal{D}(d) = \{(s, s') \in \mathcal{S}^2 : \ & \min(|s_1 - s'_1|, L_1 - |s_1 - s'_1|) + \\ & \min(|s_2 - s'_2|, L_2 - |s_2 - s'_2|) = d\}. \end{aligned} \tag{2.35}$$

where L_1 and L_2 are the lengths of the sides of the image we treat. The exponential correlation function has the form $u = e^{-\frac{d}{r}}$, where u is the correlation value, d the distance (between 0 and N) and r the range to estimate. There is no constant in front of the exponential since the GMRF has been standardized, so we will estimate $r \in \mathbb{R}^{+*}$ by fitting the exponential correlation function to the data points $\hat{C}(d)$ (Bjørnstad and Falck, 2001). Using the LLS estimator again we get:

$$\hat{r} = ((\boldsymbol{e}^T \boldsymbol{e})^{-1} \boldsymbol{e}^T \boldsymbol{w})^{-1}, \tag{2.36}$$

with $\boldsymbol{e}^T = \begin{pmatrix} O & \dots & N \end{pmatrix}$ and $\boldsymbol{w}^T = \begin{pmatrix} \log \hat{C}(0) & \dots & \log \hat{C}(N) \end{pmatrix}$.

In the context of latent variables, we only possess the observations and not the complete data. Thus the SPE algorithm successively repeats two steps. The first consists in simulating an estimation of the hidden layer, $\hat{\boldsymbol{x}}$, which is then used to form the *completed* data $(\hat{\boldsymbol{x}}, \boldsymbol{y})$, to approximate the complete data[4]. The second step uses the parameter estimators defined previously over the completed data. Algorithm 2.1 details the steps of the procedure.

SPE is a variant of the Stochastic Expectation Maximization (SEM) (Celeux, 1985). SPE differs from SEM because β and r are not estimated using a ML estimator. In SPE, at each iteration we sample from the posterior law to get the completed data, using Gibbs sampling, run during $K(t)$ iterations. We set $K(t) = t$ as proposed in (Carreira-Perpiñan and Hinton, 2005). In practice we stopped the SPE algorithm after 30 iterations, such a number was enough for the estimators to converge to a value.

Algorithm 2.1: The Stochastic Parameter Estimation (SPE) procedure to train the GPMF model.

Data: $\boldsymbol{\theta}^0 = \{\beta^0, \mu_0^0, \mu_1^0, \sigma^0, r^0\}$, the initial set of parameters,
$\boldsymbol{y}$, the observations,
$\hat{\boldsymbol{x}}^0$, the initial configuration for Gibbs sampler.
Result: $\hat{\boldsymbol{\theta}} = \{\hat{\beta}, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}, \hat{r}\}$, the estimated parameters.
$t \leftarrow 1$
while *convergence is not attained* **do**
  /* Posterior sampling with a Gibbs sampler initialized at $\hat{\boldsymbol{x}}^{t-1}$ and run during $K(t)$ steps */
  $\hat{\boldsymbol{x}}^t \sim p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}^{t-1})$
  /* Estimation */
  LLS estimator for β^t (Eq.2.31).
  ML estimator for μ_0^t and μ_1^t (Eq.2.32).
  ML estimator for σ^t (Eq.2.33).
  Estimation via correlogram for r^t (Eq.2.36).
  $\boldsymbol{\theta}^t \leftarrow \{\beta^t, \mu_0^t, \mu_1^t, \sigma^t, r^t\}$
  $t \leftarrow t + 1$
end

Remark: The same estimation procedure can be used to estimate the parameters of the other related models described in Section 2.3.

2.5. Experiments and Results

This section illustrates the models from the PMF family in the practical task of image segmentation in situations of various complexity. These models are also compared to other methods from the literature of unsupervised image segmentation.

[4]This step of inference is here carried stochastically since an exact analytical expression has not been established.

In this section we consider real-world images but also *semi-real* images. The latter are real binary images artificially corrupted by correlated noise, in order to test the relevance and tractability of the new model and its counterparts. The images used come from the "1070-Binary Shape Database"[5]. We want to evaluate the capacity of our model to handle correlated noise, the main purpose of this work, but also to test the parameter estimation procedures with the proposed SPE algorithm (Algorithm 2.1). Thus, from the natural binary image we construct an observation with an additive correlated Gaussian noise over the image. Such a noise is modeled by a GRF.

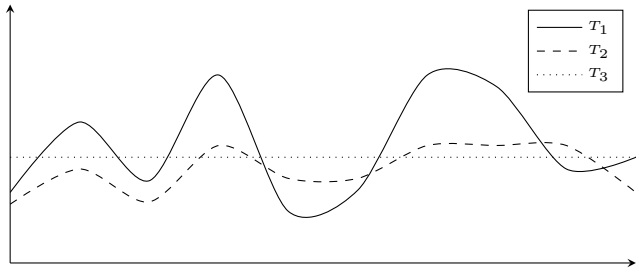
Beforehand, we discuss the efficiency of sampling from the probabilistic models.

2.5.1. Improved sampling with Tempered-Gibbs sampler

In this section we study more in depth the problematic of sampling from the new distributions. Indeed, it is known that the classic Gibbs sampler procedure (or other MCMC-based sampling approaches) suffers from a poor exploration of the probability distribution. The algorithm is moreover dependent on the initialization of the Gibbs sampler. Running several Gibbs sampler with different random initializations might not be satisfactory. These issues are discussed in (Neal, 1996) which also introduces *parallel tempering*. The idea of parallel tempering is to run in parallel several Gibbs samplers at different temperatures (in the same meaning as in Simulated Annealing (S. Geman and D. Geman, 1984) presented in Algorithm A.3). Samples from the Gibbs samplers at high temperature can swap and become the current state of a Gibbs sampler of a lower temperature. At high temperature the distribution is less severely peaked and the Gibbs sampler might explore more easily this distribution. This idea at the origin of parallel tempering is represented in Figure 2.2. This approach is called the Tempered Gibbs sampler (T-Gibbs sampler).

In our case we use the parallel tempering approach to improve the sampling procedures from the P-GMRF and GPMF models. We develop a similar methodology as (Cho et al., 2010) where the temperature factor was used to rescale some precise terms of the energy function rather than the whole energy. In the algorithm that we propose, the energy is tempered (multiplied by a positive scalar inferior to one) so that, at high temperature the probability distribution of the complex model tends towards the distribution of the P-IN model. We follow this idea since sampling is easier in the P-IN model. This uses the fact that the PMF models are more general than the P-IN model. To do so we modify the potentials and make them dependent on a temperature parameter that we explicitly add in the notation:

[5] https://vision.lems.brown.edu/content/available-software-and-databases



$$T_1 < T_2 < T_3 = \infty$$

Figure 2.2.: The parallel tempering approach: as the temperature increases, the probability distribution (of the family Equation 1.4) is flattened. It ultimately tends towards an uniform probability distribution as the temperature tends towards infinity. In a less multimodal distribution, the Gibbs sampler might more easily explore the distribution and produce interesting samples.

- Equations 2.6 and 2.7 become:

$$\tilde{p}_T(x_s, y_s | \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^X}) = \exp\Bigg(-\Big(2 \sum_{s' \in \mathcal{N}_s^X} V_T(x_s, y_s) + \frac{1}{2} Q_{s,s} \bar{y}_s^2 + \frac{1}{T} \sum_{s' \in \mathcal{N}_s^Y} Q_{s,s'} \bar{y}_s \bar{y}_{s'} \Big) \Bigg), \tag{2.37}$$

  where

$$V_T(x_s, y_s) = -\delta_{x_s}^{x_{s'}} \beta \left(1 - \frac{1}{2T} (\bar{y}_s - \bar{y}_{s'})^2 \right). \tag{2.38}$$

- Equation 2.19 becomes:

$$\tilde{p}_T(y_s | x_s, \boldsymbol{x}_{\mathcal{N}_s^X}, \boldsymbol{y}_{\mathcal{N}_s^Y}) = \exp\left(-\ln\left(\sqrt{2\pi V_s}\right) - \frac{(y_s - M_{T,s})^2}{2V_s} \right), \tag{2.39}$$

  with

$$M_{T,s} = \mu_{x_s} - \frac{V_s}{T} \sum_{\substack{s' \in \mathcal{S} \\ s' \neq s}} Q_{s,s'} \bar{y}_{s'} \text{ and } V_s = Q_{s,s}^{-1}. \tag{2.40}$$

The tempered versions of the potentials we have just defined are used to sample from the GPMF and P-GMRF model, in the T-Gibbs sampler, according to Table 2.2.

Finally, Algorithm 2.2 details the T-Gibbs sampler. In this algorithm, let K be the number of Gibbs samplers that we run in parallel. We then have K

Model	Equations used in T-Gibbs
P-GMRF	Eq. 2.16 and Eq. 2.39
GPMF	Eq.2.37

Table 2.2.: Equations used to sample from the models with the T-Gibbs.

associated and ordered temperatures starting from the temperature at which the probabilistic model is defined; which is 1 in our case. Two chains with successive temperatures can swap with probability:

$$\alpha^k(\boldsymbol{x}_k^{i+1}, \boldsymbol{x}_{k+1}^{i+1}) = \min\left(\exp\left((\frac{1}{T_k} - \frac{1}{T_{k+1}})(E(\boldsymbol{x}_k^{i+1}) - E(\boldsymbol{x}_{k+1}^{i+1}))\right), 1\right). \quad (2.41)$$

In our experiments, the set of temperatures is fixed, for all the PMF models, to 16 linearly spaced temperatures ranging from 1 to 20 (Salakhutdinov, 2009).

Algorithm 2.2: Tempered Gibbs sampler

Data: $\{T_k\}_{k=1}^{k=K}$, a set of ordered temperatures,
$\{\boldsymbol{x}_k^1\}_{k=1}^{k=K}$, initial states of the parallel Markov chains.
Result: $\boldsymbol{x}_1^i$, the sample at the model temperature.
$i \leftarrow 1$
while *not converged* **do**
  /* Gibbs sampler for each chain */
  for $k \in \{K, \ldots, 1\}$ **do**
    $\boldsymbol{x}_k^{i+1}$ drawn by Gibbs sampler from $p_{T_k}(\boldsymbol{x})$ with initialization at $\boldsymbol{x}_k^i$
  end
  /* Test to swap chains */
  for $k \in \{K-1, \ldots, 1\}$ **do**
    $\boldsymbol{x}_k^{i+1} \leftarrow \boldsymbol{x}_{k+1}^{i+1}$ with probability $\alpha^k(\boldsymbol{x}_k^{i+1}, \boldsymbol{x}_{k+1}^{i+1})$
  end
  $i \leftarrow i + 1$
end

The T-Gibbs sampler can replace the classic Gibbs sampler in the Marroquin algorithm for MPM computation (Marroquin et al., 1987).

Remark: The tempered Gibbs sampler is clearly much more computationally intensive than the classical Gibbs sampler. This constitutes an important current drawback despite the improved performance that we will study in the next section.

Remark: An additional experiment characterizing the T-Gibbs sampler is available in Appendix C.2.

2.5.2. Supervised segmentation of semi-real images with the PMF models

In this section the task of image segmentation is performed with the models from the PMF family in the context of supervised segmentation. The pair of complete data $(\boldsymbol{x}, \boldsymbol{y})$ is then available to estimate $\hat{\boldsymbol{\theta}}$ from it (with the estimators described in Section 2.4) before estimating $\hat{\boldsymbol{x}}$ using $\boldsymbol{y}$ and $\hat{\boldsymbol{\theta}}$.

The PMF models are here tested in the case of supervised image segmentation according to the two segmentation algorithms MAP, MPM. The MPM will be computed either with the Marroquin's algorithm (Marroquin et al., 1987) or the T-Gibbs algorithm. We compute the average error rate for each PMF model in the segmentation of a series of images from the dataset with varying noise levels. Recall that the noise is modeled by the realization of a GRF for the semi-real image formation. In the first case, the noise parameter that is varying is $\Delta\mu = |\mu_1 - \mu_0|$, in the second case, the range r of the noise is the variable. The experiments are summed up in Figures 2.3 and 2.4 and we now describe and interpret the results.

First of all, we note the overall reduced error rate thanks to the new pairwise models over the classical P-IN model. The GPMF model leading to the best segmentations for all noise levels. This behavior illustrates the capacity for the new model to take into account the correlated noise.

Note that in the case of varying ranges, for the two segmentation criteria (especially for the MAP), the GPMF and P-GMRF models perform worse than the P-IN model at the smallest correlation ranges. Two reasons might explain the phenomenon. First, the model definitions: when $r \to 0$, the GPMF model does not totally tend towards the P-IN model, for example. Second, convergence issues with the Gibbs sampler can happen: the error rates vary importantly between the three algorithms used to compute the criteria[6]. The MPM computed with T-Gibbs exhibits the best error rates for GPMF and P-GMRF (as expected from an improved sampling procedure). Notably, in the critical zone where $r < 2$, the averaged error rates are increased by several points for these two models. This issue is, however, unlikely to appear in practical cases when the noise is sure to be correlated, such as our applications. Indeed, we here highlight a theoretical drawback linked with the limits of stochastic Gibbs sampling.

2.5.3. Unsupervised segmentation on semi-real images

In this section, we address the problem of unsupervised image segmentation. We use the semi-real images previously introduced but consider only the observations $\boldsymbol{Y} = \boldsymbol{y}$. We here consider the GPMF model and the MPM criterion computed with Marroquin's algorithm: it is the best compromise between segmentation performance and aomputation time. To compare the GPMF model, we used three approaches developed for unsupervised image segmentation:

[6] MPM has already been recognized as more stable in practical applications that the MAP estimator (Courbot, Mazet, et al., 2019).

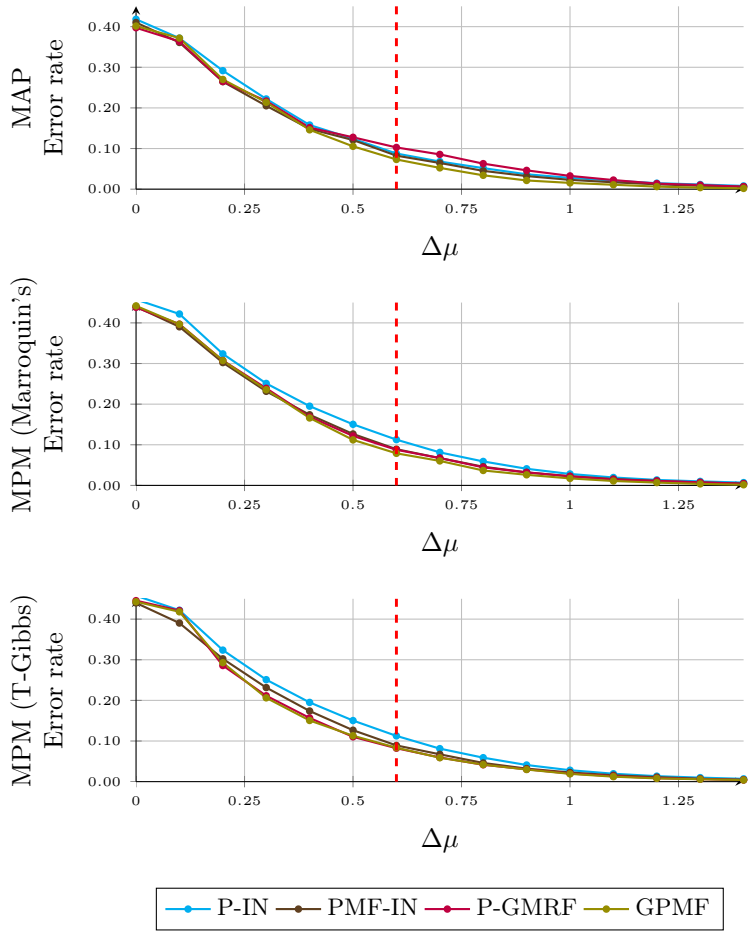


Figure 2.3.: Error rate in function of varying $\Delta\mu$ parameters and of several estimators in the supervised segmentation of the 'dude' images of the dataset. Each row, from top to bottom, respectively corresponds to the MAP, MPM (Marroquin's algorithm) and MPM (T-Gibbs algorithm) segmentation criteria. The other parameters of the GRF noise were fixed to $\sigma = 0.5$ and $r = 2$. The dashed red lines are common between the graphs of Figure 2.4, for each row respectively.

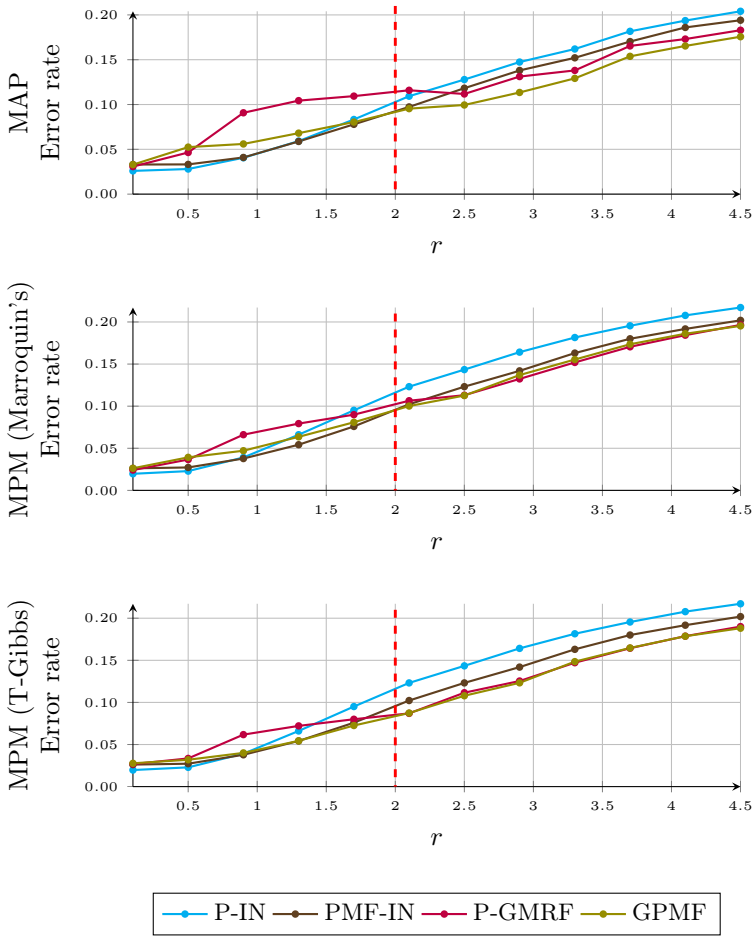


Figure 2.4.: Error rate in function of varying r parameters and of several estimators in the supervised segmentation of the 'dude' images of the dataset. Each row, from top to bottom, respectively corresponds to the MAP, MPM (Marroquin's algorithm) and MPM (T-Gibbs algorithm) segmentation criteria. The other parameters were fixed to $\mu_0 = 0, \mu_1 = 0.6$ and $\sigma = 0.5$. The dashed red lines are common between the graphs of Figure 2.3, for each row respectively.

- The classical Hidden Markov Field model with Potts prior and Independent Noise (P-IN) (S. Z. Li, 2009) (Kato and Zerubia, 2012), which is also the member of the PMF model family with the less direct dependencies between random variables (see Section 2.3).
- The KMeans clustering[7] algorithm (Arthur and Vassilvitskii, 2007).
- The pyImSegm[8] (pyIS) segmentation algorithm proposed in (Borovec et al., 2017). The core of this technique relies on a Markov Random Field energy minimization problem using super-pixel based and graph-cut based approaches.
- Based on the recent developments of (Rubel et al., 2018) for correlated noise reduction (which were not available online), we propose to combine the Block-Matching 3D filter[9] from (Dabov et al., 2009) and the graph-cut algorithm[10] from (Boykov and Kolmogorov, 2004). This approach is then the combination of a popular noise reduction technique and a popular graph-cut based segmentation. We call this approach BM3D+GC.

Let us now compare the unsupervised segmentations obtained with the GPMF model with the results given by other models. Figure 2.5 illustrates the segmentation performance of the models for a varying noise level. Figure 2.6 depicts some segmentations of images from the database. We notice, in all cases, the superiority of the new probabilistic model GPMF in its capacity to handle the correlated noise, improving the segmentation error and giving stable results. The GPMF model always performs best or equally best at all noise levels. In the best scenarii, the GPMF model increases the average error rate by about four points. The BM3D+GC and pyIS approaches are more unstable and perform worse when dealing with an image corrupted with correlated noise, despite manually tuned hyperparameters. Finally, note that, as expected, the overall error rates are higher in this unsupervised segmentation experiment than in the supervised segmentation experiment of Section 2.5.2.

2.5.4. On real world images

We now present the model in a real world application from the medical field, where unsupervised image segmentation needs to be done on strongly spatially corrupted data. The context of these experiments has been presented in Section 2.1. To the best of our knowledge, there is no counterpart of our model in the medical literature involving spatially-correlated noise for stent segmentation in medical images. We treat 512×512-pixel 2D images.

The goal of the experiment is to segment precisely the organic material from the background of the images. The two classes we wish to distinguish are visible by the naked eye, but they are very corrupted by the artifacts and

[7] Implementation from https://opencv.org.
[8] Implementation from https://github.com/Borda/pyImSegm.
[9] Implementation from http://www.cs.tut.fi/~foi/GCF-BM3D/.
[10] Implementation from https://github.com/shaibagon/GCMex/.

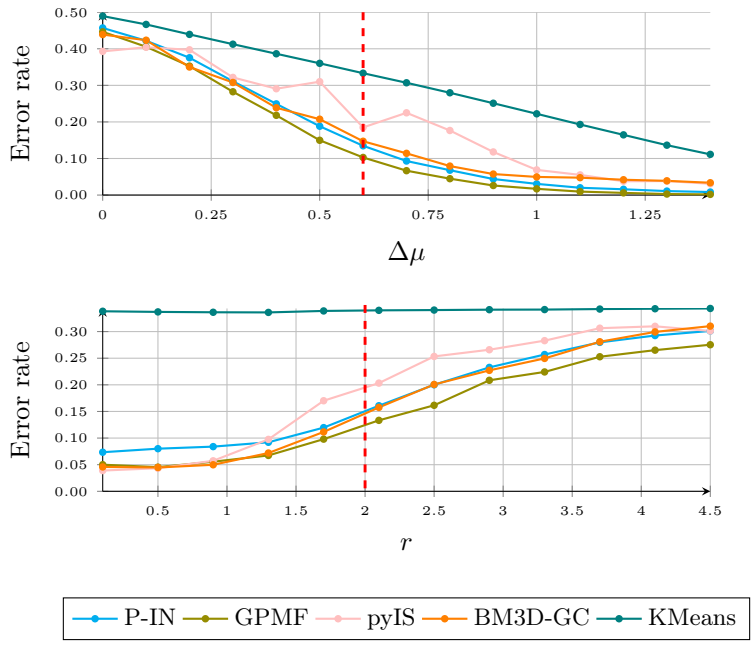


Figure 2.5.: MPM (Marroquin's algorithm) unsupervised segmentation over the 'dude' images of the dataset. For the $\Delta\mu$ row, the other parameters of the GRF noise were fixed to $\sigma = 0.5$ and $r = 2$. For the r row, they were fixed to $\mu_0 = 0, \mu_1 = 0.6$ and $\sigma = 0.5$. The dashed red lines highlight the situations when the parameter configuration is identical in both graphs. As expected, in such cases, we notice that the models have identical performances relative to each other.

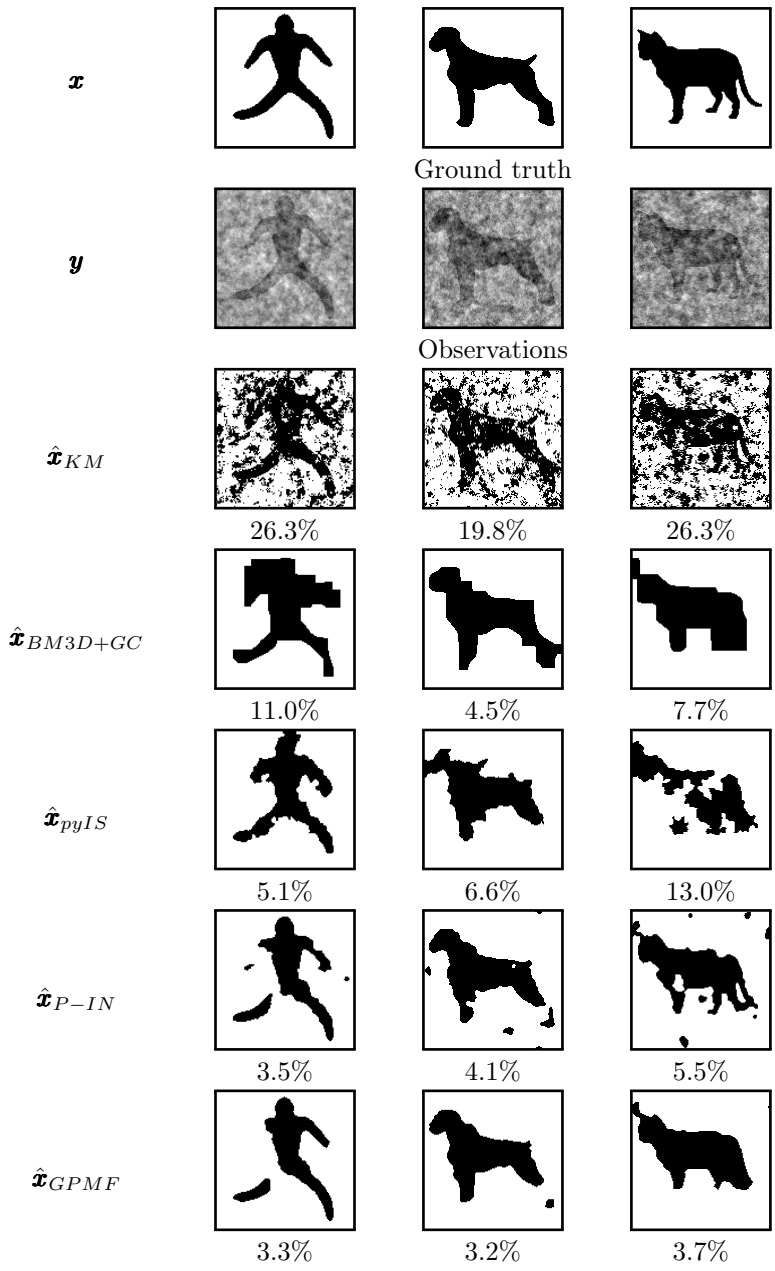


Figure 2.6.: Unsupervised segmentation of images from the dataset with the new model and 4 other models. The segmentation criterion is MPM with Marroquin's algorithm. The error rates with respect to the ground truth appear below each sample. For the 3 images, the noise parameters are $\mu_0 = 0$, $\mu_1 = 2$, $\sigma = 1$, $r = 3$.

	FN	FP
BM3D+GC	0.14	**0.01**
P-IN	**0.08**	0.08
GPMF	**0.08**	0.04

(a) Case 1

	FN	FP
BM3D+GC	0.05	**0.07**
P-IN	**0.02**	0.14
GPMF	**0.02**	**0.07**

(b) Case 2

Table 2.3.: FN and FP rates computed in the blue areas, for each model, for each case of Figure 2.7.

are challenging to segment using automatic unsupervised algorithms. Such a processing would help identify the nature of the elements in contact with the stent, to precisely understand the possible interactions of the stent that could have lead to its failure. A dedicated model needs to be created since the artifacts hide the underlying material.

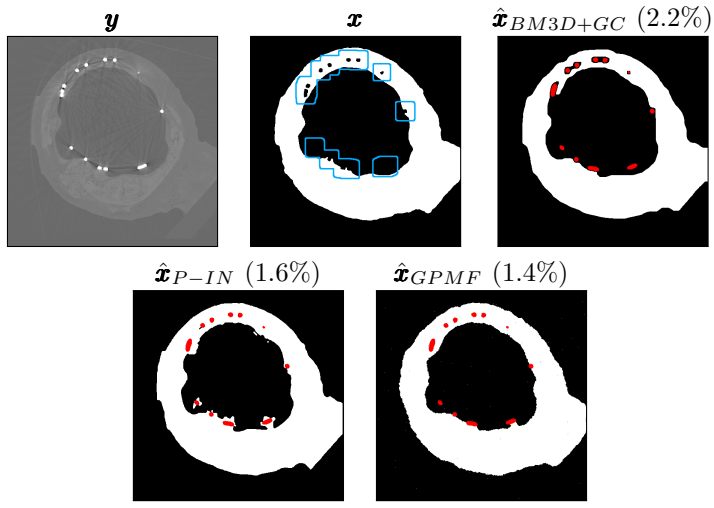
Figure 2.7 depicts the results obtained by three methods described from the previous section (BM3D+GC, P-IN and GPMF). The results of the pyIS method were essentially similar to the results of BM3D+GC and are omitted here for brevity, as well as the results form the KMeans algorithm which are omitted because of the poor performance of the algorithm studied in Section 2.5.3. The P-IN and the BM3D+GC results are particularly prone to misclassifications because of the artifacts.

While the overall error rate is in favor of the GPMF model, it does not truly reflects the capacity of the model to resolve correlated noise and to offer a proper segmentation. Therefore we consider the organic material as the *true* class and the background as the *false* class. We then compute the False Negative (FN) and False Positive (FP) percentages of pixels in the areas of interest, around the stent where the correlations are the strongest. Those data are given in Table 2.3. We can see that the GPMF model best captures the correlated noise and resolves much of the stent artifact problem: it offers the best compromise between the FN and FP scores. In addition the results obtained with the GPMF model were also judged very satisfying by the pathologists.

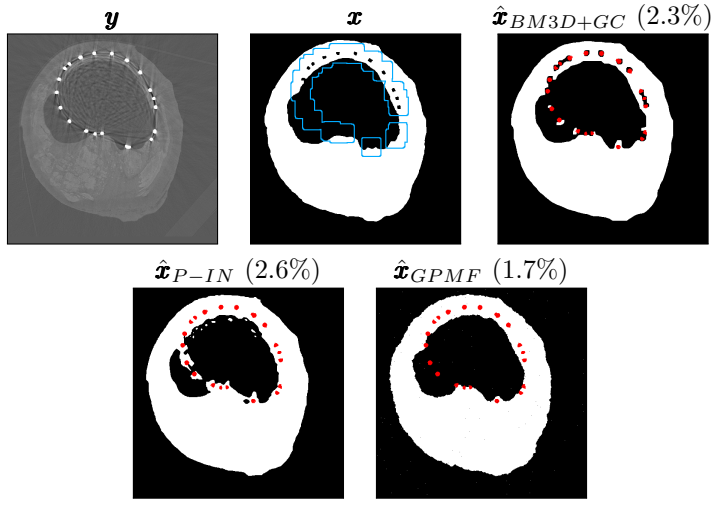
The GPMF model and its application to real world images is further discussed in Chapter 5.

2.6. Conclusion

In this chapter we proposed a new probabilistic model and two new submodels which belong to the PMF family. We have seen that the PMF assumption can lead to modelization embedding much more complex correlations between random variables than the modelizations based on HMFs. This was used to construct a generative latent variable model with a GMRF based likelihood. By proposing a new parameter estimation procedure, the GPMF model then offered the best results, in the unsupervised segmentation of images corrupted with strongly correlated noise, when compared with other classical unsupervised segmentation techniques.



(a) Case 1



(b) Case 2

Figure 2.7.: Unsupervised segmentations of organic material in corrupted X-rays images. For these 2 cases, which illustrate different explanted stents, we have: the real image $\boldsymbol{y}$, the ground truth image $\boldsymbol{x}$, the BM3D-GC segmentation, the P-IN segmentation and the GPMF segmentation. The segmentation criterion for the probabilistic models is the MPM. The stent parts (brightest pixels in $\boldsymbol{y}$) were segmented beforehand by a thresholding technique and then considered as image borders. They appear in red on the segmented images and did not take part in the segmentation. The ground truth could be provided by experts since a histological analysis is available. Error rates with respect to the ground truth on the whole image appear in parenthesis.

As an application of this model to real world data, we were able to solve the problem of organic material segmentation in an image where strongly correlated noise is caused by metallic biomaterials. The GPMF model deals well with correlated noise and limits the number of misclassifications, as opposed to methods unsensitive to the spatial context of pixels.

Chapter 3.

Spatial Triplet Markov Trees

Contents

3.1. Introduction

This chapter introduces probabilistic models that belong to the Directed Graphical Model (DGM) family. We will study both direct and approximate inference. We have seen in the previous chapters some difficulties that arise when inference must be approximate. That is why, the main focus of this chapter is to develop a new model called Spatial Triplet Markov Tree (STMT) in which we elaborate enhanced correlations with the help of auxiliary random variables, but we preserve the possibility of performing the inference directly.

The enhancement of correlations in the Markov Tree (MT) structure are *local*, *spatial*, in the same sense of the correlations found in classical Markov Random Fields models (such as the models from the previous chapter). The STMT model is then particularly suited for image segmentation, we show that it performs better than the classical MT model. The presentation of the new model and its application to image segmentation is done in the first part of this chapter.

In the second part of the chapter, Variational Inference (VI) is developed more in depth. VI (Lauritzen and Spiegelhalter, 1988) is an approach to perform (approximate) inference with a complex probability distribution where exact computations are not available. VI recasts the inference problem as an optimization problem which relies on a simpler distribution, the *variational distribution*, used to approximate the complex one. Additional approaches to VI consist in constructing a variational distribution with a certain complexity in itself (direct dependencies between random variables) which might, moreover, involve auxiliary random variables. Such developments might lead to further improvements in the inference (Agakov and Barber, 2004). We will apply these principles and we will also use VI as a way to better understand the STMT model.

3.2. Markov Tree models

This section focuses on particular MT models. The purpose of this section is to present the models that will be used in the context of Bayesian image segmentation (Section 1.5.2). We continue to denote $\boldsymbol{X}$ as the hidden process and $\boldsymbol{Y}$ as the observed process. We focus our work on dyadic and quadtree MTs whose layers of sites are $\mathcal{S} = \{\mathcal{S}^1, \ldots, \mathcal{S}^L\}$.

3.2.1. Hidden Markov Trees

Recall the introduction of HMTs made in Section 1.6.1. The joint distribution of a HMT with Independent Noise (HMT-IN) is given in Equation 1.24. Figure 1.5 depicts the graphical model of a dyadic HMT-IN.

As mentioned earlier, the inference tasks can be done with exact computations in HMTs. This can be done with the Upward Downward (UD) algorithm[1] (Monfrini, Lecomte, et al., 2003) (J.-B. Durand, Goncalves, et al.,

[1] The UD algorithm can be interpreted in two ways. Either as a generalization of the Forward

Algorithm 3.1: Upward-Downward algorithm

Data: $\boldsymbol{y}$, a realization of the observed process,
$p(x_s|x_{s^-}), \forall s \in \bar{\mathcal{S}}$, transitions of the hidden process,
$p(x_r), r \in \mathcal{S}^1$, distribution at the root vertice.

Result: $p(x_s|x_{s^-}, \boldsymbol{y}), \forall s \in \bar{\mathcal{S}}$, posterior transitions,
$p(x_s|\boldsymbol{y}), \forall s \in \mathcal{S}$, posterior marginals.

/* Compute recursively $\beta(x_s) = p(\boldsymbol{y}_{\boldsymbol{s}^{++}}|x_s)$, $\forall x_s \in \Omega$, $\forall s \in \mathcal{S}$ */

$$\begin{cases} \beta(x_s) = p(y_s|x_s), \text{ if } s \in \mathcal{S}^N, \\ \beta(x_s) = \prod_{t \in \boldsymbol{s}^+} \left(\sum_{x_t} \beta(x_t) p(x_t|x_s) \right) \text{ otherwise.} \end{cases} \tag{3.1}$$

/* Compute the remaining posterior transitions $\beta(x_s)$, $\forall (x_s, x_{s^-}) \in \Omega^2$, $\forall s \in \bar{\mathcal{S}}$ */

$$p(x_s|x_{s-}, \boldsymbol{y}) = \frac{\beta(x_s) p(x_s|x_{s-})}{\sum_{x_s} \beta(x_s) p(x_s|x_{s-})}, \tag{3.2}$$

/* Compute the posterior marginal at root, $\forall x_r \in \Omega$ */

$$p(x_r|\boldsymbol{y}) = \frac{\beta(x_r) p(x_r)}{\sum_{x_r} \beta(x_r) p(x_r)}. \tag{3.3}$$

/* Compute the remaining posterior marginals, $\forall x_s \in \Omega, \forall s \in \bar{\mathcal{S}}$ */

$$p(x_s|\boldsymbol{y}) = \sum_{x_{s^-}} p(x_{s^-}|\boldsymbol{y}) p(x_s|x_{s^-}, \boldsymbol{y}). \tag{3.4}$$

2004). It enables structured parallel retrieval of the posterior transition distributions $(p(x_s|x_{s^-}, \boldsymbol{y}), \forall s \in \mathcal{S})$ and of the posterior marginal distributions $(p(x_s|\boldsymbol{y}), \forall s \in \mathcal{S})$. Algorithm 3.1 describes the UD algorithm which makes use of an intermediate quantity $\beta(x_s) = p(\boldsymbol{y}_{\boldsymbol{s}^{++}}|x_s)$, $\forall x_s \in \Omega$, $\forall s \in \mathcal{S}$, where $\boldsymbol{s}^{++}$ denotes the set of all the descendants of s (vertices that can be reached from s by following the directed edges) (Pieczynski, 2002).

Remark: Algorithm 3.1 is written for the HMT-IN model (Monfrini, Lecomte, et al., 2003) (Laferté et al., 2000) where observed variables are associated to hidden variables on the last layer only (as drawn in Figure 1.5). This algorithm is easily adapted if there is no, more or less observed variables (J.-B. Durand, Goncalves, et al., 2004).

3.2.2. Spatial Triplet Markov Trees

Spatial Triplet Markov Trees (STMTs) is a model of MTs based on a triplet Markov process that we have defined in Section 1.6.2. Along with the previously

Backward algorithm used for Hidden Markov Chains, or as a principled message passing schedule in the Belief Propagation algorithm.

introduced processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, let $\boldsymbol{V}$ be a discrete-valued random process with values in Λ^N. $\boldsymbol{V}$ plays the role of an *auxiliary* process. In the Bayesian approach to image segmentation that we follow, $\boldsymbol{V}$ is part of the *hidden* process, along with $\boldsymbol{X}$. Let $\boldsymbol{T}$ be a process such that $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{Y})$. STMTs then have for general distribution:

$$p(\boldsymbol{t}) = p(\boldsymbol{t}_r) \prod_{s \in \bar{\mathcal{S}}} p(\boldsymbol{t}_s | \boldsymbol{t}_{s^-}). \tag{3.5}$$

Equation 3.5 reveals (similarly to Equation 1.27 in the case of chains) the triplet assumption: $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{Y})$ is a Markov process.

A generalized version of the UD algorithm (Algorithm 3.1) can be proposed for STMTs: it suffices to replace the occurences of the $\boldsymbol{X}$ process by the new hidden process $(\boldsymbol{X}, \boldsymbol{V})$. Then, to recover information on the original process $\boldsymbol{X}$, one needs to marginalize over the auxiliary random variables, $\forall s \in \mathcal{S}$:

$$p(x_s) = \sum_{v_s \in \Lambda} p(x_s, V_s = v_s). \tag{3.6}$$

The rest of this section is dedicated to the definition of the STMT model in its dyadic and quadtree versions[2]. The STMT model takes advantage of the triplet structure for exact inference with UD. It also takes advantage of auxiliary random variables to introduce much richer correlations between random variables as compared to the classical HMT model. More precisely, the model aims at introducing *local*, *spatial* correlations similarly to MRFs (Section 1.6.1). However, as already mentioned, the model preserves direct computations of the inference tasks, as opposed to MRFs.

Remark: The STMT model will be described in a general form containing the visible process $\boldsymbol{Y}$ only at the last resolution and the Independent Noise hypothesis will always hold.

The dyadic model

We now describe the construction of the dyadic STMT model whose graphical representation is given in Figure 3.1 (in case of Independent Noise). We consider that the observations are only associated with vertices of the lowest layer, with the hypothesis of the independent noise.

A node at site $s \in \bar{\mathcal{S}} \setminus \mathcal{S}^L$ is a triplet $(X_s, V_s^{\leftarrow}, V_s^{\rightarrow})$; we then assume that we have the factorization:

$$\begin{aligned} p(\boldsymbol{t}_s | \boldsymbol{t}_{s^-}) &= p(x_s, \boldsymbol{v}_s | x_{s^-}, \boldsymbol{v}_{s^-}), \\ &= p(x_s | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\leftarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\rightarrow} | x_{s^-}, \boldsymbol{v}_{s^-}). \end{aligned} \tag{3.7}$$

However, a node at site $s \in \mathcal{S}^L$ is a quadruplet $(X_s, V_s^{\leftarrow}, V_s^{\rightarrow}, Y_s)$; we then

[2]This section is a new model which builds upon and redefines the only occurence of the STMT prior to our work, found in (Courbot, Monfrini, et al., 2018).

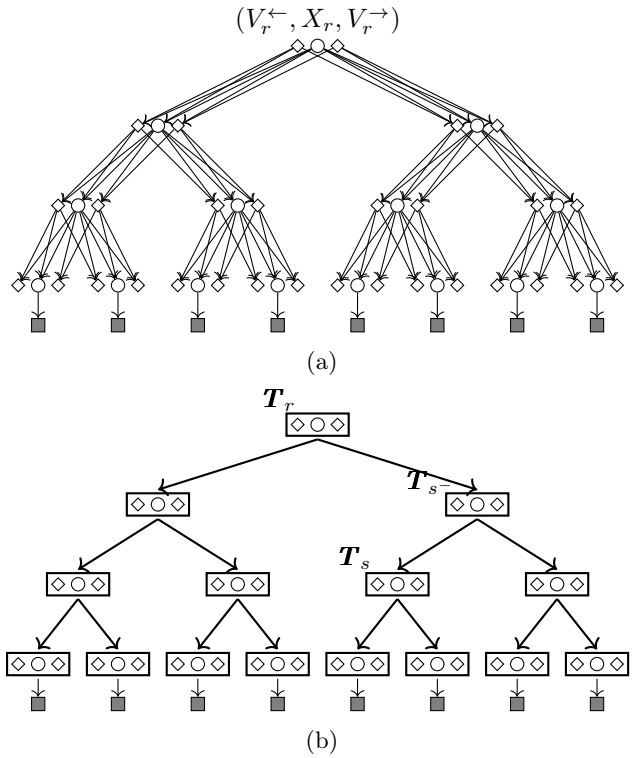


Figure 3.1.: Graphical model corresponding to a dyadic STMT with Independent Noise. (a) depicts all the correlations, (b) gives a condensed view highlighting the preserved MT structure. In (b), a node $\boldsymbol{T}_s$, its father $\boldsymbol{T}_{s^-}$ and the root node $\boldsymbol{T}_r$ have been annotated. The shadowed squares correspond to the variables from the observed process.

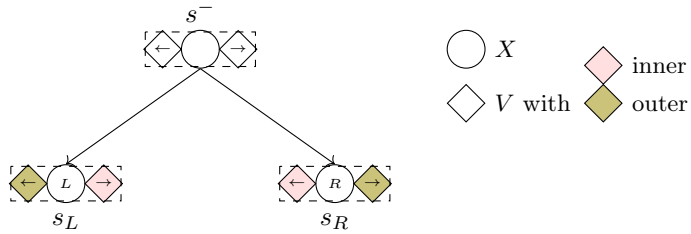


Figure 3.2.: Details of the STMT construction: a father node s^- linked to its 2 sons (s^L, s^R). The directionality of the V_s is specified as well as their type (inner or outer).

assume that we have the factorization:

$$\begin{aligned} p(\boldsymbol{t}_s|\boldsymbol{t}_{s-}) &= p(x_s, \boldsymbol{v}_s, y_s|x_{s^-}, \boldsymbol{v}_{s^-}, y_{s^-}), \\ &= p(x_s, \boldsymbol{v}_s|x_{s^-}, \boldsymbol{v}_{s^-}, y_{s^-})p(y_s|x_s, \boldsymbol{v}_s, x_{s^-}, \boldsymbol{v}_{s^-}, y_{s^-}), \\ &= p(x_s, \boldsymbol{v}_s|x_{s^-}, \boldsymbol{v}_{s^-})p(y_s|x_s), \\ &= p(x_s|x_{s^-}, \boldsymbol{v}_{s^-})p(v_s^{\leftarrow}|x_{s^-}, \boldsymbol{v}_{s^-})p(v_s^{\rightarrow}|x_{s^-}, \boldsymbol{v}_{s^-})p(y_s|x_s). \end{aligned} \tag{3.8}$$

To further define the transition laws of each of the variables, we define the notion of *inner* and *outer* variables for the V variables. We define that, within *Left* (L) (resp. *Right* (R)) sons, $V_{s^L}^{\leftarrow}$ (resp. $V_{s^R}^{\rightarrow}$) is an outer variable and $V_{s^L}^{\rightarrow}$ (resp. $V_{s^L}^{\leftarrow}$) is an inner variable. Figure 3.2 illustrates these concepts for a particular node.

We now detail Equations 3.7 and 3.8, according to the variable type (*left*, *right*, *inner* or *outer*). For X_s sons:

$$\begin{cases} p(x_s^L|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(x_s^L|X_{s^-} = x_{s^-}, V_{s^-} = v_{s^-}^{\leftarrow}), \\ p(x_s^R|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(x_s^R|X_{s^-} = x_{s^-}, V_{s^-} = v_{s^-}^{\rightarrow}), \end{cases} \tag{3.9}$$

for inner V_s sons:

$$\begin{cases} p(v_{s^L}^{\rightarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(v_{s^L}^{\rightarrow}|x_{s^-} = x_{s^-}, V_{s^-} = v_{s^-}^{\rightarrow}), \\ p(v_{s^R}^{\leftarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(v_{s^R}^{\leftarrow}|X_{s^-} = x_{s^-}, V_{s^-} = v_{s^-}^{\leftarrow}), \end{cases} \tag{3.10}$$

and for outer V_s sons:

$$\begin{cases} p(v_{s^L}^{\leftarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(v_{s^L}^{\leftarrow}|X_{s^-} = v_{s^-}^{\leftarrow}, V_{s^-} = x_{s^-}), \\ p(v_{s^R}^{\rightarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) &= p(v_{s^R}^{\rightarrow}|X_{s^-} = v_{s^-}^{\rightarrow}, V_{s^-} = x_{s^-}). \end{cases} \tag{3.11}$$

Remark: The conditioning of the V_s variables can be seen as being the same conditioning the nearest X neighbor of X_s on a same resolution. The triplet tree framework then provides a way to simulate X variables conditionally to the realizations of V variables which behave similarly to the neighboring X variables. This is notable since such links are strictly impossible in classical

Type of son at s	Inner V_s sons at s	Outer I V_s sons at s	Outer II V_s sons at s
NW	$V_s^{\rightarrow}, V_s^{\searrow}, V_s^{\downarrow}$	$V_s^{\leftarrow}, V_s^{\nwarrow}, V_s^{\uparrow}$	$V_s^{\nearrow}, V_s^{\swarrow}$
NE	$V_s^{\downarrow}, V_s^{\swarrow}, V_s^{\leftarrow}$	$V_s^{\uparrow}, V_s^{\nearrow}, V_s^{\rightarrow}$	$V_s^{\nwarrow}, V_s^{\searrow}$
SE	$V_s^{\leftarrow}, V_s^{\nwarrow}, V_s^{\uparrow}$	$V_s^{\rightarrow}, V_s^{\searrow}, V_s^{\downarrow}$	$V_s^{\nearrow}, V_s^{\swarrow}$
SW	$V_s^{\uparrow}, V_s^{\nearrow}, V_s^{\rightarrow}$	$V_s^{\leftarrow}, V_s^{\downarrow}, V_s^{\swarrow}$	$V_s^{\nwarrow}, V_s^{\searrow}$

Table 3.1.: Definitions of the inner, outer I and outer II V_s sons for the quadtree model.

MTs.

Finally for the root node $s \in \mathcal{S}^1$, $p(x_s, \boldsymbol{v}_s)$ is chosen as a discrete distribution on $\Omega \times \Lambda^2$.

The quadtree model

We now describe the construction of the quadtree STMT. Again, we consider that the observations are only associated with vertices of the lowest layer, with the hypothesis of the independent noise.

Each father s^- has four sons s_{NW}, s_{NE}, s_{SE} and s_{SW}, where NW, NE, SE and SW stand respectly for *North West*, *North East*, *South East* and *South West*. For example, at site s_{NW}, the random variables are $X_{s_{NW}}$ and $\boldsymbol{V}_{s_{NW}}$. The composition of the octuplet $\boldsymbol{V}_{s_{NW}}$ is again spatially described by $\boldsymbol{V}_{s_{NW}} = (V^{\leftarrow}_{s_{NW}}, V^{\nwarrow}_{s_{NW}}, V^{\uparrow}_{s_{NW}}, V^{\nearrow}_{s_{NW}}, V^{\rightarrow}_{s_{NW}}, V^{\searrow}_{s_{NW}}, V^{\downarrow}_{s_{NW}}, V^{\swarrow}_{s_{NW}})$. Similarly to the dyadic case we define *inner* and *outer* V variables for each type of father node. However the *outer* type is now divided in two categories *outer I* and *outer II*. The definitions are given in Table 3.1.

At a site $s \in \bar{\mathcal{S}}$, should it be *North West* or *North East* or *South East* or *South West*, we find the couple $(X_s, \boldsymbol{V}_s)$. The transition distributions are similar to Equation 3.7 (when $s \in \bar{\mathcal{S}} \setminus \mathcal{S}^L$) and Equation 3.8 (when $s \in \mathcal{S}^L$, case with observed random variables[3]), however $p(x_s, \boldsymbol{v}_s | x_{s^-}, \boldsymbol{v}_{s^-})$ is now assumed to factorize as:

$$\begin{aligned} p(x_s, \boldsymbol{v}_s | x_{s^-}, \boldsymbol{v}_{s^-}) = & p(x_s | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\leftarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\nwarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) \\ & p(v_s^{\uparrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\nearrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\rightarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) \\ & p(v_s^{\searrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\downarrow} | x_{s^-}, \boldsymbol{v}_{s^-}) p(v_s^{\swarrow} | x_{s^-}, \boldsymbol{v}_{s^-}). \end{aligned} \tag{3.12}$$

This factorization is independent of the type of the father. However, we will now specify the conditioning variables in these transition laws and the latter will be dependent on the type of the father as in the dyadic case. Thus, we will propagate the homogeneity of the realizations at one resolution.

[3] In such case the Y process is taken into account and we consider the probability of a decuplet $(X_s, \boldsymbol{V}_s, Y_s), \forall s \in \mathcal{S}^L$.

Each father s^- has four sons s_{NW}, s_{NE}, s_{SE} and s_{SW}, with, for the X_s sons:

$$\begin{cases} p(x_{s_{NW}}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(x_{s_{NW}}|x_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\leftarrow}, v_{s^-}^{\nwarrow}, v_{s^-}^{\uparrow})), \\ p(x_{s_{NE}}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(x_{s_{NE}}|x_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\uparrow}, v_{s^-}^{\nearrow}, v_{s^-}^{\rightarrow})), \\ p(x_{s_{SE}}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(x_{s_{SE}}|x_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\rightarrow}, v_{s^-}^{\searrow}, v_{s^-}^{\downarrow})), \\ p(x_{s_{SW}}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(x_{s_{SW}}|x_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\downarrow}, v_{s^-}^{\swarrow}, v_{s^-}^{\leftarrow})). \end{cases} \tag{3.13}$$

We now detail the transition laws for the inner V_s sons at site s_{NW}:

$$\begin{cases} p(v_{s_{NW}}^{\rightarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\rightarrow}|X_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\uparrow}, v_{s^-}^{\nearrow}, v_{s^-}^{\rightarrow})), \\ p(v_{s_{NW}}^{\searrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\searrow}|X_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\rightarrow}, v_{s^-}^{\searrow}, v_{s^-}^{\downarrow})), \\ p(v_{s_{NW}}^{\downarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\downarrow}|X_{s^-} = x_{s^-}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\downarrow}, v_{s^-}^{\swarrow}, v_{s^-}^{\leftarrow})). \end{cases} \tag{3.14}$$

The same remark as for Equation 3.10 needs to be made for Equation 3.14: given a father s^- and two sons s, s', $\forall s \in \{NW, NE, SE, SW\}$, the inner V_s sons take the same conditioning as the nearest $X_{s'}$ (neighbor of X_s) on the same resolution. Since inner V_s will always have the same conditioning as a $X_{s'}$ whose father is also at site s^-, we have $s \neq s'$.

Outer I V_s sons take the same conditioning as the nearest $X_{s'}$ but under a *switching* condition concerning the conditioning: the V_{s^-} with the same direction as V_s is switched with X_{s^-}. This gives the transition laws of the outer I V_s sons at site s_{NW}:

$$\begin{cases} p(v_{s_{NW}}^{\leftarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\leftarrow}|X_{s^-} = v_{s^-}^{\leftarrow}, \boldsymbol{V}_{s^-} = (x_{s^-}, v_{s^-}^{\nwarrow}, v_{s^-}^{\uparrow})), \\ p(v_{s_{NW}}^{\nwarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\nwarrow}|X_{s^-} = v_{s^-}^{\nwarrow}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\leftarrow}, x_{s^-}, v_{s^-}^{\uparrow})), \\ p(v_{s_{NW}}^{\uparrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\uparrow}|X_{s^-} = v_{s^-}^{\uparrow}, \boldsymbol{V}_{s^-} = (v_{s^-}^{\leftarrow}, v_{s^-}^{\nwarrow}, x_{s^-})). \end{cases} \tag{3.15}$$

Outer II V_s sons take the same conditioning as the nearest $X_{s'}$, for s and s' two nodes which are sons of s^- (leading again to $s \neq s'$). Moreover we also have a *switching* condition: the only V_{s^-} variable which plays a role in the conditioning of both X_s and $X_{s'}$ is switched with X_{s^-}. This gives the transition laws for the outer II V_s sons at site s_{NW}:

$$\begin{cases} p(v_{s_{NW}}^{\nearrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\nearrow}|X_{s^-} = v_{s^-}^{\uparrow}, \boldsymbol{V}_{s^-} = (x_{s^-}, v_{s^-}^{\nearrow}, v_{s^-}^{\rightarrow})), \\ p(v_{s_{NW}}^{\swarrow}|x_{s^-}, \boldsymbol{v}_{s^-}) = p(v_{s_{NW}}^{\swarrow}|X_{s^-} = v_{s^-}^{\leftarrow}, \boldsymbol{V}_{s^-} = (x_{s^-}, v_{s^-}^{\downarrow}, v_{s^-}^{\swarrow})), \end{cases} \tag{3.16}$$

To illustrate Equations 3.13, 3.14, 3.15 and 3.16 (and the other transition laws that have not been written), Figure 3.3 illustrates a transition between a node s^- and its 4 sons s_{NW}, s_{NE}, s_{SE} and s_{NW} with the spatial context.

Finally for the root node $s \in \mathcal{S}^1$, $p(x_s, \boldsymbol{v}_s)$ is chosen as a discrete distribution on $\Omega \times \Lambda^8$.

Remark: In all the thesis, when dealing with image segmentation, the quadtree version of the STMT model (as well as the HMT model) will be used.

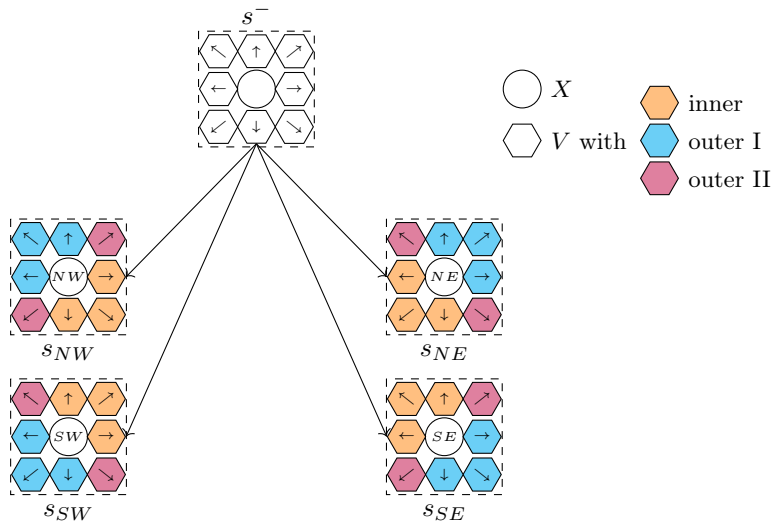


Figure 3.3.: A father node s^- linked to its 4 sons $(s_{NW}, s_{NE}, s_{SE}, s_{SW})$ in the STMT model. Each node is composed of 9 random variables X_s and V_s. The directionality of the V_s is specified as well as their type (inner, outer I or outer II).

3.3. Image segmentation

The models introduced in the previous section will be tested in the context of image segmentation. In this section we present the family of functions used to define the transitions in HMTs and in STMTs, as well as a method to estimate the parameters of these distributions in an unsupervised context.

3.3.1. The Potts-like transition distributions

Definition

Recall the Potts potential function (J. Besag, 1986) defined in Equation 2.16. Let us define conditional probabilities in arborescences that follow this function. Then, in the HMT model, we can form a transition distribution that we call Potts-like distribution:

$$p(x_s|x_{s^-}) \propto \exp(\alpha \delta_{x_s}^{x_{s^-}}), \tag{3.17}$$

where $\alpha \in \mathbb{R}_+$ is a parameter which regulates the similarity between two linked realizations of the hidden process in the HMT model. The higher α the more probable is the fact that the realizations will be equal. A Potts-like distribution for the STMT model is defined by:

$$p(x_s|x_{s^-}, \boldsymbol{v}_{s^-}) \propto \exp\left(\alpha \delta_{x_s}^{x_{s^-}} + \sum_{v_{s^-} \in \boldsymbol{v}_{s^-}} \beta \delta_{x_s}^{v_{s^-}}\right), \tag{3.18}$$

where $\alpha \in \mathbb{R}_+$ and $\beta \in \mathbb{R}_+$ are parameters with the same role as before. In all the following work, the transitions of the HMT and STMT models will use the Potts-like transition distributions. Moreover, the transitions will be considered homogeneous: $\forall s \in \bar{\mathcal{S}}$, they are Potts-like transitions equally parametrized.

Remark: Note that when $\beta = 0$, Equations 3.17 and 3.18 are equal and the HMT and STMT models thus become equivalent.

Linear Least-Square estimator of the Potts-like transition parameter

We now establish an estimator for the α and the β parameters based on the Linear Least-Square (LLS) estimator first proposed in (Derin and H. Elliott, 1987). The method requires the complete data, hidden and observed, which correspond to the pair $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})$ for HMT and the triplet $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{Y} = \boldsymbol{y})$ for STMT. Note that the derivation of the LLS estimator for the granularity coefficient of the GPMF model (Section 2.4.1) is close to the equations that we now establish.

In the case of the α estimation for HMTs, we can write, $\forall s \in \bar{\mathcal{S}}$:

$$\frac{p(x_s, x_{s^-}|\boldsymbol{y})}{p(x_{s^-}|\boldsymbol{y})} = p(x_s|x_{s^-}, \boldsymbol{y}), \tag{3.19}$$

which gives, $\forall s \in \bar{\mathcal{S}}, \forall (x_s, x'_s) \in \Omega^2$:

$$\frac{p(x_s|x_{s^-}, \boldsymbol{y})}{p(x'_s|x_{s^-}, \boldsymbol{y})} = \frac{p(x_s, x_{s^-}|\boldsymbol{y})}{p(x'_s, x_{s^-}|\boldsymbol{y})}. \tag{3.20}$$

From now on, only the sites s such that $s \in \mathcal{S}^L$ will be considered[4]. From the Upward Downward algorithm (Algorithm 3.1), we know that in the present case of the independent noise, we have $\forall s \in \mathcal{S}^L$:

$$p(x_s|x_{s^-}, \boldsymbol{y}) = \frac{p(y_s|x_s)p(x_s|x_{s^-})}{\sum_{x_s} p(y_s|x_s)p(x_s|x_{s^-})}. \tag{3.21}$$

Then Equation 3.20 becomes, $\forall s \in \mathcal{S}^L, \forall (x_s, x'_s) \in \Omega^2$:

$$\frac{p(y_s|x_s)p(x_s|x_{s^-})}{p(y_s|x'_s)p(x'_s|x_{s^-})} = \frac{p(x_s, x_{s^-}|\boldsymbol{y})}{p(x'_s, x_{s^-}|\boldsymbol{y})}. \tag{3.22}$$

Taking the natural logarithm on both sides, plugging the potential from Equation 3.17, using the definition of the independent Gaussian noise and rearranging, we finally get:

$$\begin{aligned}\alpha(\delta_{x_s}^{x_{s^-}} - \delta_{x'_s}^{x_{s^-}}) = \ln \frac{p(x_s, x_{s^-}|\boldsymbol{y})}{p(x'_s, x_{s^-}|\boldsymbol{y})} - \ln \frac{1}{\sqrt{2\pi\sigma_{x_s}^2}} + \frac{(y_s - \mu_{x_s})^2}{2\sigma_{x_s}^2} + \\ \ln \frac{1}{\sqrt{2\pi\sigma_{x'_s}^2}} + \frac{(y_s - \mu_{x'_s})^2}{2\sigma_{x_s}^2}.\end{aligned} \tag{3.23}$$

[4] Equations from sites from other resolutions could be derived and integrated to the LLS however we found out that they do not contribute significantly to the final solution.

The probabilities of the type $p(x_s, x_{s^-}|\boldsymbol{y}), \forall (x_s, x_{s^-}) \in \Omega^2$, are estimated with the frequency estimator. Then the same LLS procedure as described in Section 2.4.1 can be used to find $\hat{\alpha}$.

Then, following strictly the same procedure for the STMT model, starting from, $\forall s \in \bar{\mathcal{S}}$:

$$\frac{p(x_s, x_{s^-}, \boldsymbol{v}_{s^-}|\boldsymbol{y})}{p(x_{s^-}, \boldsymbol{v}_{s^-}|\boldsymbol{y})} = p(x_s|x_{s^-}, \boldsymbol{v}_{s^-}, \boldsymbol{y}), \tag{3.24}$$

we can write, in the case of the independent Gaussian noise, $\forall s \in \bar{\mathcal{S}}, \forall (x_s, x'_s) \in \Omega^2$:

$$\begin{aligned} \alpha(\delta_{x_s}^{x_{s^-}} - \delta_{x'_s}^{x_{s^-}}) + \beta \left(\sum_{v_{s^-} \in \boldsymbol{v}_{s^-}} \left(\delta_{x_s}^{v_{s^-}} - \delta_{x'_s}^{v_{s^-}} \right) \right) = \ln \frac{p(x_s, x_{s^-}, \boldsymbol{v}_{s^-}|\boldsymbol{y})}{p(x'_s, x_{s^-}, \boldsymbol{v}_{s^-}|\boldsymbol{y})} - \\ \ln \frac{1}{\sqrt{2\pi\sigma_{x_s}^2}} + \frac{(y_s - \mu_{x_s})^2}{2\sigma_{x_s}^2} + \ln \frac{1}{\sqrt{2\pi\sigma_{x'_s}^2}} + \frac{(y_s - \mu_{x'_s})^2}{2\sigma_{x_s}^2}. \end{aligned} \tag{3.25}$$

Again, the probabilities of the type $p(x_s, x_{s^-}, \boldsymbol{v}_{s^-}|\boldsymbol{y})$ are estimated with the frequency etimator. The LLS procedure presented in Section 2.4.1 is adapted to handle two unknown variables and to estimate $\hat{\alpha}$ and $\hat{\beta}$.

3.3.2. Iterative Parameter Estimation for Trees

This section describes fully the Iterative Parameter Estimation for Trees (IPET) algorithm which will be used to estimate the parameters in unsupervised problems of image segmentation. This procedure is a deterministic iterative procedure close to the SEM algorithm (Celeux, 1985). Moreover, we describe the IPET method for the STMT model, its adaptation for the HMT model is straightforward.

Without loss of generality, in the following developments we consider a two class segmentation problem, *i.e.*, $\Omega = \{\omega_0, \omega_1\}$, and the case of Gaussian independent noise parametrized by the class of the realization of the underlying hidden variable. This noise model adds four parameters to the modelizations: $\mu_0 \triangleq \mu_{\omega_0}$, $\mu_1 \triangleq \mu_{\omega_1}$, $\sigma_0 \triangleq \sigma_{\omega_0}$, $\sigma_1 \triangleq \sigma_{\omega_1}$. Finally for the STMT model, this leads to six parameters to estimate. Let $\boldsymbol{\theta} \in \Theta$ be the vector of parameters, then:

$$\begin{aligned} \boldsymbol{\theta} &= \{\alpha, \beta, \mu_0, \mu_1, \sigma_0, \sigma_1\}, \\ \text{with } \Theta &= (\mathbb{R}_+)^2 \times (\mathbb{R})^2 \times (\mathbb{R}_+^*)^3. \end{aligned} \tag{3.26}$$

We use the Maximum Likelihood (ML) estimators μ_0 and μ_1, $\forall i \in \{0, 1\}$:

$$\hat{\mu}_i = \frac{1}{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{x_s = i\}}} \sum_{s \in \mathcal{S}} y_s \mathbb{1}_{\{x_s = i\}}. \tag{3.27}$$

The ML estimators are also used for σ_0 and σ_1, $\forall i \in \{0, 1\}$:

$$\hat{\sigma}_i = \left(\frac{1}{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{x_s = i\}}} \sum_{s \in \mathcal{S}} \mathbb{1}_{\{x_s = i\}} (y_s - \hat{\mu}_{x_s})^2 \right)^{\frac{1}{2}}. \tag{3.28}$$

The IPET algorithm for unsupervised parameter estimation in STMTs is then described in Algorithm 3.2.

Remark: The main difference between the SPE algorithm for parameter estimation in Markov Random Fields (Section 2.4.1) and the IPET algorithm for parameter estimation in MTs is that the expectation step is done in an exact manner. It is taken equal to the result of the MPM criterion (Section 1.5.2) readily accessible in MTs with the UD algorithm (Equation 3.4). The IPET algorithm is deterministic.

Algorithm 3.2: Iterative Parameter Estimation for Trees (IPET) for STMTs.

Data: $\boldsymbol{\theta}^0 = \{\alpha^0, \beta^0, \mu_0^0, \mu_1^0, \sigma_0^0, \sigma_1^0\}$, an initial set of parameters,
$\boldsymbol{y}$, the observations.
Result: $\boldsymbol{\theta}^* = \{\alpha^*, \beta^*, \mu_0^*, \mu_1^*, \sigma_0^*, \sigma_1^*\}$, the estimated parameters.
$t \leftarrow 1$
while *convergence is not attained* **do**
1. MPM estimation:
$\hat{x}_s^{MPM,t} = \text{argmax}_{x_s} p(x_s|\boldsymbol{y}, \boldsymbol{\theta}^{t-1})$, $\forall s \in \mathcal{S}$
2. Estimation with the complete data $(\hat{\boldsymbol{x}}^{MPM,t}, \boldsymbol{y})$:
• LLS estimator for α^t and β^t (Eq.3.25).
• ML estimator for μ_0^t and μ_1^t (Eq.3.27).
• ML estimator for σ_0^t and σ_1^t (Eq.3.28).
$\boldsymbol{\theta}^t \leftarrow= \{\alpha^t, \beta^t, \mu_0^t, \mu_1^t, \sigma_0^t, \sigma_1^t\}$
$t \leftarrow t + 1$
end

3.3.3. Experiments and Results

Synthetic data experiment

In this first experiment, we assess experimentally the strict generalization of STMTs over HMTs-IN in the case of the Potts-like potential. We sample realizations of the STMT model and segment the observed images with both models. The segmentation criterion is the MPM and the parameters used for the segmentation are the parameters used for the simulation, the true parameters. We are in the context of supervised segmentation. Again the hypothesis of the independent Gaussian noise parametrized by the underlying class (Section 3.3.2) is done for both models. Let us denote the segmentation error rate of the HMT-IN model, resp. STMT model, err_{HMT-IN}, resp. err_{STMT}.

In the first experiment, α and β are the varying parameters in the STMT samples. We plot the quantity $\text{err}_{HMT-IN} - \text{err}_{STMT}$ as a function of α and β in Figure 3.4. We first notice that the values taken by the surface are always positive which means that the STMT segmentations are always better or equal to the HMT-IN segmentations (the maximal value is a 4-point improvement in

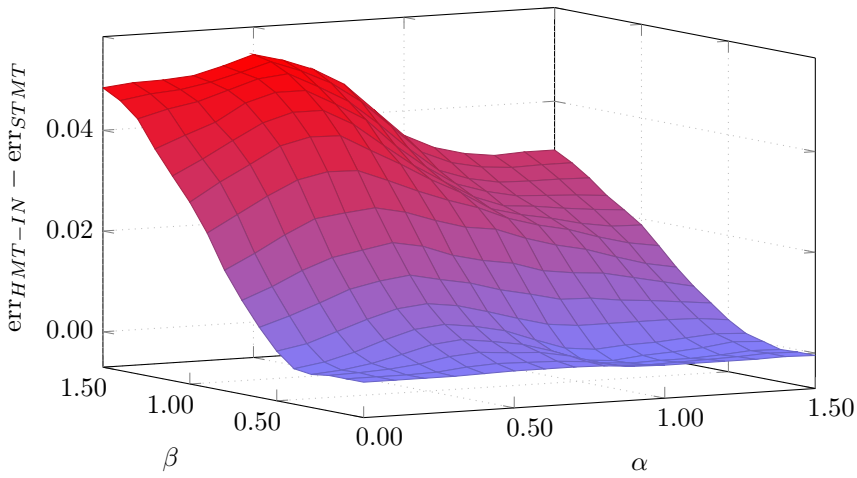


Figure 3.4.: $\mathrm{err}_{HMT-IN} - \mathrm{err}_{STMT}$ as a function of α and β. The other parameters were fixed to $\mu_0 = 0, \mu_1 = 0.6$ and $\sigma_0 = \sigma_1 = 0.5$. We note a segmentation gain (in terms of error rates) raising up to 4 points in favour of the STMT model.

favor of STMT). Note the strict equivalence between the models when $\beta = 0$, this fact already mentioned earlier is now confirmed empirically. Furthermore we can see that as β increases, *i.e.*, the importance of the realization of the random variables which are neighbors to the father increases, the advantage of the STMT model increases. This was to be expected since the HMT-IN model cannot take into account such spatial and local correlations. However, a bigger α, *i.e.*, the dependence on the realization of the father random variable increases, leads to a smaller advantage for the STMT model. This is due to the fact that, with a bigger α, the importance of the neighbors is lowered in the Potts-like transitions. In such cases the HMT-IN model is more relevant.

In the second experiment, let us fix $\sigma = \sigma_0 = \sigma_1$ and $\Delta\mu = |\mu_0 - \mu_1|$. Then σ and $\Delta\mu$ are the varying parameters in the STMT samples. We plot in Figure 3.5 the quantity $\mathrm{err}_{HMT-IN} - \mathrm{err}_{STMT}$ as a function of $\Delta\mu$ and σ. We can see that for all the noise levels, the error rates are in favor of the STMT model. Moreover we note that the smaller the noise levels, the more similar the performances of both models.

Semi-real image experiment

In this section, we work in the context of unsupervised image segmentation on semi-real images from the "1070-Binary Shape Database"[5] already used in Chapter 2. We aim at highlighting the apparent capacity for the STMT model to take into account the spatial context. The observations $\boldsymbol{y}$ in this section

[5] https://vision.lems.brown.edu/content/available-software-and-databases

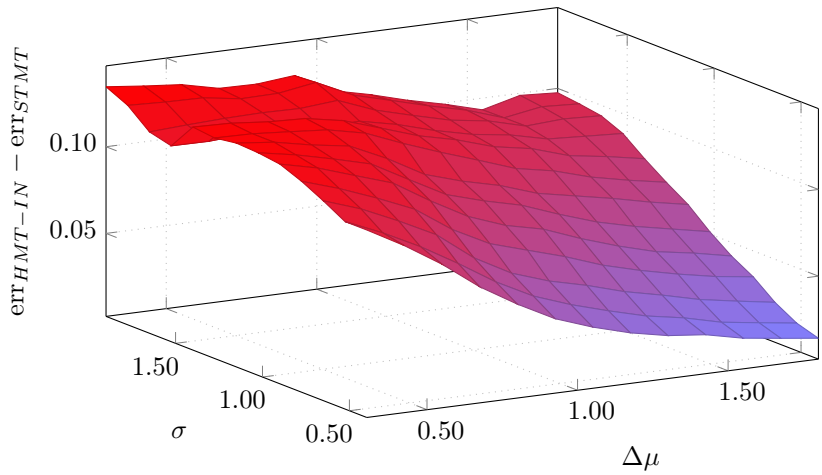


Figure 3.5.: $\mathrm{err}_{HMT-IN} - \mathrm{err}_{STMT}$ as a function of $\Delta\mu$ and σ. We note a segmentation gain (in terms of error rates) raising up to 14 points.

are then the binary images from the data corrupted with an additive Gaussian noise. The noise level will be varying during our experiment.

The parameters of the HMT and STMT models are computed unsupervisedly with the IPET algorithm (Algorithm 3.2). The tree models are then compared to the results of a Hidden Markov Field with Independent Noise (HMF-IN). This probabilistic model is indeed a reference to model local spatial correlations and the motivation behind the STMT model. In the binary classification task we consider ($\Omega = \{\omega_0, \omega_1\}$), the HMF-IN model is defined with a prior factor taken as a Potts factor (J. Besag, 1986) with bias:

$$\tilde{p}(x_s|\boldsymbol{x}_{\mathcal{N}_s^X}) = \exp\left(-\delta_{x_s}^{\omega_0}\alpha - 2\sum_{s'\in\mathcal{N}_s^X} -\delta_{x_s}^{x_{s'}}\beta\right), \tag{3.29}$$

where $\alpha \in \mathbb{R}$ is the bias parameter[6] and $\beta \in \mathbb{R}_+^*$ is the granularity parameter, and by the conditional likelihood factor:

$$\tilde{p}(y_s|x_s) = \exp\left(-\ln(\sqrt{2\pi}\sigma) - \frac{\bar{y}_s^2}{2\sigma^2}\right). \tag{3.30}$$

The description and parameter estimation for a HMF-IN model without bias has been developed in Section 2.3. The extension to estimate the bias parameter α is straightforward by considering a two-unknown-variable LLS as already mentioned in Section 3.3.1.

[6] Strictly speaking, there should be a bias parameter for each of the class. However, in this binary classification context, the proposed parametrization can be seen as forcing the bias parameter for class ω_1 to be 0. α is then the bias parameter for class ω_0 and it is estimated under this constraint.

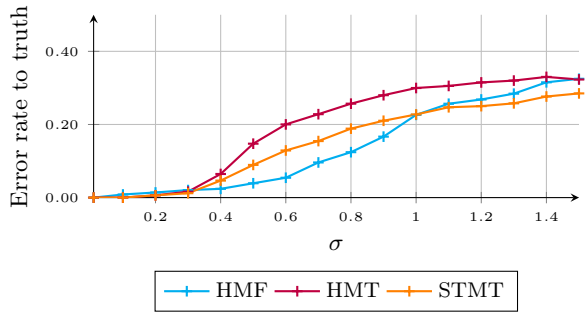


Figure 3.6.: Error rate in unsupervised segmentation

Figure 3.7.: Unsupervised segmentation of semi-real images with HMTs, STMTs and HMFs. The figure depicts the error rate of the three models with a varying additive Gaussian noise level. It appears that the STMT model always performs equally or better than the HMT model, offering up to a 5 point improvement in the error rates with respect to the HMT model. The HMF model is the best performing model offering up to a 10 point gain over the STMT model.

Figure 3.7 illustrates, for a varying noise level, the error rates in the unsupervised segmentation task for each model. This experiment clearly attests the advantage of the STMT model over the HMT model in unsupervised segmentation, the STMT model always performs equivalently or better than the HMT model. This result confirms the conclusions made on simulated data in Section 3.3.3. The HMFs remain the best performing model for $\sigma \leq 1.0$. However, it then performs worse than STMTs at high noise levels, this might be due to the failures of the stochastic algorithms that we use for the models. Indeed, HMFs rely on indirect stochastic inference and parameter estimation which might fail at such noise levels. It is then notable that the STMT improvement is obtained while keeping a direct and analytical inference process. Figure 3.8 reports selected graphical examples of the experiment described above.

Experiments of atherosclerotic calcification segmentation are given in Chapter 5.

3.4. Spatial Triplet Markov Trees for Auxiliary Variational Inference in Spatial Bayes Networks

This section now offers a more theoretical study of the distribution of the original hidden process $\boldsymbol{X}$ within the STMT model. In order to understand the much more complex and richer correlations that the STMT model introduces,

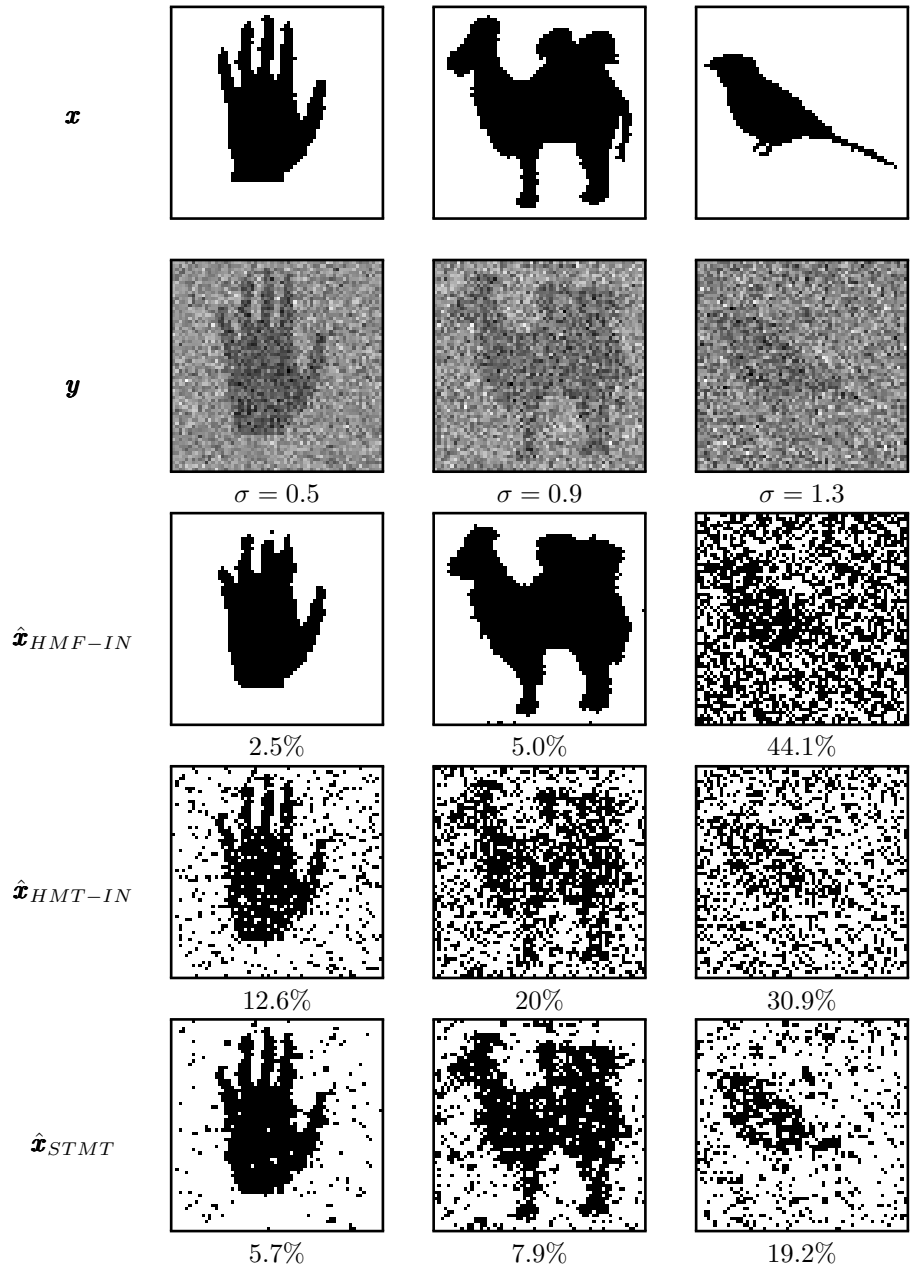


Figure 3.8.: Unsupervised segmentation of semi real images with the HMF, HMT and STMT models. Ground truths are on the first row. The percentages indicate the segmentation error rate.

we first define a new model of Bayesian network (or Directed Acyclic Graph (DAG)) which differs from the previous MT models. We call this new model Spatial Bayes Network (SBN) and we will study its relationship to the STMT model.

More precisely, as will see in the next section about the model definition, SBNs exhibit directly the spatial correlations that we want to model (but cannot directly model without breaking the Markovian assumption on the process) in the STMT construction. This comes at the price of an inference process that cannot be done through direct computations. However, since the direct computations are available within the STMT model, the pair of models SBN/STMT can be used in the context of Variational Inference (VI) as we will see in Section 3.4.2.

Remark: The following development focuses on a study of the hidden and auxiliary processes, therefore, there will be no mention of a visible process, and the latter will not be drawn in the graphical representations. This can also be interpreted as working in graphical models where all the variables are visible.

3.4.1. Spatial Bayes Networks

The SBN model is a DAG based on the MT model but SBNs contain semi cycles. SBNs are not arborescences anymore and the computation for the inference step must be approximate. The additional edges have the purpose to better capture local correlations between random variables.

The dyadic model

In order to distinguish between the two sons of a father node s^- in a dyadic tree: let s_L and s_R be respectively the *left* and *right* son of s^-. We also define $s^{\leftarrow}$ (resp. $s^{\rightarrow}$) as the left (resp. right) neighboring node of s. Let $v\colon \bar{\mathcal{S}} \to \bar{\mathcal{S}}$ be a mapping from a node to a neighboring node of its father, such that:

$$v\colon s \mapsto \begin{cases} (s^-)^{\leftarrow} \text{ if } s \text{ is a left node,} \\ (s^-)^{\rightarrow} \text{ if } s \text{ is a right node.} \end{cases} \tag{3.31}$$

The SBN model (illustrated in Figure 3.9) has the following distribution:

$$p(\boldsymbol{x}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}, x_{v(s)}). \tag{3.32}$$

The quadtree model

The development above, dedicated to dyadic trees, can be straightforwardly extended to quadtrees, to form the quadtree SBN. There are 4 types of nodes (except from the root), which we call *North West* (NW), *North East* (NE),

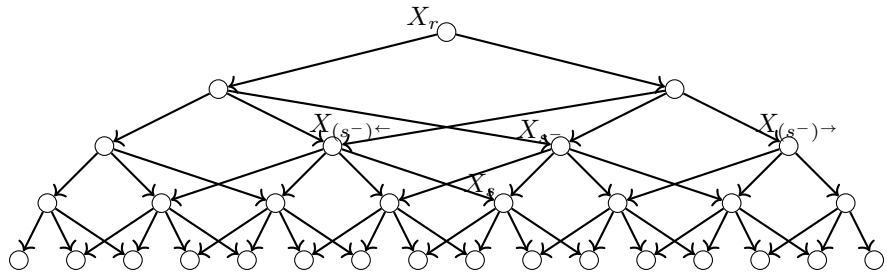


Figure 3.9.: Graphical model corresponding to a dyadic SBN. As an illustration, a node s, its father s^-, the left neighbor of its father $(s^-)^\leftarrow$, the right neighbor of its father $(s^-)^\rightarrow$ and the root node r have been annotated.

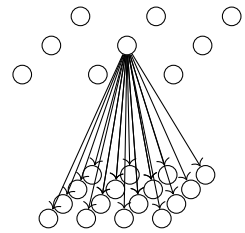


Figure 3.10.: Local illustrations of the correlations within the quadtree SBN model: all the sons of a father vertice.

South East (SE) and *South West* (SW). Now the v function is redefined as a mapping from $\bar{\mathcal{S}}$ to $\bar{\mathcal{S}}^3$ with:

$$v\colon s \mapsto \begin{cases} ((s^-)^\leftarrow, (s^-)^\nwarrow, (s^-)^\uparrow) \text{ if } (s^-) \text{ NW node,} \\ ((s^-)^\uparrow, (s^-)^\nearrow, (s^-)^\rightarrow) \text{ if } (s^-) \text{ NE node,} \\ ((s^-)^\rightarrow, (s^-)^\searrow, (s^-)^\downarrow) \text{ if } (s^-) \text{ SE node,} \\ ((s^-)^\downarrow, (s^-)^\swarrow, (s^-)^\leftarrow) \text{ if } (s^-) \text{ SW node.} \end{cases}$$

The joint distribution $p(\boldsymbol{x})$ of the quadtree SBN uses this newly defined v function. It is then written as:

$$p(\boldsymbol{x}) = p(x_r) \prod_{s \in \bar{\mathcal{S}}} p(x_s | x_{s^-}, \boldsymbol{x}_{v(s)}). \tag{3.33}$$

For clarity, we only illustrate locally the quadtree SBN model. Figure 3.10 depicts the relations between a father vertice and all its sons.

Exploring the model similarities by sampling

We propose to explore the capacity of the STMT, SBN and MT models to take into account the spatial context by *clamped* sampling in these models. Recall

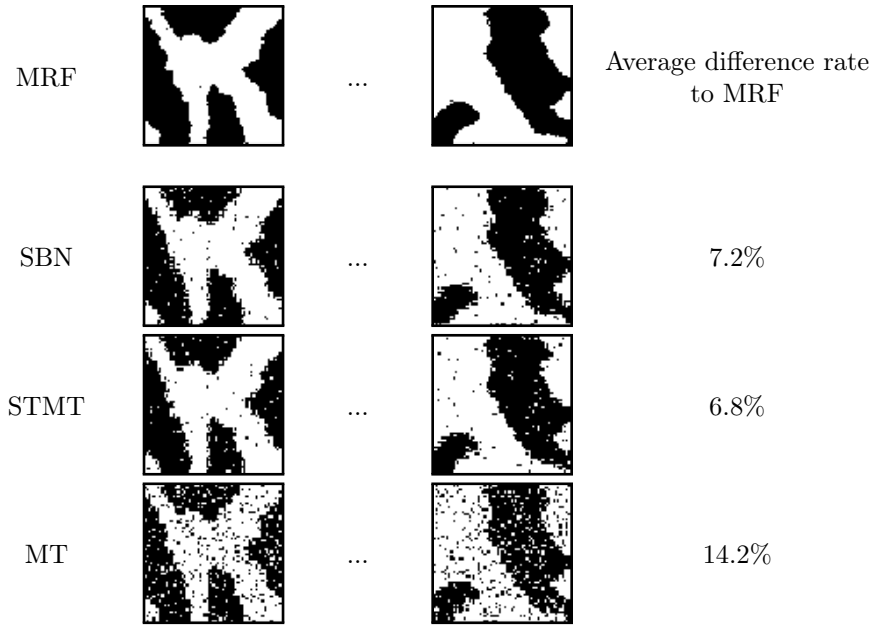


Figure 3.11.: Clamped samplings of the original $\boldsymbol{X}$ process (last layer only) from the SBN, STMT and MT models. The first row represents the MRF reference simulations. The other rows represent the clamped samplings. The average difference rate between the realizations of each model and the MRF reference realization is computed over 100 simulations (from which only two are graphically depicted).

that in all this section, we do not consider any visible process ($\boldsymbol{Y}$). Clamped sampling in these layered models means that we fix a layer to a given realization and sample the next generation(s) by ancestral sampling. We fix a realization to be that of a MRF with a Potts potential since it is our reference for modeling local spatial interactions.

Figure 3.11 illustrates the experiment and provides an averaged difference rate. From this score, and also the two visual examples of the figure, we can say that the SBN and STMT models seem to better capture the spatial context than the MT model whose realizations are farther away from the MRF reference. Indeed, the difference rate between the MT realizations and MRF reference is more than twice the difference rate between the STMT realizations and the MRF reference.

This experiment provides a first insight on the similarity between STMT and SBN. The next section and the rest of the chapter study this similarity more in depth in the context of Variational Inference.

3.4.2. Variational inference

Variational Inference (VI) (Lauritzen and Spiegelhalter, 1988) (Michael I Jordan et al., 1999) (Murphy, 2012) (Blei et al., 2017) (C. Zhang et al., 2018) is an approach to perform inference in probabilistic models where direct computations are not feasible. In the general context the goal is to find a *variational distribution* $q(\boldsymbol{x})$ which approximates well a posterior $p(\boldsymbol{x}|\boldsymbol{y})$ in which computations for inference need to be approximated. To do so, VI casts an optimization problem which aims at minimizing the Kullback Leibler (KL) divergence between the two distributions; it is denoted $\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y}))$. By definition:

$$\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log q(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{y})]. \tag{3.34}$$

The variables of this optimization problem are the factors of the q distribution. The structure of q is chosen so that inference in q is easier than in p. Most of the time, q is simple enough for inference tasks in q to be directly computable. Therefore, at the end of the optimization problem, when q is fitted to p in the sense of the KL divergence, the complex inference in p will be approximated by an easier inference in q. It can be shown (Blei et al., 2017) that minimizing Equation 3.34 is equivalent to maximizing the Evidential Lower BOund (ELBO) quantity defined as:

$$\begin{aligned} ELBO(q) &= \log p(\boldsymbol{y}) - \mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})), \\ &= \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{y})] - \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log q(\boldsymbol{x})]. \end{aligned} \tag{3.35}$$

Remark: As stated above, the rest of this section will not consider any visible process $\boldsymbol{Y}$, hence the conditioning on $\boldsymbol{y}$ in the equations, such as Equation 3.34, will drop. Note however that all the results could be readily extended to integrate a visible process. Then, we have that:

$$\begin{aligned} ELBO(q) &= \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log p(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log q(\boldsymbol{x})], \\ &= -\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})). \end{aligned} \tag{3.36}$$

Therefore, in what follows, we will refer to the optimization problem as maximizing the opposite of the KL divergence:

$$-\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log p(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log q(\boldsymbol{x})]. \tag{3.37}$$

Remark: The rest of this section will consider the MT and SBN models in their dyadic version. All the results can also be written for the quadtree versions of the models.

In our case, $p(\boldsymbol{x})$ is the joint distribution of a dyadic SBN given in Equation 3.32. In the context of structured VI (introduced in the next paragraph), we will reparametrize and work with another representation of Equation 3.32. Indeed, we need to use a concise notation embedding the notion of clusters of

variables for each of the terms of the product. The notations are adapted from (C. Bishop and Winn, 2003). We can write Equation 3.32 as:

$$p(\boldsymbol{x}) = \prod_{d_s} p_{d_s} \tag{3.38}$$

where d_s represents the cluster of variables $(x_s, x_{s^-}, x_{v(s^-)})$ (note that these clusters overlap) and p_{d_s} is a shortcut for $p(d_s)$. In the following we develop the necessary material to approximate Equation 3.32 by a variational distribution:

1. with totally independent random variables (Section 3.4.3).
2. with a MT structure (Section 3.4.4).
3. with a STMT structure (Section 3.4.5).

The first option is the most popular and is called Mean Field VI (Michael I Jordan et al., 1999). The two other options give the variational distribution a structure, it is called structured VI (Wiegerinck, 2000) (C. Bishop and Winn, 2003) (Z. Ghahramani and M. I. Jordan, 1996) (Olariu et al., 2009). It consists in using a structured variational distribution, but which is still simpler than the distribution p, in order to better approximate the correlations in the original intricate distribution. Structured VI can lead to dramatic increases in performances thanks to the improved modeling of the correlations. In the structured VI scenarii of this study, we take advantage of the MT and STMT structures and of the direct inference computations.

Algorithmically, the steps of the VI procedure for a variational distribution q can be summarized:

1. Initialize all the factors of q.
2. Update each factor of q with the expression that minimizes the KL divergence.
3. Check convergence and repeat step 2 if needed.

The convergence can be assessed, *e.g.*, by monitoring the values of the cost function (the KL divergence) or monitoring stationarity in the estimated variational parameters.

Remark: Once the terms of the variational distribution are learnt, a final inference task still needs to be done. For example, one could compute the MPM of the variational distribution, which would be used to approximate the MPM of the original intractable distribution.

3.4.3. Mean Field Variational Inference in SBNs

Let q be the variational distribution. If q is subject to the Mean Field (MF) assumption, all the random variables are independent:

$$q(\boldsymbol{x}) = \prod_i q_i(x_i). \tag{3.39}$$

The quantity to maximize (Equation 3.37) becomes in this case:

$$\begin{aligned} -\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})) &= \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}[\log p(\boldsymbol{x}) - \log q(\boldsymbol{x})], \\ &= \sum_{\boldsymbol{x}} \prod_i q_i(x_i) \left[\log p(\boldsymbol{x}) - \sum_k \log q_k(x_k) \right], \end{aligned} \tag{3.40}$$

where p is the joint dyadic SBN distribution 3.32. MF is the most popular form of VI. The derivation of the update equations to solve the MF optimization problem is done in Appendix D.1.

3.4.4. Markov Tree Variational Inference in SBNs

Let r be the variational distribution. r is here defined with the structure of a MT, it then follows a distribution of the form of Equation 1.23. Using a MT structure as a posterior to approximate the complex posterior p is a compromise between additional structure that might better reflect the correlations in p and tractability. Indeed, once the posterior transitions are learnt through the VI optimization, the posterior marginals are easily computable with the UD algorithm.

We then reparametrize the MT expression in terms of clusters:

$$r(\boldsymbol{x}) = \prod_{c_s} r_{c_s}, \tag{3.41}$$

where c_s represents the cluster of variables (x_s, x_{s^-}). Note that the clusters overlap.

The quantity to maximize (Equation 3.37) becomes in this section:

$$\begin{aligned} -\mathbb{KL}(r(\boldsymbol{x})||p(\boldsymbol{x})) &= \mathbb{E}_{\boldsymbol{x}\sim r(\boldsymbol{x})}[\log p(\boldsymbol{x}) - \log r(\boldsymbol{x})], \\ &= \sum_{\boldsymbol{x}} \prod_{c_s} r_{c_s} \left(\log p(\boldsymbol{x}) - \sum_{c_k} \log r_{c_k} \right). \end{aligned} \tag{3.42}$$

The derivation of the update equations for this optimization problem is given in Appendix D.2.

3.4.5. Auxiliary variable Variational Inference

We now study VI with auxiliary random variables whose first occurence in the literature is in (Agakov and Barber, 2004). Let $\boldsymbol{V}$ be the auxiliary process, both the target distribution, p, and the variational distribution, which we call here t, will be augmented with variables from $\boldsymbol{V}$[7].

Remark: As stated in (Agakov and Barber, 2004), in this thesis, we augment p with the condition that $p(\boldsymbol{x}, \boldsymbol{v}) = p(\boldsymbol{x})p(\boldsymbol{v}|\boldsymbol{x})$. This ensures an easy

[7] Note that if p is not augmented with auxiliary random variables, certain terms of t involving the distributions of auxiliary random variables cannot be updated.

marginalization of the $\boldsymbol{V}$ process in order to compare the distributions of the $\boldsymbol{X}$ process. In Section 3.4.7 we show how to integrate auxiliary random variables in p, following the previous joint distribution decomposition when p is a SBN.

Then, considering again temporarily observed random variables for generality, the new KL divergence with auxiliary random variables to minimize is:

$$\begin{aligned}\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v}|\boldsymbol{y})) &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{x},\boldsymbol{v})] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x},\boldsymbol{v}|\boldsymbol{y})],\\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{x},\boldsymbol{v})] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x}|\boldsymbol{y})] -\\ &\quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{v}|\boldsymbol{x},\boldsymbol{y})].\end{aligned} \tag{3.43}$$

Following (Kingma, 2017), we give the relation with the previous objective (without auxiliary variables), *i.e.*, the KL divergence of Equation 3.34:

$$\begin{aligned}\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v}|\boldsymbol{y})) &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{x},\boldsymbol{v})] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x}|\boldsymbol{y})] -\\ &\quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{v}|\boldsymbol{x},\boldsymbol{y})]\\ &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{x})] + \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{v}|\boldsymbol{x})] -\\ &\quad \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x}|\boldsymbol{y})] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{v}|\boldsymbol{x},\boldsymbol{y})]\\ &= \mathbb{KL}(t(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})) + \underbrace{\mathbb{KL}(t(\boldsymbol{v}|\boldsymbol{x})||p(\boldsymbol{v}|\boldsymbol{x},\boldsymbol{y}))}_{\geq 0 \text{ (non-negativity)}},\\ &\geq \mathbb{KL}(t(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})).\end{aligned} \tag{3.44}$$

This allows us to draw a conclusion similar to (Kingma, 2017): the use of auxiliary random variables leads to the minimization of an objective function that is *worse* as concerns the original hidden variables of interest. However, the use of auxiliary variables offers much more flexibility in the modelization and enables the practitioner to add rich correlations between the variables of interest. The use of a well structured distribution also reduces local optima, which often helps since VI leads, in general, to a non convex optimization problem (C. Zhang et al., 2018). Those advantages outweight, in practice, the fact that we are optimizing an objective function with a worse lower bound.

Remark: Note that, in Equation 3.44, we also used the factorization $t(\boldsymbol{x},\boldsymbol{v}) = t(\boldsymbol{x})t(\boldsymbol{v}|\boldsymbol{x})$. However, in the case of a STMT variational distribution t, an explicit expression $t(\boldsymbol{x})$ as not been established yet and will be part of further research on the topic. Therefore Equation 3.44 constitutes here a theoretical insight but cannot be used directly.

Remark: The conditioning on $\boldsymbol{y}$ that appears in Equations 3.45 and 3.44 will disappear in the following development. Then, for reasons seen in the

previous section, we will refer to the optimization problem as maximizing the opposite of the KL divergence with auxiliary random variables:

$$-\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v})) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x},\boldsymbol{v})] - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log t(\boldsymbol{x},\boldsymbol{v})]. \tag{3.45}$$

3.4.6. STMT auxiliary variable Variational Inference in SBNs

In this section we propose a STMT distribution t as the variational distribution for auxiliary variable VI in SBNs. We do not include the observed process in the modelization. Then we have:

$$\begin{aligned} t(\boldsymbol{x},\boldsymbol{v}) &= \prod_{s\in\mathcal{S}} t(x_s,\boldsymbol{v}_s|x_{s^-},\boldsymbol{v}_{s^-}), \\ &= \prod_{s\in\mathcal{S}} t(x_s|x_{s^-},\boldsymbol{v}^1_{s^-})t(v_s^{\leftarrow}|x_s,\boldsymbol{v}^2_{s^-})t(v_s^{\rightarrow}|x_s,\boldsymbol{v}^3_{s^-}), \end{aligned} \tag{3.46}$$

which corresponds to Equation 3.7 without the observed process. $\boldsymbol{V}^1_{s^-}$, $\boldsymbol{V}^2_{s^-}$ and $\boldsymbol{V}^3_{s^-}$ are different subsets of $\boldsymbol{V}_{s^-}$. Recall that at the end of the VI procedure, exact marginal computation can be done, again with a generalized UD algorithm.

For the cluster parametrization, we have:

$$t(\boldsymbol{x},\boldsymbol{v}) = \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s}, \tag{3.47}$$

with $c_s = (x_s, x_{s^-}, \boldsymbol{v}^1_{s^-})$, $c'_s = (v_s^{\leftarrow}, x_{s^-}, \boldsymbol{v}^2_{s^-})$ and $c''_s = (v_s^{\rightarrow}, x_{s^-}, \boldsymbol{v}^3_{s^-})$. As previously mentioned, the clusters overlap.

Remark: We separated the auxiliary process, at each site, using their *Left* or *Right* attribute. However this choice is arbitrary and it is only done here for clarity. It can be omitted without modifying the optimization problem.

We want to maximize Equation 3.45, which becomes, with p the distribution of a SBN augmented with auxiliary random variables[8]:

$$\begin{aligned} -\mathbb{KL}(t(\boldsymbol{x},\boldsymbol{v})||p(\boldsymbol{x},\boldsymbol{v})) &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{v})\sim t(\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x},\boldsymbol{v}) - \log t(\boldsymbol{x},\boldsymbol{v})], \\ &= \sum_{\boldsymbol{x},\boldsymbol{v}} \prod_{c_s} t_{c_s} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \Bigg(\log p(\boldsymbol{x}) - \\ &\qquad \Bigg(\sum_{c_k} \log t_{c_k} + \sum_{c'_k} \log t_{c'_k} + \sum_{c''_k} \log t_{c''_k} \Bigg) \Bigg). \end{aligned} \tag{3.48}$$

The derivations of the update equations are given in Appendix D.3.

[8] In Section 3.4.7, we show how p can integrate auxiliary random variables in the SBN case.

3.4.7. Experiments & Results

Small SBN example

We now consider variational inference on the small SBN network given in Figure 3.12a. Similar experiments have been conducted in the same context, to evaluate a VI approximation, for example in (Lauritzen and Spiegelhalter, 1988). Due to its small size the SBN of Figure 3.12a represents a slightly modified probability distribution from that the SBN used up to now. Indeed, we needed to treat in a specific fashion the root node to induce SBN-like correlations on a network with three layers only. It follows that p has the following distribution:

$$\begin{aligned} p(a, a^{\leftarrow}, a^{\rightarrow}, b, c, d, e, f, g) =& p(a)p(a^{\leftarrow})p(a^{\rightarrow})p(b|a, a^{\leftarrow})p(c|a, a^{\rightarrow}) \\ & p(d|b)p(e|b, c)p(f|c, b)p(G|c). \end{aligned} \tag{3.49}$$

We are interested in retrieving the marginals in the SBN using VI. We successively consider three VI techniques: MF VI (Figure 3.13a), MT VI (Figure 3.13b) and STMT VI (with auxiliary nodes) (Figure 3.13c).

The developments of the previous sections are used to conduct VI with MT and STMT. Note that for the STMT VI, we also need to provide the SBN with auxiliary nodes (Agakov and Barber, 2004). We need to keep the property that $p(\boldsymbol{x}, \boldsymbol{v}) = p(\boldsymbol{x})p(\boldsymbol{v}|\boldsymbol{x})$, where $\boldsymbol{x} = \{a, a^{\leftarrow}, a^{\rightarrow}, b, c, d, e, f, g\}$ and $\boldsymbol{v} = \{b^{\leftarrow}, b^{\rightarrow}, c^{\leftarrow}, c^{\rightarrow}, d^{\leftarrow}, d^{\rightarrow}, e^{\leftarrow}, e^{\rightarrow}, f^{\leftarrow}, f^{\rightarrow}, g^{\leftarrow}, g^{\rightarrow}\}$ in order to ensure that $p(\boldsymbol{x})$ (Equation 3.49) is the same between the three VI procedures. In STMT VI, p is then defined so that the auxiliary random variables have the same structure of dependencies in p as in t, the STMT variational distribution. We then have:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{v}) =& p(\boldsymbol{x})p(b^{\leftarrow}|b, a^{\rightarrow})p(b^{\rightarrow}|a, a^{\rightarrow})p(c^{\leftarrow}|a, a^{\leftarrow})p(c^{\rightarrow}|a, a^{\leftarrow}) \\ & p(d^{\leftarrow}|c^{\rightarrow}, b^{\leftarrow})p(d^{\rightarrow}|b^{\leftarrow}, c^{\rightarrow})p(e^{\leftarrow}|b^{\rightarrow}, c^{\leftarrow})p(e^{\rightarrow}|c^{\leftarrow}, b^{\rightarrow}) \\ & p(g^{\leftarrow}|c^{\rightarrow}, b^{\leftarrow})p(g^{\rightarrow}|b^{\leftarrow}, c^{\rightarrow})p(f^{\rightarrow}|b^{\rightarrow}, c^{\leftarrow})p(f^{\leftarrow}|c^{\leftarrow}, b^{\rightarrow}), \end{aligned} \tag{3.50}$$

with

$$\begin{cases} p(b|a, a^{\leftarrow}) & = p(b^{\leftarrow}|a, a^{\leftarrow}) = p(c^{\leftarrow}|a, a^{\leftarrow}), \\ p(c|a, a^{\rightarrow}) & = p(b^{\rightarrow}|a, a^{\rightarrow}) = p(c^{\rightarrow}|a, a^{\rightarrow}), \\ p(d|b) & = p(d^{\leftarrow}|b^{\leftarrow}) = p(e^{\leftarrow}|b^{\leftarrow}), \\ p(e|b, c) & = p(d^{\rightarrow}|b^{\leftarrow}, b^{\rightarrow}) = p(f^{\leftarrow}|c^{\leftarrow}, b^{\rightarrow}), \\ p(f|c, b) & = p(g^{\leftarrow}|c^{\rightarrow}, c^{\leftarrow}) = p(e^{\rightarrow}|b^{\rightarrow}, c^{\leftarrow}), \\ p(g|c) & = p(f^{\rightarrow}|c^{\rightarrow}) = p(g^{\rightarrow}|c^{\rightarrow}). \end{cases} \tag{3.51}$$

The model p with auxiliary nodes is described in Figure 3.12b.

The random variables are chosen with values in $\{0, 1\}$. Our goal is to estimate the *true* marginals $p(x = 1), \forall x \in \boldsymbol{x}$ of the SBN of Figure 3.12a. In this synthetic example those true marginals are given as well as the transitions in p (we do not cover the parameter estimation problem). Moreover, the transitions of Equation 3.49 which also totally define Equation 3.50 are taken randomly.

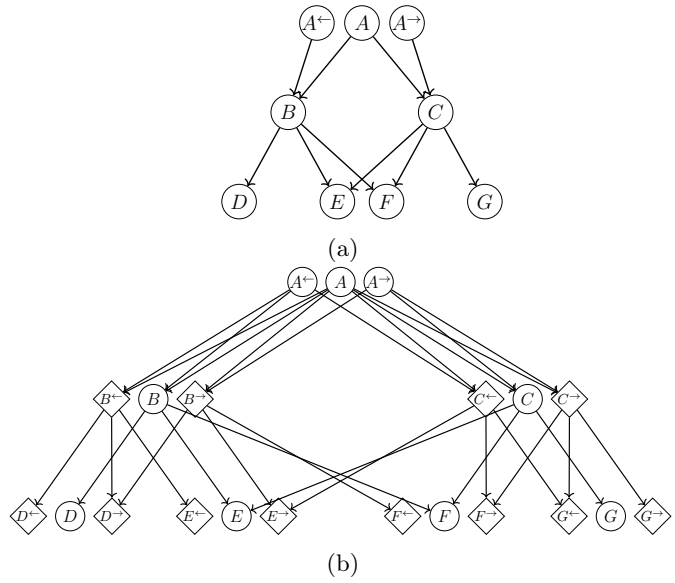


Figure 3.12.: Target distributions in the VI procedures: (a) SBN, (b) Adapted SBN with auxiliary nodes.

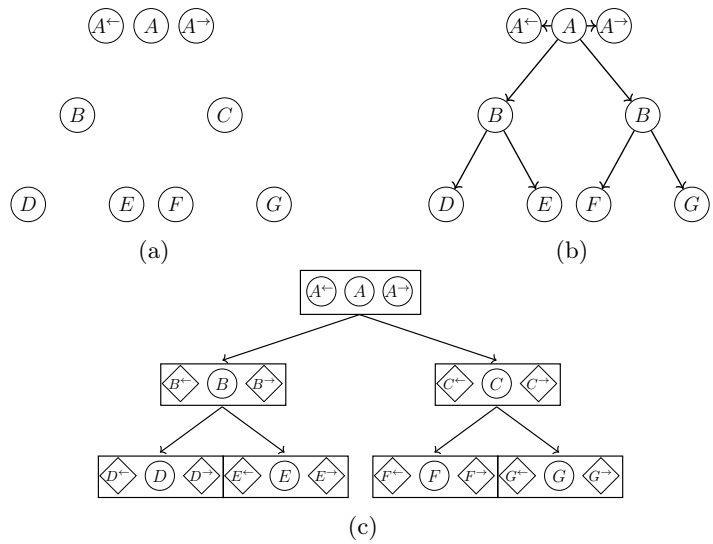


Figure 3.13.: Variational distributions for the VI procedure: (a) MF VI, (b) MT VI and (c) STMT VI.

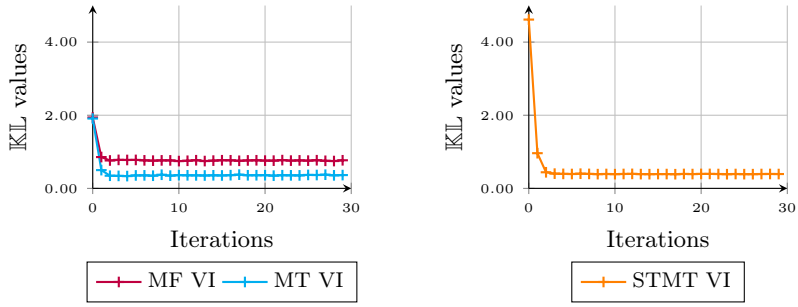


Figure 3.14.: Values of the cost function (KL divergence values) of the minimization problem for the three VI procedures. Note that the STMT VI cost function integrates auxiliary variables and is not comparable with the others, hence it is plotted in another graph.

Results

We define a marginal *error* for a variational distribution q and a random variable x by $e_q(x) = p(x = 1) - q(x = 1)$. These errors are computed and stored over 1000 different SBNs p whose transitions are randomly chosen. Figure 3.14 depicts the values of the KL divergence of the three VI procedures. We see a rapid convergence which is related to the small size of the considered SBN. We then choose to stop the VI process after 30 iterations. The value for MF VI and MT VI are comparable and are shown on the same graph. The minimization is better in the case of MT VI, the structured VI; this is reflected in Figure 3.15 which illustrates the goodness of the estimated marginals.

The boxplots of Figure 3.15 illustrates the interest of the STMT structure to approximate the marginals of SBN: this VI procedure exhibits in all cases the smallest error with respect to the true marginals. In absolute value, this error is remarkably very small. We can also conclude that progressively adding structure increases the quality of the approximation since MF VI performs worse than MT VI which performs in turn worse than STMT VI. Note that the left/right symmetry of the SBN can be found in the behavior of the error rates: D is similar to G, E is similar to F, and so on.

3.5. Conclusion

This chapter was dedicated to the development, study and application of a new probabilistic model, the STMT model. This model combines the benefits of the triplet structure and auxiliary random variables modeling to introduce rich correlations between random variables while preserving exact inference computations. More precisely, we have seen that the correlations modeled are close to the spatial correlations of MRFs, making the new model particularly suited for image processing. In particular, the STMT model outperforms the classical HMT model.

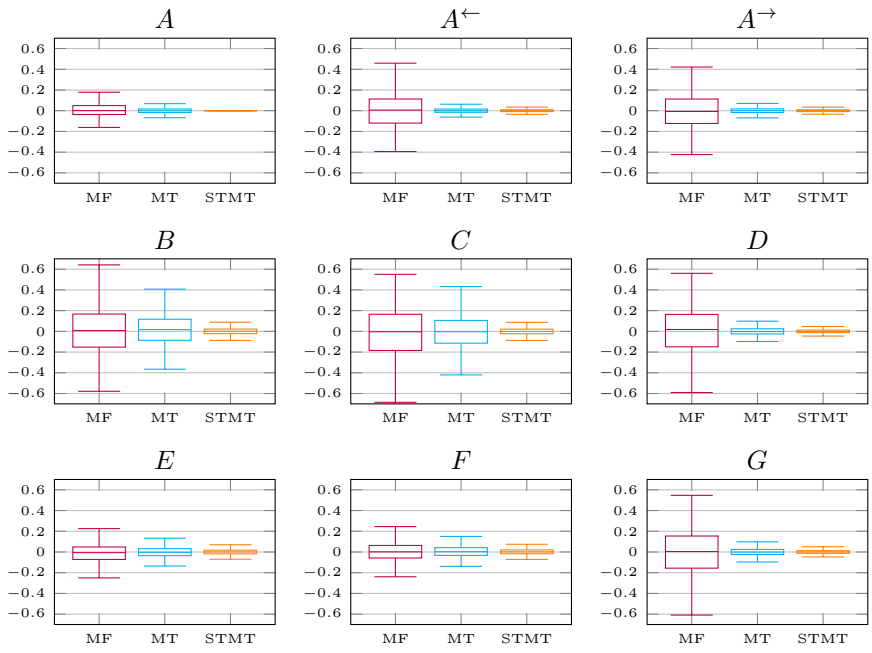


Figure 3.15.: Boxplots of the dispersion of the error around the true marginal, for each vertice, for the three VI procedures. We show the results over 1000 different experiments (different p transitions randomly chosen).

In the second part of this chapter, the developement of the new SBN model and the use of the STMT/SBN pair of models in the context of VI offers an even better understanding of the correlations among the STMT model. The correlations of the $\boldsymbol{X}$ process in SBNs are experimentally found to be very close to the correlations of $\boldsymbol{X}$ in STMTs. Hence the richness of the STMT model.

Chapter 4.

Image segmentation with Deep Learning and Conditional Random Fields

Contents

4.1. Introduction

Supervised segmentation

In this chapter we study some *supervised* segmentation approaches. To this extent, the developed methodologies will differ from those of the previous chapters which focused on unsupervised segmentation methods. Deep Learning (DL) (Goodfellow et al., 2016) will also be introduced in this chapter; it is the core segmentation approach for the practical problem that we will address. Subsequent to the DL approach, in order to refine the results, we rely on a discriminative UGM model called Conditional Random Fields (CRFs) (Sutton and McCallum, 2012).

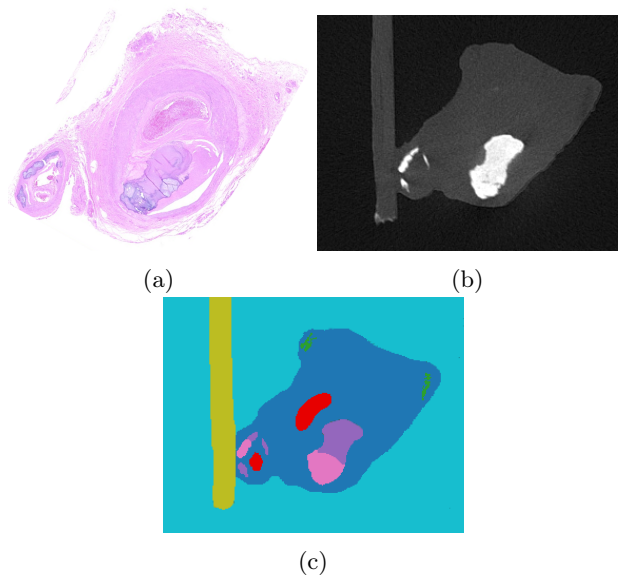
Remark: The content of this chapter has been developed after the creation of an annotated dataset of mCT images during the course of the thesis. Thanks to this valuable dataset we were able to study the complementarity of DL approaches with probabilistic graphical models. As we will see in this chapter, state of the art approaches in medical image segmentation often rely on mixing both of these approaches.

The purpose of the dataset is to gather knowledge to answer the problem of histological classification of the biological components on the mCT images of femoropopliteal arteries (see Figure 0.9 for the location of these arteries in the lower limbs). The end goal of the project is to perform a first automatic 3D histological analysis of the explant by just using a mCT image. Such an algorithm could save time, money and help process a growing number of biological data. Figure 4.1 depicts the dataset. To the best of our knowledge, there exists no similar study in the literature.

The whole project was created from scratch at the Geprovas laboratory since 2018: starting from the explantations of the arteries from human bodies to the output of the segmentation algorithms. The project is presented in two parts in this thesis. The rest of this chapter focuses on the mathematical aspects of this study, while details about the medical background of the project, the protocols and the full results are given in Chapter 5.

Deep Learning approaches for histopathological image analysis

During the last decade, the interest of the researchers towards DL approaches have skyrocketed, particularly in the field of image segmentation and analysis. The mathematical basics of DL is given in Section 4.2.1. Our study deals with the application of DL to medical image processing, and more precisely, histopathological image processing. Reviews of image processing techniques in this field are published on a regular basis, and more frequently since the beginning of the DL era which was popularized in (Krizhevsky et al., 2012). One can compare a pre-DL era review of histopathological image analysis such



(a) (b)

(c)

Figure 4.1.: A specimen from the created dataset: (a) the histological ground truth (optical microscopy), (b) the mCT scan and (c) the annotated mCT scan. Recall that the histological ground truth is not used by the methods described in this chapter. It is only needed for the human expert to correctly annotate the mCT image. The labels on the annotated image are *soft tissue*, *fatty tissue*, *sheet calcification*, *nodular calcification*, *thrombus*, *specimen holder* and *background*, respectively in blue, green, purple, pink, red, lime and cyan (see Figure 5.15). In the segmentation problem described in Section 4.5.1, the *thrombus* and *soft tissue* classes are merged.

as (Gurcan et al., 2009) with a recent one such as (Komura and Ishikawa, 2018) to understand the important role that DL now plays in the field.

In the context of histopathological image segmentation with DL, Convolutional Neural Networks (CNNs) are the reference approach which yields state of the art results (Srinidhi et al., 2019) (Haque and Neubert, 2020). The seminal paper (Krizhevsky et al., 2012) demonstrated that these approaches can learn features in the data that can outperfom expert modelizations from researchers which were traditionally proposed in machine learning. The mathematical foundations and a description of the CNN we use in our study are given in Section 4.2.

3D data segmentation with deep learning approaches

To conclude this introduction, we now focus on a particular point which is crucial to our application and for which CNNs currently reach their limits: the problem of volumetric semantic segmentation. Most often in the literature, CNNs work hand-in-hand with another machine learning approach in order to perform the volumetric segmentation. The final goal of this study is also to reconstruct a 3D volume using the 2D annotations that we possess.

This is unsurprisingly a common topic of interest in medical imaging. In general, the two main difficulties in this problem are, first, that the 3D annotations are almost inexistent, mainly because they are too complicated to make for human experts. Second, working with fully 3D CNNs is also computationally prohibitive. Several approaches built on 2D CNNs or related models have been proposed to provide the final 3D segmentations, we review them now.

The first type of approaches is to train directly a CNN to perform 3D convolutions to reconstruct a final 3D volume. The methodology can use a single fully 3D CNN (Çiçek et al., 2016) or use the 3D convolutions in hybrid approaches making use of previous results from 2D CNNs (Yang et al., 2018) (X. Li et al., 2018). Then because of the computational cost of 3D convolutions, one can note 2.5D approaches which aim at integrating spatial data with strong restrictions. For example, one can find propositions to consider successively small 3D chunks of consecutive slices (Ben-Cohen et al., 2016). One also finds the idea of combining segmentations from the three axes of the data cube to produce a segmentation that can take into account more spatial context than a single 2D segmentation (Zhao et al., 2017). Another type of popular approaches are based on Recurrent Neural Networks (RNNs) (Goodfellow et al., 2016) which process iteratively the slices through the datacube in successive passes to provide a segmentation taking into account all the axes (Novikov et al., 2018) (Cai et al., 2017). The final methodology that has been popularized during the last years and that we will follow is the 3D refinement with fully-connected Gaussian CRFs (Krähenbühl and Koltun, 2011) of 2D segmentations produced by a CNN. The first occurence of such a work can be found in (Kamnitsas et al., 2017). In Section 4.3, we present an approach built on a 3D refinement of the 2D dimensional segmentations with the fully-connected Gaussian CRFs.

Remark: A mathematical equivalence between computations in RNNs and Mean Field inference computations in CRFs has been demonstrated (Zheng et al., 2015). This thus brings these two approaches closer.

4.2. Convolutional Neural Networks

4.2.1. The Deep-Learning approach

DL is based on the paradigm that computers learn by themselves from the data the features that optimally answer the problem (Litjens et al., 2017). A DL model is composed of many layers that transform the input data up to an output layer which realizes the required task. Among DL models in the context of unsupervised learning (not treated in this thesis) one can cite Restricted Boltzmann Machines (Hinton, 2012), Deep Belief Networks (Bengio et al., 2007), or Variational Autoencoders (Kingma and Welling, 2013). CNNs (Fukushima and Miyake, 1982) (Krizhevsky et al., 2012) are deep models trained in a supervised setting. CNNs are the most popular deep models for medical image processing, they are described in 4.2.2. Another supervised technique for this task are the RNN (Hochreiter and Schmidhuber, 1997) approach which is gaining popularity (Litjens et al., 2017). RNNs are nonetheless out of the scope of this thesis.

4.2.2. Network definition

The U-Net network

In this work we construct a CNN based on the U-Net architecture (Ronneberger et al., 2015). Since its conception, the U-Net network has yielded state of the art results in the field of hispathological slice analysis and has become a reference approach (Falk et al., 2019) upon which many other architectures were proposed (Jégou et al., 2017) (Oktay, Schlemper, et al., 2018) (Zhou et al., 2018). The U-Net architecture is depicted in Figure 4.2.

The U-Net network retains the key elements of a CNN, particularly in the first part of the network, the *contractive path* or *encoder*. Convolution layers are the foundations of the network, they are parametrized by weights and biases and play the role of filters that extract features from the data. Each convolution layer is followed by a non-linear activation layer and a max-pooling layer. The latter successively reduces the dimension of the previous filtering operation, or *feature map*. This dimension reduction enables, in particular, to learn higher-level features (at different depths), which are likely to be more invariant to each training sample. Thus, the CNN predictions are more prone to generalize on unseen data. The *expansive path*, or *decoder*, is the second part of the network specific to network architectures from the U-Net family. In the decoder we find upsampling and concatenation operations which combine, respectively, the contextual, higher-order, features acquired in the encoder part with a precise higher resolution image. The network ends with a feature map of the size of the original image where all the pixels are assigned a score of

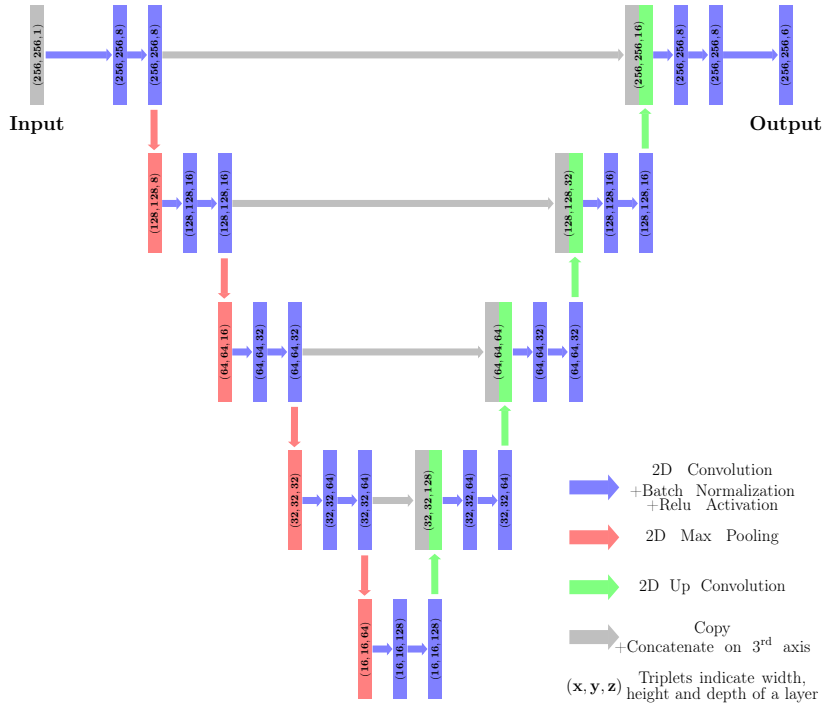


Figure 4.2.: The U-Net architecture. Data flow from left to right and first go through the contractive path (*encoder*) and then the expansive path (*decoder*). The network presented here has input images of size 256×256 (which is adjustable). Note that the Batch Normalization layer is not present in the original article (Ronneberger et al., 2015). Following the contractive path, the layers lose width and height but gain depth: we build a more condensed and higher-level representation of the data.

belonging to a certain class.

Remark: In our U-Net modelization we added Batch Normalization layers (Ioffe and Szegedy, 2015) after each convolution layer.

U-Net training and evaluation

The U-Net network is trained in a supervised manner by solving an optimization problem over the parameters (weights and biases) of the network with the help of the back-propagation algorithm (Rumelhart et al., 1986). The loss function of the optimization problem varies depending on the problem. The Dice loss function is a popular approach in the case of imbalanced dataset[1] such as ours (Salehi et al., 2017) (Sudre et al., 2017). The Dice score D is defined by:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{4.1}$$

where A and B are two images. We have the property that $\forall A, B, D(A, B) \in [0, 1]$. In the case where A and B are two segmentation maps for a same semantic class, a perfect segmentation, *i.e.*, a perfect superposition of A and B, yields $D(A, B) = 1$. In the worst case, *i.e.*, when A and B have no common pixel, we have $D(A, B) = 0$. From the Dice score we form the Dice loss $\bar{D}$ as $\forall A, B, \bar{D}(A, B) = 1 - D(A, B) \in [0, 1]$. The Dice metric has the advantage of being insensitive to the absolute number of pixels in a class, hence its popularity in case of imbalanced data set.

For a multi-class segmentation problem, we consider a generalized Dice score, the mean Dice score which is defined by the mean of the Dice scores (Fidon et al., 2017). For a L-label problem, where the segmentation maps to evaluate are organized as $\boldsymbol{A} = (A_1, \ldots, A_L)$ and $\boldsymbol{B} = (B_1, \ldots, B_L)$, the mean Dice score is defined by:

$$mD(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{L} \sum_{l \in L} D(A_l, B_l). \tag{4.2}$$

Since $\forall A_l, B_l, D(A_l, B_l) \in [0, 1]$, we have that $mD(\boldsymbol{A}, \boldsymbol{B}) \in [0, 1]$. The mean Dice loss is defined by $\forall \boldsymbol{A}, \boldsymbol{B}, m\bar{D}(\boldsymbol{A}, \boldsymbol{B}) = 1 - mD(\boldsymbol{A}, \boldsymbol{B})$. The latter is the loss function we use in the rest of this thesis.

The training is done on the *training set* which is composed of original slices but also augmented slices which come from data augmentation techniques (Shorten and Khoshgoftaar, 2019) (rotation, rescaling and gamma correction) applied to the original slices. Medical applications rely heavily on data augmentation techniques since medical data is often scarce, hard to gather and complex to annotate (Morra et al., 2019). A small portion[2] of the training set, the *validation set*, is kept apart and is not used in the back-propagation algorithm. It serves as a means to tune hyperparameters of the neural network such as the number of iterations, or *epoches*, in the learning process.

[1] In the sense that some semantic classes appear much less often than others.
[2] Typically 10% of the training set.

Once the training is achieved the network is evaluated on a third dataset, the *test set*, which have been totally held out during the training process not to introduce any biases in the performance evaluation. The score obtained on the test set then reflects the generalization capability of the network to segment unseen data.

4.3. Fully-connected Conditional Random Fields

4.3.1. Model definition

In this section we define the CRF model from (Krähenbühl and Koltun, 2011) that we use in a post-processing step to reconstruct segmented 3D volumes. We call this model fully-connected CRF (fcCRF). Recall the introduction on discriminative models from Section 1.5.3. The fcCRF model is a discriminative model that we then define with a Gibbs distribution:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{x}|\boldsymbol{y})\right), \tag{4.3}$$

where Z is a normalization constant. The energy function $E(\boldsymbol{x}|\boldsymbol{y})$ is composed of unary and pairwise factors:

$$E(\boldsymbol{x}|\boldsymbol{y}) = \sum_{s \in \mathcal{S}} \psi_u(x_s) + \sum_{(s,s') \in \mathcal{S}^2} \psi_p(x_s, x_{s'}). \tag{4.4}$$

To avoid complicated expressions, the dependency of the factors on the observations $\boldsymbol{Y} = \boldsymbol{y}$ will not explicitly appear. The unary potentials $\psi_u(x_s), \forall s \in \mathcal{S}, \forall x_s \in \Omega$, are taken as the label map outputs from the CNN classifier described in the previous section. The pairwise potentials are defined, $\forall (s, s') \in \mathcal{S}^2$, $\forall (x_s, x_{s'}) \in \Omega^2$, by:

$$\begin{aligned}
\psi_p(x_s, x'_s) &= (1 - \delta_{x_s}^{x_{s'}}) \sum_{m=1}^{2} w_m k_m(\mathbf{f}_s, \mathbf{f}_{s'}), \\
&= (1 - \delta_{x_s}^{x_{s'}}) \Bigg[w_1 \underbrace{\exp\left(-\frac{|s - s'|^2}{2\theta_\alpha^2} - \frac{|y_s - y_{s'}|^2}{2\theta_\beta^2} \right)}_{k_1(\mathbf{f}_s, \mathbf{f}_{s'})} + \\
&\quad w_2 \underbrace{\exp\left(-\frac{|s - s'|^2}{2\theta_\gamma^2} \right)}_{k_2(\mathbf{f}_s, \mathbf{f}_{s'})} \Bigg].
\end{aligned} \tag{4.5}$$

$\mathbf{f}_s, \forall s \in \mathcal{S}$, is a feature vector whose components are the location, s, and the observed value at this location, y_s. The weights w_1 and w_2 multiply, respectively, an *appearance* kernel, k_1, parameterized by θ_α and θ_β, and a *smoothness* kernel, k_2 parametrized by θ_γ. More precisely, the appearance kernel is the kernel used in bilateral filters (Paris, Kornprobst, et al., 2009) which is the

product of a Gaussian parametrized by the spatial distance and a Gaussian parametrized on the pixel intensity distance. This results in a filter capable of smoothing while preserving edges.

4.3.2. Optimized Mean Field Variational Inference for fcCRFs

The Mean Field Variational Inference algorithm

In this section, we study an approximate inference procedure for estimating marginals in the fcCRF model. The fully-connected property of the networks forbids any kind of exact computations and requires several approximations for the problem to be tractable. The standard approach is to perform Mean Field (MF) Variational Inference (VI) in the fcCRF model. We already introduced MF VI in Section 3.4.3. Here, observed random variables need to be considered. We minimize the quantity $\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y}))$, or, equivalently, maximize its opposite, with $p(\boldsymbol{x}|\boldsymbol{y})$ given in Equation 4.3 and $q(\boldsymbol{x}) = \prod_i q_i(x_i)$.

By following similar steps as in Appendix D.1 but integrating visible random variables, and using the energy from Equation 4.4, we find that[3], $\forall j \in \mathcal{S}$, $\forall x_j \in \Omega$:

$$q_j(x_j) = \frac{1}{Z_j} \exp\left(-\psi_u(x_j) - \sum_{x' \in \Omega} (1 - \delta_{x_j}^{x'}) \sum_{m=1}^{2} w_m \sum_{i \neq j} k_m(\mathbf{f}_j, \mathbf{f}_i) q_i(x') \right), \tag{4.6}$$

where Z_j is a normalization constant. The segmentation is then performed following the MPM criterion such that, $\forall j \in \mathcal{S}$:

$$\hat{x}_j^{MFVI} = \text{argmax}_{x \in \Omega} q_j(x_j). \tag{4.7}$$

The particular ordering of the summations in Equation 4.6 enables a MF algorithm divided in three steps which can be further optimized. The MF algorithm is given in Algorithm 4.1.

Remark: The update of all the q_j terms in Algorithm 4.1 can be done in a parallel fashion.

Optimization via approximate convolutions

In this section, we study the optimization proposed in (Krähenbühl and Koltun, 2011) for an efficient MF inference in fcCRFs. In terms of computational performances, the bottleneck of Algorithm 4.1 is the message passing step.

[3]The full demonstration can also be found in the supplementary material of (Krähenbühl and Koltun, 2011).

Algorithm 4.1: MF VI for fcCRFs (Krähenbühl and Koltun, 2011)

Data: $w_1, w_2, \theta_\alpha, \theta_\beta, \theta_\gamma$, the model parameters,
$\boldsymbol{y}$, the observations.
Result: $q(\boldsymbol{x})$, the MF approximation of $p(\boldsymbol{x}|\boldsymbol{y})$.
/* Initialize q */
$q_j(x_j) \leftarrow \exp\left(-\psi_u(x_j)\right), \forall j \in \mathcal{S}, \forall x_j \in \Omega$
while *convergence is not attained* **do**
  for $j \in \mathcal{S}$ **do**
    /* Message passing */
    for $x' \in \Omega$ **do**
      for $m \in \{1, 2\}$ **do**
        $\tilde{q}_{j,m}(x') \leftarrow \sum_{i \neq j} k_m(\mathbf{f}_j, \mathbf{f}_i) q_i(x')$ (Optimized in Section 4.3.2)
      end
    end
    for $x_j \in \Omega$ **do**
      /* Compatibility transform */
      $\hat{q}_j(x_j) \leftarrow \sum_{x' \in \Omega} (1 - \delta_{x_j}^{x'}) \sum_{m=1}^{2} w_m \tilde{q}_{j,m}(x')$
      /* Local update */
      $q_j(x_j) \leftarrow \exp\left(-\psi_u(x_j) - \hat{q}_j(x_j)\right)$
    end
    Normalize q_j
  end
end

We first start by rewriting the message passing step in terms of a convolution. For $m = 1$ we have:

$$\begin{aligned}\tilde{q}_{j,1}(x') &= \sum_{i \neq j} k_1(\mathbf{f}_j, \mathbf{f}_i) q_i(x'), \\ &= \sum_{i \in \mathcal{S}} k_1(\mathbf{f}_j, \mathbf{f}_i) q_i(x') - q_j(x'), \\ &= [\dot{G}_{\theta_\alpha, \theta_\beta} * \dot{q}(x')](\mathbf{f}_j) - q_j(x'),\end{aligned} \tag{4.8}$$

where $*$ is the convolution operator, $\dot{G}_{\theta_\alpha, \theta_\beta}$ is a Gaussian kernel in the *augmented space* and $\dot{q}$ is the q distribution in the *augmented space*. The notion of augmented space is needed to interpret the appearance kernel as a convolution. Details about the augmented space are given in Appendix E.1. For $m = 2$ we have:

$$\begin{aligned}\tilde{q}_{j,2}(x') &= \sum_{i \neq j} k_2(\mathbf{f}_j, \mathbf{f}_i) q_i(x'), \\ &= \sum_{i \in \mathcal{S}} k_2(\mathbf{f}_j, \mathbf{f}_i) q_i(x') - q_j(x'), \\ &= [G_{\theta_\gamma} * q(x')](\mathbf{f}_j) - q_j(x'),\end{aligned} \tag{4.9}$$

where G_{θ_γ} is a Gaussian kernel of parameter θ_γ. In this case of the smoothing kernel, the interpretation in terms of a convolution is straightforward.

For tractability of the MF VI algorithm, the methodology of (Paris and F. Durand, 2006) is used to approximate the convolutions. The observation is that a convolution acts as a low-pass filter, hence its result can be approximated by performing a convolution on a downsampled version of the filtered object. The result is then upsampled again. In practice, the downsampling step is taken equal to the kernel standard-deviation and the downsampling is done via averaging.

Remark: A further optimization consists in performing the multidimensional Gaussian filtering independently on each axis due to the separability of the Gaussian kernel. Note also that the parameters θ_α and θ_γ can be vary on each axis such that we have $\boldsymbol{\theta}_\alpha = (\theta_{\alpha,x}, \theta_{\alpha,y}, \theta_{\alpha,z})$ and $\boldsymbol{\theta}_\gamma = (\theta_{\gamma,x}, \theta_{\gamma,y}, \theta_{\gamma,z})$.

4.4. Markov Chain Variational Inference for fcCRFs

Chapter 3 illustrated the dramatic increase in performances in structured VI procedures over MF VI procedures. That is why in this section we propose to study the gain in the classical MF VI for fcCRFs with a Markov Chain (MC) structured VI inference approach. This is an original approach that we propose; to the best of our knowledge there is no study in the literature that deals with structured VI in fcCRFs.

4.4.1. Markov Chains for image processing

While MCs are intrinsically unidimensional data structures, they have been numerous examples of their use to process data of higher dimension (Bricq et al., 2008) (Meriem Yahiaoui et al., 2016) (Giordana and Pieczynski, 1997) (Lanchantin et al., 2011) (Fjortoft et al., 2003). The advantage of the MC structure resides in the small computational cost and the exact inference procedures that are available. The key step is to scan the higher dimensional space with a well designed path. The most common approach is to scan the space with space filling curves such as the Hilbert-Peano scan (Sagan, 2012). This is used in a 2D context in (Giordana and Pieczynski, 1997) and in a 3D context in (Bricq et al., 2008). However the path could be adapted, according to the specific application, for a better statistical segmentation (Courbot, Rust, et al., 2015) (Meriem Yahiaoui et al., 2016).

4.4.2. Scanning the data with Markov Chains

We developed the MC VI approach for inference in fcCRFs to answer the strong and limiting assumption of fully independent random variables in the variational distribution of the MF approach. Indeed, with the goal in mind to refine 3D segmentations produced by a simple stacking of 2D slices segmented by the CNN, it seems important to take as much spatial context as possible in the post-processing step. We introduced the concept of a MC scan in Section 4.4.1. In this section, we define another way to use MCs in the processing of higher dimensional data which is based on parallel and independent MCs. Indeed, for computational reasons, it is prohibitive to perform a Peano scan such as defined in (Bricq et al., 2008) for 3D volumes inside a VI procedure.

The variational distribution that we propose is made of independent MCs that goes through the data cube following one of the cube axes. This follows an idea also found in the Factorial Hidden Markov Models introduced in (Z. Ghahramani and M. I. Jordan, 1996). There is, however, no reason to privilege an axis over another, we then propose to perform parallel MC VIs along each axis and then to merge their result. More precisely, for a 3D image with dimensions $M \times N \times K$, respectively the height, width and depth. Then, we propose to perform six parallel MC VIs, one in each direction of each axis. This is illustrated in Figure 4.3.

For a given MC VI procedure, *e.g.*, in the cases *top-down* or *down-top* ($N \times K$ MCs of length M), the variational distribution is (recall the definition of a MC given in Equation 1.20):

$$l(\boldsymbol{x}) = \prod_{n=1}^{N \times K} l_1^n(x_1^n) \prod_{m=2}^{M} l_m^n(x_m^n | x_{m-1}^n). \tag{4.10}$$

The quantity to maximize, $-\mathbb{KL}(l(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y}))$, then becomes in this section:

$$-\mathbb{KL}(l(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{y})) = \mathbb{E}_{\boldsymbol{x}\sim l(\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{y}) - \log l(\boldsymbol{x})]. \tag{4.11}$$

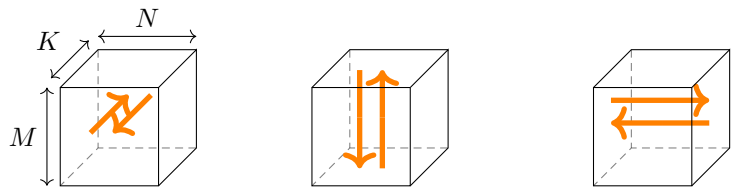


Figure 4.3.: The different MC VIs in the image data cube that can be run in parallel. (a) depicts the MC VIs of types *front-back* and *back-front*, each involving $M \times N$ MCs of length K in both directions of the third dimension. (b) depicts the MC VIs of types *top-down* and *down-top* ($N \times K$ MCs of length M in both directions of the first dimension). (c) depicts the MC VIs of types *left-right* and *right-left* ($M \times K$ MCs of length N in both directions of the second dimension).

The details of the derivation of the update equations for the MC VI approach are given in Appendix E.2. Once the update equations are found we use the same approach as shown in Section 4.3.2, that is, we interpret the bottleneck equation (summation over all the sites) as a convolution. Algorithm 4.2 summarizes the MC VI for fcCRFs. The update of all the terms of the variational distribution in Algorithm 4.2 can be done in a parallel fashion; this is the foundation of the practical MC VI methodology we present in Section 4.4.2.

Performing the MC VI procedures yields six posterior marginal probability distributions. Let these distributions be $\forall s \in \mathcal{S}, \forall x_s \in \Omega$, $l_s^{bf}(x_s)$, $l_s^{fb}(x_s)$, $l_s^{td}(x_s)$, $l_s^{dt}(x_s)$, $l_s^{lr}(x_s)$ and $l_s^{rl}(x_s)$, associated, respectively to the *front-back*, *back-front*, *top-down*, *down-top*, *left-right* and *right-left* MC VI procedure. Then the final posterior marginal probability distribution results from taking the product of each MC VI, at each site and for each class. That is, the final marginal probability distribution is defined by, $\forall s \in \mathcal{S}, \forall x_s \in \Omega$:

$$\mathring{l}_s(x_s) = l_s^{bf}(x_s) l_s^{fb}(x_s) l_s^{td}(x_s) l_s^{dt}(x_s) l_s^{lr}(x_s) l_s^{rl}(x_s). \tag{4.12}$$

From $\mathring{l}$ we perform the final segmentation following the MPM criterion such that, $\forall s \in \mathcal{S}$:

$$\hat{x}_s^{MCVI} = \text{argmax}_{x_s \in \Omega} \mathring{l}_s(x_s). \tag{4.13}$$

Remark: Other merging rules were tested such as a majority voting for each pixel based on the segmentations provided by each MC VI, or taking the maximal probability between each MC VI, at each site and for each class. Those criteria performed worse than Equation 4.12 and their results will not be reported.

Remark: The approach we propose is straightforwardly transposed in the context of 2D images by performing four parallel MC VIs (*top-down*, *down-*

Algorithm 4.2: MC VI for fcCRFs (example of *top-down* or *down-top* cases)

Data: $w_1, w_2, \theta_\alpha, \theta_\beta, \theta_\gamma$, the model parameters,
$\boldsymbol{y}$, the observations.
Result: $l(\boldsymbol{x})$, the MC approximation of $p(\boldsymbol{x}|\boldsymbol{y})$.
/* Initialize l */
$l_m^n(x_m^n|x_{m-1}^n) \leftarrow \exp\left(-\psi_u(x_m^n)\right), \forall n \in \{1,\dots,N\times K\}, \forall m \in \{1,\dots,M\}, \forall (x_m^n, x_{m-1}^n) \in \Omega^2$
Forward Backward (Algorithm A.2) to get the marginals of l
while *convergence is not attained* **do**
  for $n \in \{1,\dots,N\times K\}$ **do**
    for $m \in \{1,\dots,M\}$ **do**
      for $x' \in \Omega$ **do**
        for $r \in \{1,2\}$ **do**
          $\tilde{l}_{m,r}^n(x') \leftarrow$
          $\sum_{(m',n')\neq(m,n)} k_r(\mathbf{f}_m^n, \mathbf{f}_{m'}^{n'}) \sum_{x''\in\Omega} l_{m'-1}^{n'}(x'|x'') l_{m'-1}^{n'}(x'')$
          (Optimized in Section 4.3.2 by recognizing a convolution)
        end
      end
      for $(x_m^n, x_{m-1}^n) \in \Omega^2$ **do**
        $\hat{l}_m^n(x_m^n|x_{m-1}^n) \leftarrow \Big[\sum_{x'\in\Omega}(1-\delta_{x_m^n}^{x'}) \sum_{r=1}^2 w_r \tilde{l}_{m,r}^n(x') +$
        $\sum_{x_{m-1}^n\in\Omega} l_{m-1}^n(x_{m-1}^n|x_{m-1}^n) l_{m-2}^n(x_{m-2}^n) \times$
        $\log l_{m-1}^n(x_{m-1}^n|x_{m-2}^n)\Big]$
        $l_m^n(x_m^n|x_{m-1}^n) \leftarrow \exp\left(-\psi_u(x_m^n) - \hat{l}_m^n(x_m^n|x_{m-1}^n)\right)$
      end
      Normalize l_m^n
    end
  end
  Forward Backward (Algorithm A.2) to get the marginals of l
end

top, *left-right* and *right-left*). The final posterior marginal distribution is also computed by adapting Equation 4.12 to a product of four terms.

4.5. Experiments and Results

4.5.1. Segmentation via Deep Learning

We here describe the experiment of histological image segmentation and its results with the CNN described in Section 4.2.

Experimental set-up

The medical context surrounding this application is described in Chapter 5, the focus in this section is on the mathematical aspects of the problem.

The annotated dataset is made of 91 original slices of the type shown in Figure 4.1. We consider a 6-class segmentation problem; the mCT image pixels are then assigned to one of the 6 labels: *soft tissue*, *fatty tissue*, *sheet calcification*, *nodular calcification*, *specimen holder* and *background*. The class acronyms are respectively ST, FT, SC, NC, SH, Ba. Among the 91 annotated slices, 10 are held apart from the beginning of the experiment, they form the test set. The remaining 81 original annotated slices form the training set. Then the train and test sets undergo data augmentation algorithms (rescaling, rotation and gamma correction) (Shorten and Khoshgoftaar, 2019). This raises the train and test set to respectively 1620 and 200 elements. Eventually the train set is split into the actual train set and a validation set whose size is set to 10% of the former training set. The network we use is the same as in Figure 4.2 but with inputs of size 512×512 (images are cropped to fit this size). The network is then trained with respect to the mean Dice loss function. The experiments are coded in Python3 using the Tensorflow[4] library.

Results

Figure 4.4 illustrates the evolution of the mean Dice loss on the train and validation set during training. We can see that, after 50 iterations, the loss is stabilized and the training stops. Table 4.1 gives the Dice score for each class on the test set. We can see some important discrepancies in the results between the classes. These discrepancies reflect the intrinsic complexity of the dataset. Indeed, the latter presents high class imbalances, high similarities between classes and notable differences between the image dynamic ranges. Note however that one key goal was to distinguish, on mCT images, the two types of calcifications namely, SC and NC. This is an important histopathological aspect. While such a task seems extremely complex for the naked eye, our DL approach can, to some extent, answer this problem. Future research might further improve this point. An example of a 2D prediction made by the CNN on a slice from the test set is given in Figure 4.5.

[4] https://www.tensorflow.org

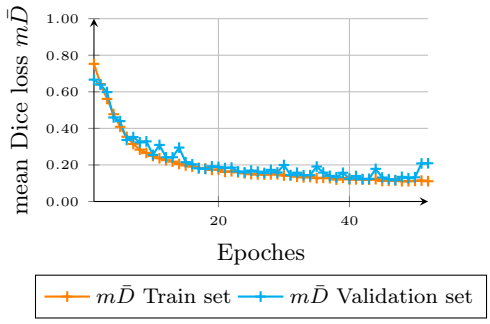


Figure 4.4.: Mean Dice loss values on the train and validation set as a function of the epoches. Recall that the training is stopped according to the behavior of the validation loss.

	ST	FT	SC	NC	SH	Ba
Dice	0.94	0.41	0.85	0.64	0.86	0.99

Table 4.1.: Dice scores on test set for each class.

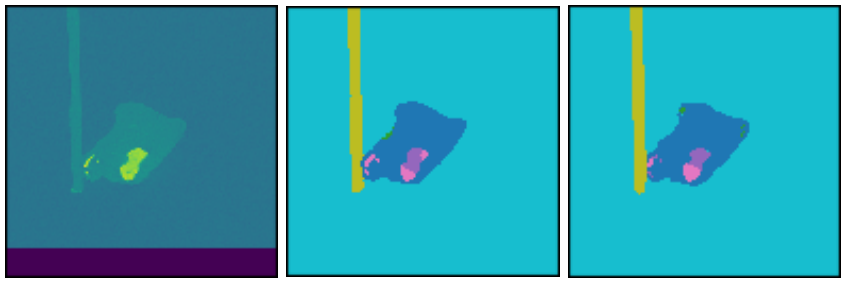
More results and in-depth discussions about this semantic segmentation problem and some histopathological interpretations are available in Chapter 5.

4.5.2. Experimental comparisons of MF VI and MC VI on semi-real images

Before applying the fcCRF VI procedures on real data, we evaluate the relative performance of MF VI and MC VI on synthetic data. We consider the case of unsupervised grayscale two-class image segmentation corrupted by additive Gaussian noise. The set of hidden classes is $\Omega = \{\omega_1, \omega_2\}$. The additive Gaussian noise is parametrized by a mean and a standard deviation for each class: $(\mu_\omega, \sigma_\omega), \forall \omega \in \Omega$. A first segmentation is performed using a Gaussian Mixture Model (GMM) (Murphy, 2012). The posterior probabilities of the GMM are used for the initialization of the unary potentials of the fcCRF model, for each pixel and for each class.

We define as err_{GMM}, $\text{err}_{MFVI-Pe}$ and $\text{err}_{MCVI-Pa}$ the error rates in the segmentation with, respectively, the GMM, the fcCRF model with MF VI, the fcCRF model with MC VI with Peano scan and the fcCRF model MC VI with parallel MCs. In all cases we fix the number of iterations in the VI procedures to 50. In the following, we provide experiments offering a general overview of the behavior of the models.

In Figure 4.6 we plot the error rates of the approaches for the unsupervised segmentation of images. First we emphasize the fact the MC VI approach appears as more efficient for inference in fcCRFs than the classical MF VI approach both with the parallel MCs and the Peano scan implementations. In particular, in the case of Figure 4.6b, MC VI offers up to a 8 point gain in the



(a) MicroCT slice (b) CNN predictions (c) Ground truth

Figure 4.5.: Example of 2D prediction by the CNN on a sample from the test set. It is the same sample as illustrated in Figure 4.1 where, as noted, the classes *soft tissue* and *thrombus* have been merged. The colormap is given in Figure 5.15.

error rate over MF VI, and provides a better or equivalent error rate for all noise levels. We also note that the GMM approach is not competitive against the three others, its main drawback lies in its inability to take into account any spatial context. The fcCRF approach appears as an efficient post-processing technique which is built upon the probability map given by the GMM method. The fcCRF approach significantly improves the GMM segmentation with more than 10 point gains in segmentation error. Figure 4.6 provides evidence of the similar performances of the parallel MCs and Peano scan implementations of the MC VI procedures in terms of error rate. However they are not equivalent in terms of computational cost. Indeed, in Figure 4.7 we plot the time taken for each procedures as a function of the size of the image[5]. The Peano scan approach suffers mainly from the fact that it cannot be parallelized and must process very long MCs. This experiment discards the use of the Peano scan in favour of our proposed parallel MCs scan.

In Figure 4.8 we plot the surface whose points are defined by $\text{err}_{MFVI} - \text{err}_{MCVI-Pa}$ where the varying parameters are the kernel weights w_1 and w_2[6]. Except for the pathologic case $w_1 = 0$, the surface error is always above 0, suggesting that the MC VI (with parallel MCs) approach that we propose always gives better results that the classical MF VI approach for inference in fcCRFs.

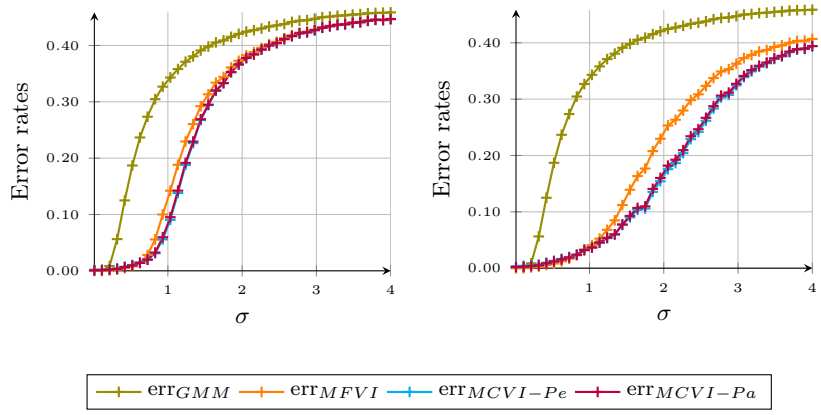
Finally, in Figure 4.9 we illustrate graphically the results on particular images. The images come from the "1070-Binary Shape Database"[7].

We conclude this section by noting the evidence of the advantage in favour of the MC VI procedures (both with Peano scan and parallel MCs) for fcCRFs in the case of image segmentation. While we did not work on finding the op-

[5] Note, moreover, that the use of the Peano scan can only be done on images whose dimensions are a power of two, thus making this approach less practical.

[6] We empirically found that these parameters are more critically related to the segmentation performance than the kernel standard deviations.

[7] https://vision.lems.brown.edu/content/available-software-and-databases



(a) $w1 = w2 = 1$ (b) $w_1 = 1, w_2 = 2$

Figure 4.6.: Error rates in unsupervised segmentation of the two VI procedures as a function of the noise level $\sigma = \sigma_{\omega_1} = \sigma_{\omega_2}$. The other parameters are fixed to $\mu_{\omega_1} = 0$, $\mu_{\omega_2} = 1$, $\theta_\alpha = \theta_\beta = \theta_\gamma = 1$. The new MC VI methodologies appear to be the best performing methods for all noise levels. The results are here averaged on 10 randomly chosen images from the dataset.

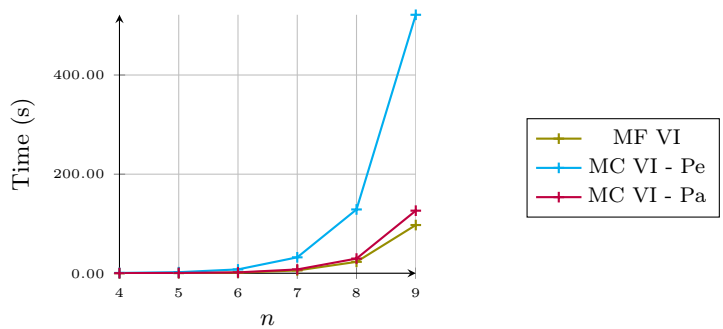


Figure 4.7.: Timing of the VI procedures as a function of the size of the image (a size n means an image of dimensions $2^n \times 2^n$).

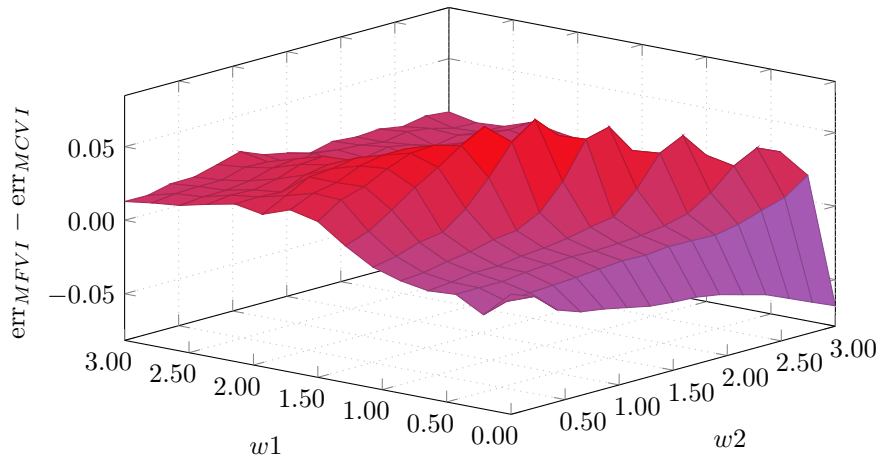


Figure 4.8.: $\mathrm{err}_{MFVI} - \mathrm{err}_{MCVI}$ as a function of w_1 and w_2. The other parameters are $\mu_{\omega_1} = 0$, $\mu_{\omega_2} = 1$, $\sigma_{\omega_1} = \sigma_{\omega_2} = 1.5$, $\theta_\alpha = \theta_\beta = \theta_\gamma = 1$. The results are averaged over 10 simulations.

timal set of parameters for the models, we performed several experiments that exhibited better results for the MC VI for any non-pathological parameter set. Similarly to the conclusion of Chapter 3, the structured VI approach seems to better take into account the spatial context thanks to the MCs, leading to better segmentation rates.

Remark: In the following, all references to MC VI will refer to the parallel MCs version of MC VI.

4.5.3. Post-processing with fcCRFs

The rest of this chapter is dedicated to improving the CNN segmentation with a post-processing method based on fcCRFs. We here compare the results in the segmentations provided by the CNN (Section 4.2.2), by a fcCRF post-processing based on MF VI (Section 4.3.2) and by a fcCRF post-processing based on the new MC VI (Section 4.4).

MF VI with fcCRFs for segmented 3D volumes

In this section we employ the fcCRFs with MF VI to segment 3D data: the whole artery can be reconstructed in 3D. Figure 4.10 illustrates the resulting segmentation on a mCT from a test artery. The post-processing enables taking into account the third axis to reconstruct a volume with spatial context from all the dimensions. More segmentation are provided in Chapter 5.

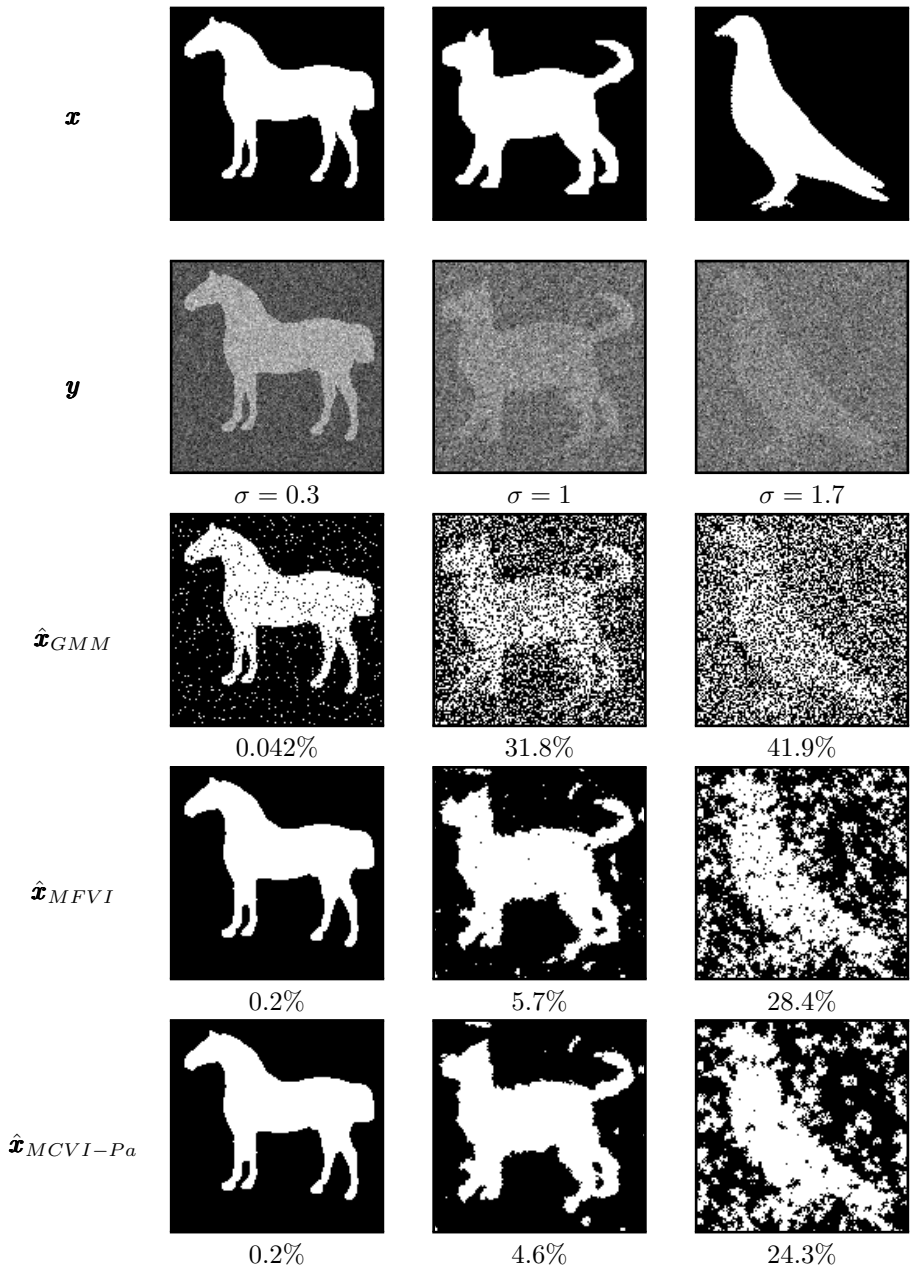


Figure 4.9.: Unsupervised segmentation of semi-real images with the two VI procedures on variable images and noise levels $\sigma = \sigma_{\omega_1} = \sigma_{\omega_2}$. We have $w_1 = 1, w_2 = 2, \mu_{\omega_1} = 0$, $\mu_{\omega_2} = 1$, and $\theta_\alpha = \theta_\beta = \theta_\gamma = 1$. The parallel MCs approach was taken for the MC VI.

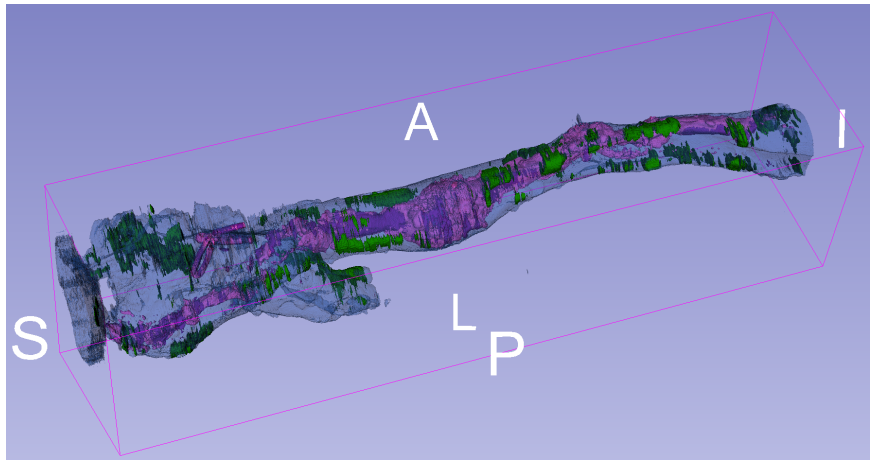


Figure 4.10.: Example of a 3D prediction by the CNN on a mCT from a test artery. We note that the irregularities found in the *fatty tissue* class in green are to be expected since this class exhibit the worst Dice score. The *support* class has been made invisible for clarity and the opacity of the *soft tissue* class has been lowered. The colormap from Figure 5.15 is used.

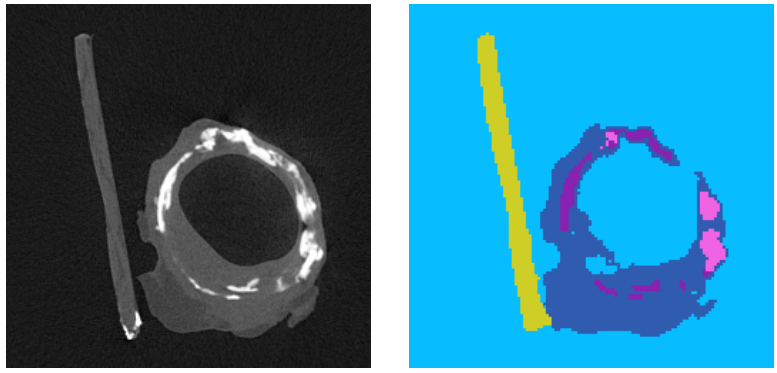
The *soft tissue* class misclassifications

We focus on the problem of correcting spurious classifications from the CNN which erratically misclassifies the *soft tissue* class into the *background* class leading to an inhomogeneous segmented volume. This problem is similar to (Kamnitsas et al., 2017), which also uses the fcCRFs to a similar end. The problem is described in Figure 4.11.

(Kamnitsas et al., 2017) uses fcCRFs and MF VI to process the CNN output. They rely on the smoothing capabilities and the modeling of the 3D context provided by the fcCRFs to solve the misclassifications. Illustration of such process is done in Figure 4.12a. However, MF VI does not seem suitable in our case: it results in the loss of the small *calcification* pixels which are essential in our application. Those unsatisfactory results made us consider a new methodology based on MC VI. It is described in the next section.

Using MC VI with fcCRFs to recover the missing *soft tissue* class

In this section we describe the processing based on MC VI. MC VI offers the additional possibility to refine the segmentation with known constraints from the application. Indeed, the VI procedure produces a non-stationary transition matrix between the states in Ω. One then has an understable view of the model that has been learnt before the final inference in the variational distribution with the Forward Backward algorithm. Since a MC transition matrix is easily comprehensive, we also propose to adjust the probabilities that are learnt in



(a) mCT image (b) CNN segmentation

Figure 4.11.: Illustration of spurious classifications CNN of the *soft tissue* class. While clearly visible by the naked eye, the *soft tissue* class is only partially segmented on this slice. A major cause of this problem might be the lack of context from the depth axis in the slice-by-slice segmentation performed by the CNN. The color code from Figure 5.15 is used.

order to correct learnt transitions that does not fit the desired application.

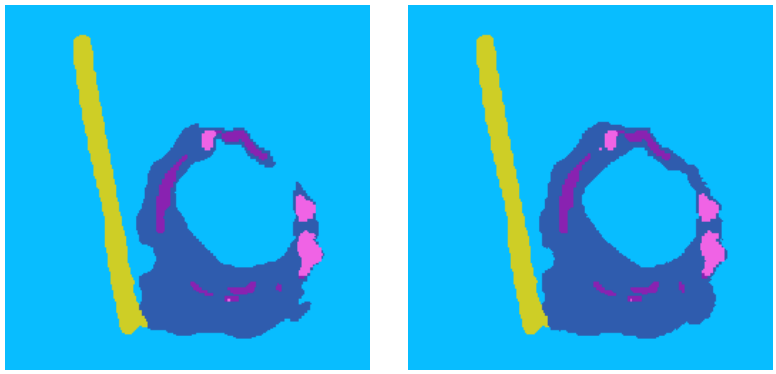
We want to *penalize* the predictions made in favour of the background class in the areas were the *soft tissue* class is missing. This can be done thanks to the transition matrix of the MCs that are learnt during the MC VIs processes. A similar idea of matrix adjustment to exhibit class relationships to improve classifications predictions can be found in (Fidon et al., 2017) for example.

We start by detecting the areas where the soft tissue is missing by using a binary closing operator. Then, within these areas denoted $\mathcal{A}$, at each iteration of the MC VI, the probability of transitioning from any class to the *background* class is lowered so that, for the transition matrix of any MC VI, with ω_{bck} being the *background* class:

$$\forall \omega \in \Omega \setminus \{\omega_{bck}\}, \forall s \in \mathcal{A}, l_s(\omega_{bck}|\omega) = \min_{\omega' \in \Omega} l_s(\omega'|\omega). \tag{4.14}$$

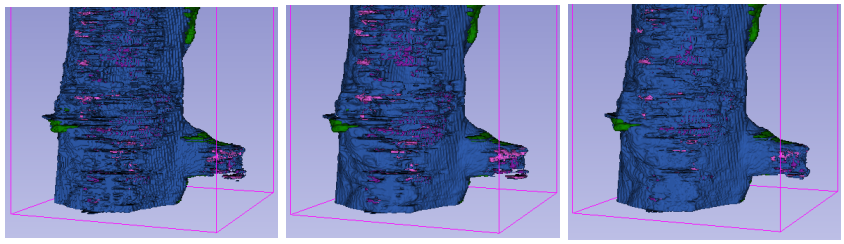
Such an adjustment improved greatly the segmentations enabling the *soft tissue* class to be recovered thanks to the spatial context brought by each VI. An example of MC VI is given in Figure 4.12b (single slice illustrated) and in Figure 4.13c (3D segmented volume). We note how the MC VI methodology leads to the most refined results. It is also computationally efficient since both the MC VI procedures and the processing of each MC inside each of the procedure are parallelized. The computational bottleneck of MC VI remains the same as in MF VI, namely, the bilateral filtering step.

Remark: Such an adjustment in the MC VI procedure does not possess an equivalent adjustment within the MF VI procedure. The closest experiment



(a) $\hat{\boldsymbol{x}}^{MFVI}$ (b) $\hat{\boldsymbol{x}}^{MCVI}$

Figure 4.12.: Illustration of fcCRF post-processings in 3D of the image in Figure 4.11. In (a), the case of the MF VI, we can see that the spatial context offers some improvement to solve the misclassification from the *soft tissue* class. However this comes at the price of high kernel variances for smoothing which leads to loss of details in the calcifications parts which are the most important components of the image. However in (b), the case of the MC VI with modified transition matrix, we can see that our proposed approach offers the best improvements while preserving the most the details within the segmented calcifications. None of the approaches could recover the missing calcification in the segmentation. For both VIs the parameters were set to $w_1 = 2, w_2 = 2, \theta_{\alpha,x} = \theta_{\alpha,y} = \theta_{\alpha,z} = 4, \theta_{\gamma,x} = \theta_{\gamma,y} = \theta_{\gamma,z} = \theta_{\beta} = 2$. The color code from Figure 5.15 is used.



(a) CNN segmentation (b) $\hat{\boldsymbol{x}}^{MFVI}$ (c) $\hat{\boldsymbol{x}}^{MCVI}$

Figure 4.13.: Illustration of fcCRF post-processings in 3D (full volumes). The same parameters and colormap as in Figure 5.15 are used. The smoothest segmentation is obtained in (c) with the proposed MC VI approach thanks to its ability to better correct the spurious CNN classifications.

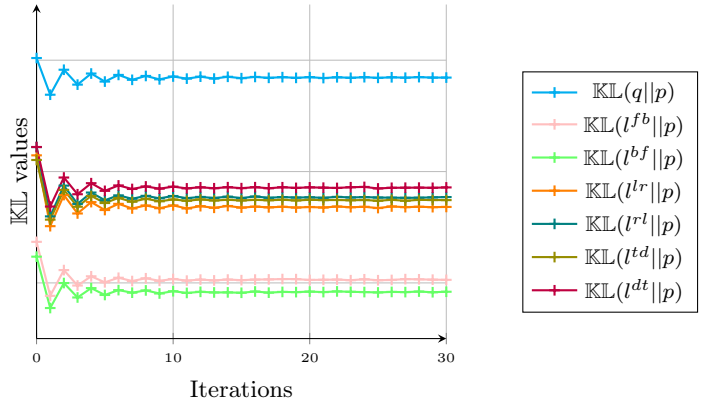


Figure 4.14.: KL values (up to a constant) for the VI procedures as a function of the iterations regarding the experiment from Figure 4.12. We can see that all the procedures converge in a few iterations.

that was carried consisted in a direct modification of the estimated marginal probabilities $q_s(x_s), \forall s \in \mathcal{S}$, in the same way as done in Equation 4.14, which resulted in bad performances.

Remark: In all this section, as suggested by (Kamnitsas et al., 2017), the parameters were manually tuned. Further developments might consider an optimization procedure for automatic fcCRF parameter selection.

Convergence of the VI procedures

Finally we give an insight about the convergence of all the VI approach. This is in done in Figure 4.14. We can see that in all cases the convergence is very quick to happen. As discussed in the supplementary material of (Krähenbühl and Koltun, 2011), the KL values (Equation 4.11) are computed up to a constant because the partition function of $p(\boldsymbol{x}|\boldsymbol{y})$ cannot be computed. The partition function can be discarded since it appears as an additive term which only depends on the observations $\boldsymbol{y}$ and the CRF parameters which are constant accross the VIs.

4.6. Conclusion

In this chapter we treated an unique dataset of mCT of atherosclerotic arteries, whose histological ground truths have been annotated by an expert during the course of the thesis. That is why we studied supervised segmentation approaches and more precisely, deep learning approaches. We also studied the complementary of deep learning and probabilistic approaches.

We first processed the data and trained a CNN to segment any new mCT data of atherosclerotic data into six histologically meaningful classes. We then

studied the problem of using the CNN outputs to reconstruct a 3D segmented volume. We used the fcCRF probabilistic model with both its classical MF VI procedure and a new structured MC VI procedure. The latter seems to enable a much better use of the spatial context than the former, leading to better results, both on synthetic and real data. Thus, we brought new solutions to the problem of volumetric semantic segmentation introduced at the beginning of the chapter.

More medical context, interpretations and results, as well as illustrations of this study are available in Chapter 5.

Chapter 5.

Applications to vascular surgery

Contents

5.1. Introduction

The previous chapters were dedicated to the mathematical presentations of new probabilistic models motivated by new challenges in the field of image processing for medical imaging. In this chapter, the focus is on the application of some of the models discussed previously on real problems and real data collected in the field of vascular surgery.

The first section describes the development of a Hidden Markov Chain model for the precise segmentation of stents corrupted by artifacts in X-ray images. We then present additional results of the applications of the Gaussian Pairwise Random Field model (Chapter 2) and of the Spatial Triplet Markov Tree model (Chapter 3) for the segmentation of organic material and calcifications in X-ray images. The third section completes the presentation and evaluation of the U-Net Convolutional Neural Network (Chapter 4) with more background on the histologic problem, on the dataset construction as well as additional illustrations.

5.2. Unsupervised segmentation of stents corrupted by artifacts

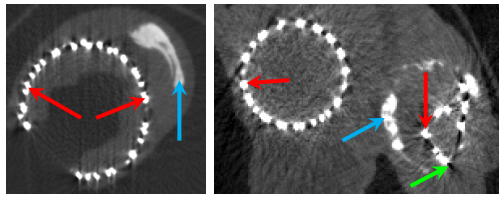
5.2.1. Context and motivation

In this section, we describe a statistical approach based on the Hidden Markov Chain (HMC) model, introduced in Section 1.6.1, to finely segment stents in medical images corrupted by strong artifacts.

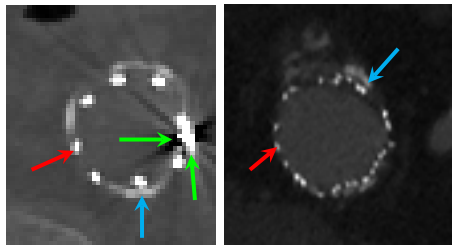
Medical context

The *in vivo* behavior of the stent is far from being fully understood (Lejay et al., 2018). However, the amount of medical images is constantly growing, offering new opportunities to learn from clinical cases. The interactions between the stent struts and the calcifications are believed to be a major cause of failure of the treatments based on the implantation of a stent. The study of such interactions by combining information of both CT and mCT scans is a new approach to the problem, it requires the development of new imaging tools which can finely segment the metallic stent components.

The most complex images to process often include a broken stent in a calcified environment. Images are also often corrupted by strong artifacts. Figure 5.1 illustrates the typical images from which one needs to segment the stent. It appears that the task of unsupervised segmentation in X-ray scans is a complex problem. Notably, artifact and calcification pixels are close to the stent pixels both in intensity, geometry and localization, which tends to produce stent False-Positive classifications with traditional segmentation methods as we will see in Section 5.2.3.



(a) mCT scan



(b) CT scan

Figure 5.1.: Example of input data (2D views), we see the complex environment in which the stent lies: artifacts and calcifications are notable. Some stent, calcification and artifact components are indicated by, respectively, red, blue and green arrows. We note the spatial resolution differences between CT scans and mCT scans.

Related literature

To the best of our knowledge, there exists no dedicated method to confidently segment stents in such images in an unsupervised fashion[1]. As a consequence, people still use simple approaches such as manual thresholding which fails in complex cases. Thus, such cases are discarded from clinical datasets despite their medical interest (Perrin et al., 2016).

Relatively few works deal directly with the topic. Works such as (Klein et al., 2012) or (Langs et al., 2011) that also address stent segmentation are not suitable for us since they do not consider the stent in a calcified environment. In our case, the calcifications and the stent are so close in appearance and sometimes in geometry that dedicated methods need to be developed to distinguish both classes during segmentation. Moreover, we have to discard approaches such as (Demirci et al., 2011) since they are stent specific and we need our method to be independent on the stent model to process our data.

However, the problem of stent segmentation is somehow very close to the problem of blood vessel segmentation (see (Lesage et al., 2009) for a review on the topic which concerns traditional approaches). Notably, some multi-scale filters have met success in the task of blood vessel segmentation (Frangi et al., 1998) (Merveille et al., 2014), these methods will be used for comparisons in the result section. During the last years, Deep Learning approaches are performing

[1]When this project was carried at the beginning the thesis, we had very few and only unannotated data at our disposal.

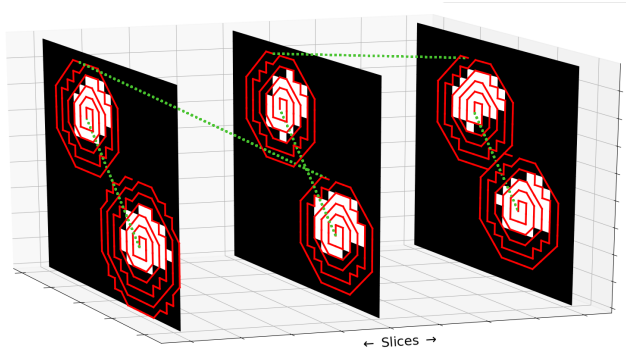


Figure 5.2.: The HMC path, to transform 3D data into 1D data, illustrated on 3 successive slices.

state of the art results in the context of supervised segmentation. Results using this technique are also available for 3D vessel segmentation (Livne et al., 2019), and could potentially be adapted to 3D stent segmentation. However, this goes out of the scope of this work which deals with unsupervised segmentation.

Finally, a point should be made about a very close and active research topic which is Metal Artifact Reduction (MAR) (Verburg and Seco, 2012) (Y. Zhang and Yu, 2018). MAR focuses on improving the quality of images corrupted by metallic artifacts. Many MAR approaches first rely on a segmentation of the metallic elements in the image, the work presented in this section contributes to the improvement of MAR techniques. The final result after the MAR step will be used to assess the quality of the stent segmentation in Section 5.2.3.

5.2.2. A HMC dedicated to handling strong artifacts

The HMC path

We already mentioned in Section 4.4.1 the importance of the transformation of higher dimensional data into one dimensional data, so that the latter can be processed by a MC.

In this section we propose a new kind of path dedicated to handling the artifacts. Each connected component (of the *stent* class from the initial segmentation) of each slice is visited using a snail path (snail paths are described in (M. Yahiaoui et al., 2014)). The sense of rotation is alternated between each snail path. This way, physical continuity in the path is simulated, gathering all the stent parts in a 1D sequence which allows efficient computation. Such a path is particularly suited to get a smoothed and regular surface for tubular structures (Courbot, Rust, et al., 2016). Moreover, in the HMC model one has to learn the transition probabilities between the classes. A small morphologic dilation is then performed over the result of the initial segmentation (see the end of this section) for the snail path to be a little bigger than the connected components isolated during the initial segmentation. Thus, we make sure that

all the voxels belonging to the *stent* class are part of the sequence of pixels to classify.

Figure 5.2 illustrates our choice for the path of the MC through the slices of our data cubes.

The HMC statistical model

The segmentation problem presented in this section aims at classifying pixels into two classes, *stent* and *rest*, in an unsupervised fashion. Recall the context of Bayesian image segmentation introduced in Section 1.5.2. Let $\boldsymbol{X}$ be the vector of hidden random variables of the class image taking their value in $\Omega = \{\text{stent}, \text{rest}\}$. Let $\boldsymbol{Y}$ be the vector of random variables representing the observed grayscale image (the X-ray scan), the elements of $\boldsymbol{Y}$ take their value in $\mathbb{R}$.

If the hypothesis of independent Gaussian noise is made then the parameters that we need to learn are the initial probabilities of the MC, the transition probabilities of the MC, the mean and the standard deviation of the Gaussian distribution associated with each class. The parameters are learnt unsupervisedly with the SEM algorithm (Algorithm A.5) similarly to the parameter estimation steps described in Chapters 2 and 3. The statistical inference step is performed by the FB algorithm (Algorithm A.2) followed by the MPM algorithm (Algorithm A.6).

Improving the noise model to handle strong artifacts

Classically, such as in (Courbot, Rust, et al., 2016), the noise model in a HMC, also called conditional likelihood, is chosen to be a mixture of Gaussian distributions (MoG). Using developments on generalized mixture models in Hidden Markov Models such as (Delignon et al., 1997) and (Pieczynski, Bouvrais, et al., 2000), we propose to work with a noise model involving mixture of exponential distributions (MoE) to improve the results. MoEs have already been explored for Hidden Markov Trees in (Monfrini and Pieczynski, 2005). Indeed, the "all-or-nothing" behavior enabled by using the exponential distribution seemed particularly suited to handle the strongest artifacts. The exponential probability density function is given by:

$$f(x;\lambda,\delta) = \begin{cases} \lambda \exp\left(-\lambda\left(x-\delta\right)\right) & x > \delta, \\ 0 & x \leq \delta, \end{cases} \tag{5.1}$$

where $(\lambda, \delta) \in (\mathbb{R}^{+*}, \mathbb{R})$ are the new parameters to estimate for each class in the SEM procedure.

Moreover, a Bayesian Information Criterion (BIC) (Wit et al., 2012) score comparison was conducted and we showed that a MoE fitted better the empirical distribution (obtained after the coarse segmentation) for the *stent* class and the *rest* class than a MoG. It is described in Appendix F.1.

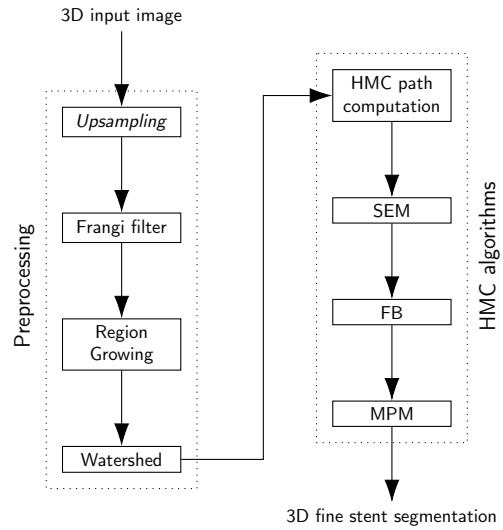


Figure 5.3.: Diagram of the fine stent segmentation step.

The complete segmentation approach

The HMC needs to be initialized by an initial segmentation, we propose that its initialization correspond to the response of a Frangi filter (Frangi et al., 1998) followed by a region growing algorithm and a watershed algorithm (Haidekker, 2011). This forms the *Preprocessing* part of Figure 5.3.

The Frangi filter is first performed and we keep only voxels whose probability to belong to a tubular structure is above a certain threshold (dependent on the nature of the image we want to segment). A region-growing algorithm is performed in 3D to select only the biggest connected component (the pixels from the *stent* class). A watershed algorithm is performed on every slice to separate connected components that should be disjoint but still appear stuck together because of the artifacts.

Figure 5.3 summarizes the complete segmentation approach.

Remark: A step of upsampling is needed for images of very low resolution such as CT scans (see, *e.g*, Figure 5.1b). The idea is that the stent parts on the image should be increased up to a certain width (we used 10 pixels in our applications) so that the successive filters used for segmentation show responses.

Remark: The Frangi filter is the most important part of the *Preprocessing* algorithm. The subsequent HMC algorithms can be seen as a statistical refinement of the Frangi filter response, increasing the robustness of the latter and making it fully automatic.

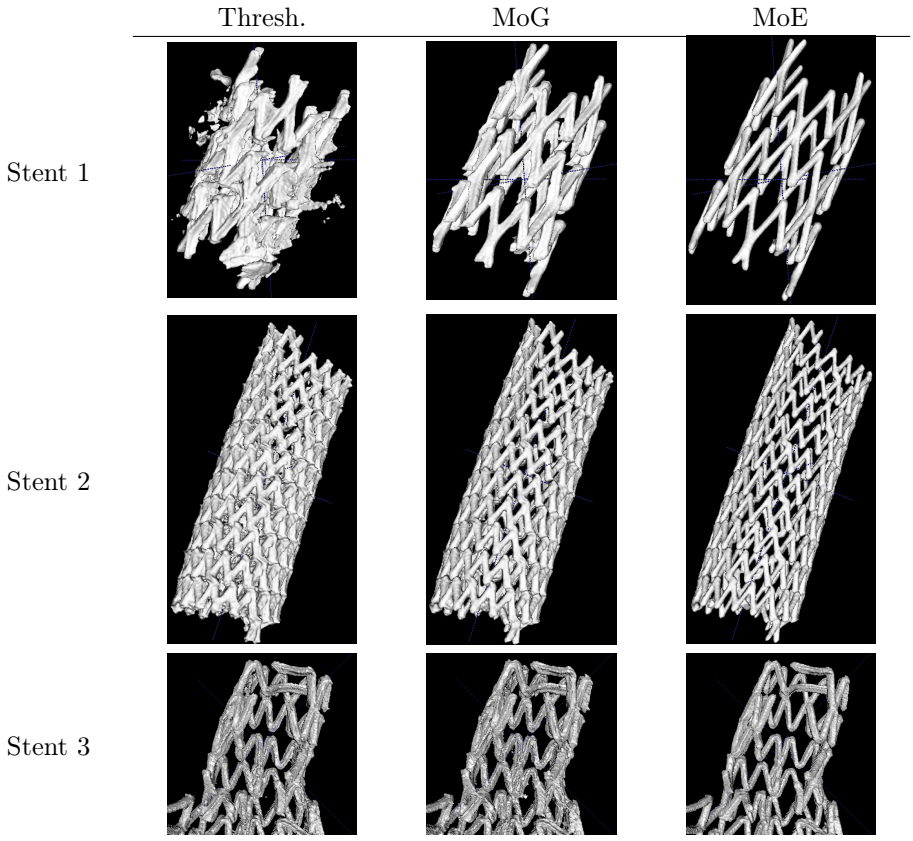


Figure 5.4.: 3D segmentation of stents via manual thresholding (Thresh.), MoG approach and MoE approach.

5.2.3. Results

We begin by illustrating segmentations by our *fine* segmentation method. We start by considering the refinement introduced by using a noise model involving a mixture of exponential distributions, which better reflects the nature of the artifacts in the images. Indeed, as depicted in Figure 5.4 on several cases with very strong artifacts, the recovered stent structure is much thinner and much more circular as the original metallic structure of the stent. Figure 5.5 illustrates some more segmented 3D meshes of broken stents.

As stated in the introduction, we can apply our new metal segmentation method to improve Metal Artifact Reduction (MAR) procedures. Beam Hardening Correction (BHC) is built upon the Linear Interpolation technique and they form the most classical approaches to MAR (Verburg and Seco, 2012) (Y. Zhang and Yu, 2018). We show, with Figure 5.6, on various CT scans, restored images with reduced artifacts using our *fine* statistical segmentation approach followed by BHC. Results are compared to a manual segmentation, to a seg-

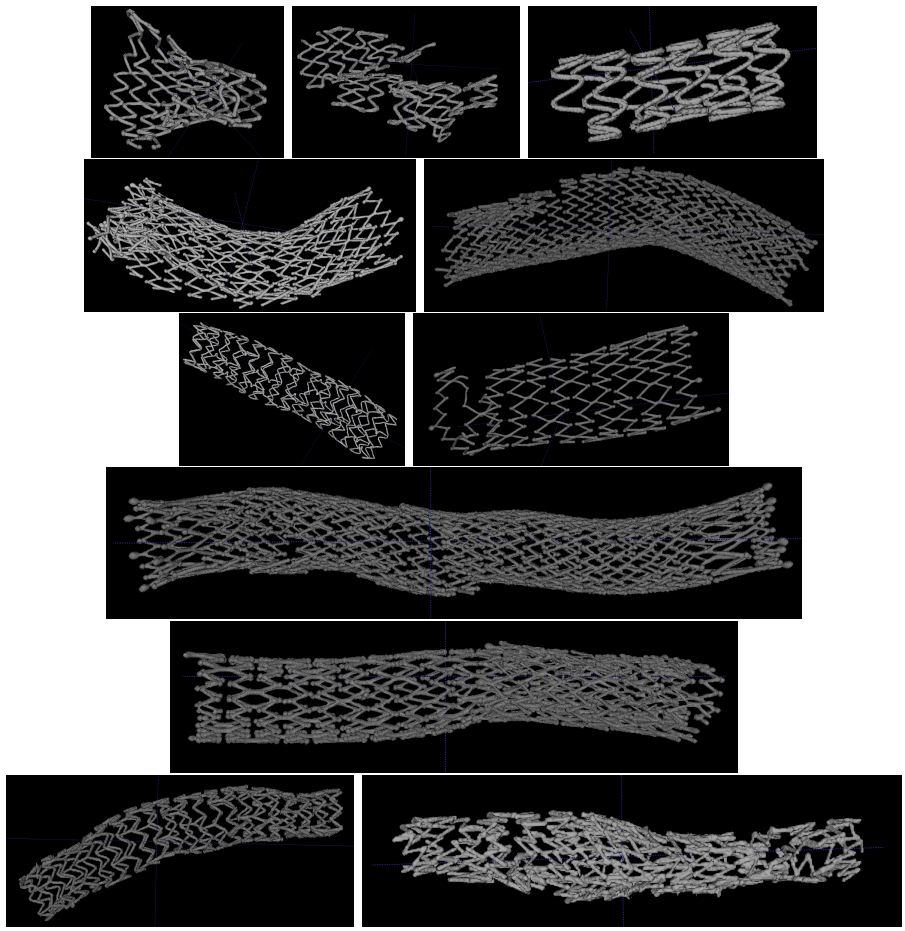


Figure 5.5.: 3D views of segmented mCT and CT stent images from the database.

mentation based on Frangi filter[2] (Frangi et al., 1998) and to a segmentation based on the RORPO filter[2] (Merveille et al., 2014), all three followed by BHC. Note that all methods operate in 3D but 2D slices are presented for a better visual assessment of the results. Such results highlight the importance of the metallic segmentation step in MAR problems.

The absence of ground truth forces us to make a qualitative assessment of the segmentations. Our algorithm is the most likely to offer a precise refinement over the *stent* pixels as well as avoiding False-Positive classifications resulting from calcifications or artifacts. The precise segmentation that we propose reduces the remaining artifacts at the end of the MAR procedure.

Remark: It is notable that the results from the three methods used for comparisons depend on a final step of manual thresholding. On the contrary, the proposed HMC does not need such manual intervention since the statistical refinement is fully automatic, avoiding a time consuming step for the expert which is also subject to the operator subjectivity.

[2]This filter response over the data cube is manually segmented afterwards, with a global threshold.

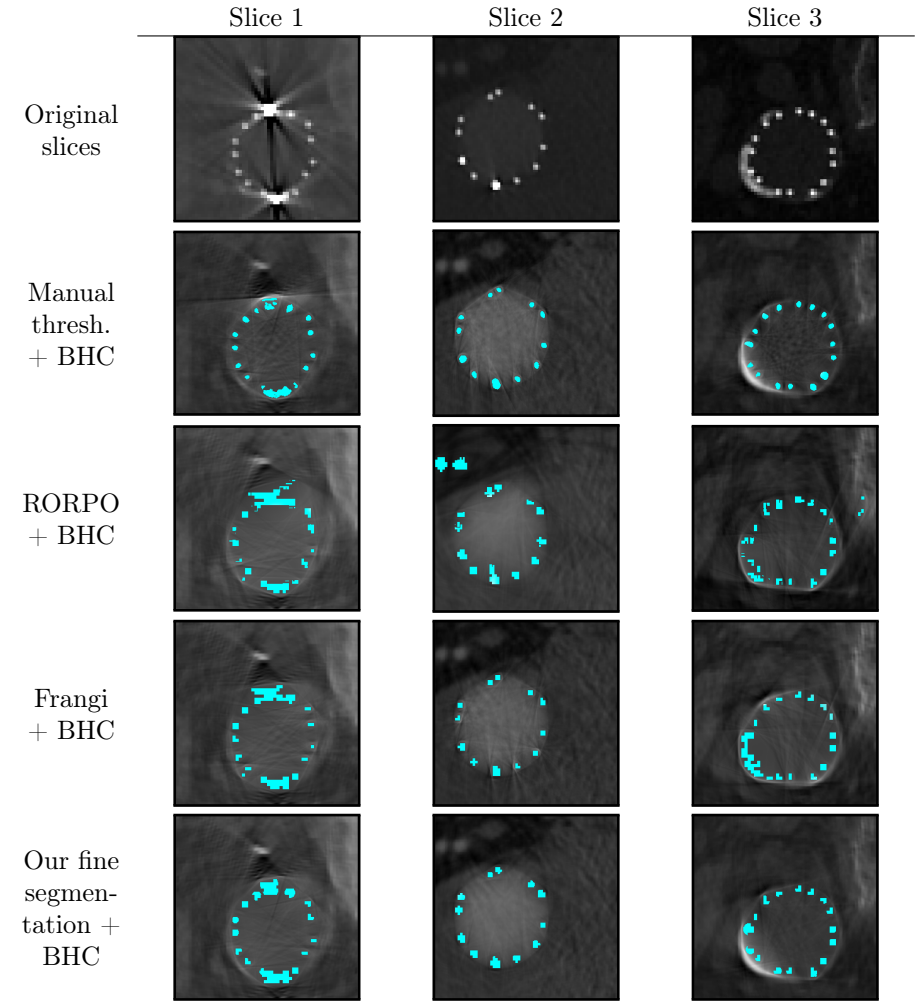


Figure 5.6.: MAR with the BHC algorithm when using different stent segmentation techniques. The last row, which is the one including our segmentation technique, shows restored images less affected by the artifacts. This can be seen especially in the center region of the stent. Artifacts may be hard to see on the original slices because of the image dynamic range. The scales used in the Frangi filter are $\{3, 6, 9\}$. The scales used in the RORPO filter are $\{100, 150, 200\}$.

		FN	FP
Case 1	P-IN	<0.01	0.16
	GPMF	0.03	0.03
Case 2	P-IN	0.10	0.02
	GPMF	0.07	0.02
Case 3	P-IN	0.03	0.11
	GPMF	0.02	0.02

(a) From Figure 5.7.

		FN	FP
Case 4	P-IN	0.03	0.03
	GPMF	0.04	0.01
Case 5	P-IN	<0.01	0.19
	GPMF	<0.01	0.08

(b) From Figure 5.8.

Table 5.1.: FN and FP rates computed in the blue areas for each model, for each case of Figures 5.7 and 5.8.

5.3. Unsupervised segmentation of the organic material and calcifications

5.3.1. Organic material segmentation with GPMFs

In this section we provide some more illustrations on real world images of segmentations using the GPMF model. We discuss the current performances of the model and explore its current limits. Recall the context of Section 2.5.4, the goal is to segment the organic biomaterial in mCTs of stented arteries corrupted by spatially correlated noise.

In this context, Figures 5.7 and 5.8 illustrate some more segmentations with the classical HMF model, the Potts-Independent Noise (P-IN) model, and the new GPMF model. It is notable that the overall performance of the GPMF model seems worst in Figure 5.8 than in Figure 5.7. Indeed, the latter segmentations are examples of the current limitations of the GPMF model that we associate with the noise stationarity assumption made in the model.

A closer examination of the data seems to reveal that a bad performance of the GPMF is mainly caused by a smoothing effect that entails a quite substancial loss of details in the segmentation. The stationarity assumption might explain the problem, indeed, the model might reflect poorly the reality in zones where there is no or very few spatial correlation.

We might argue that the most interesting zones are the ones directly around the stents. Indeed, an important biomechanical question that initiated this work is whether the stent is in contact with calcifications. Such kind of interactions are believed to be of crucial importance in many cases of the failure of the treatment such as stent fracture. Hopefully, at such locations of the images the GPMF model is performing well in all cases: see the FN/FP rates computed in Table 5.1.

Remark: It also seems that the most complex cases for the GPMF model are the ones where the artifacts are the strongest, or more spread on the image. However this last remark has not currently been explored since the literature seems to lack a good measure of the metallic artifacts on an image.

Remark: A comparison with the results we obtained on semi-real images in Section 2.5.2 corroborates the fact that the stationarity assumption in the GPMF model is the most limiting factor. Indeed, in the semi-real experiment, the noise stationarity assumption holds and we do not observe any of the problems that we observe here.

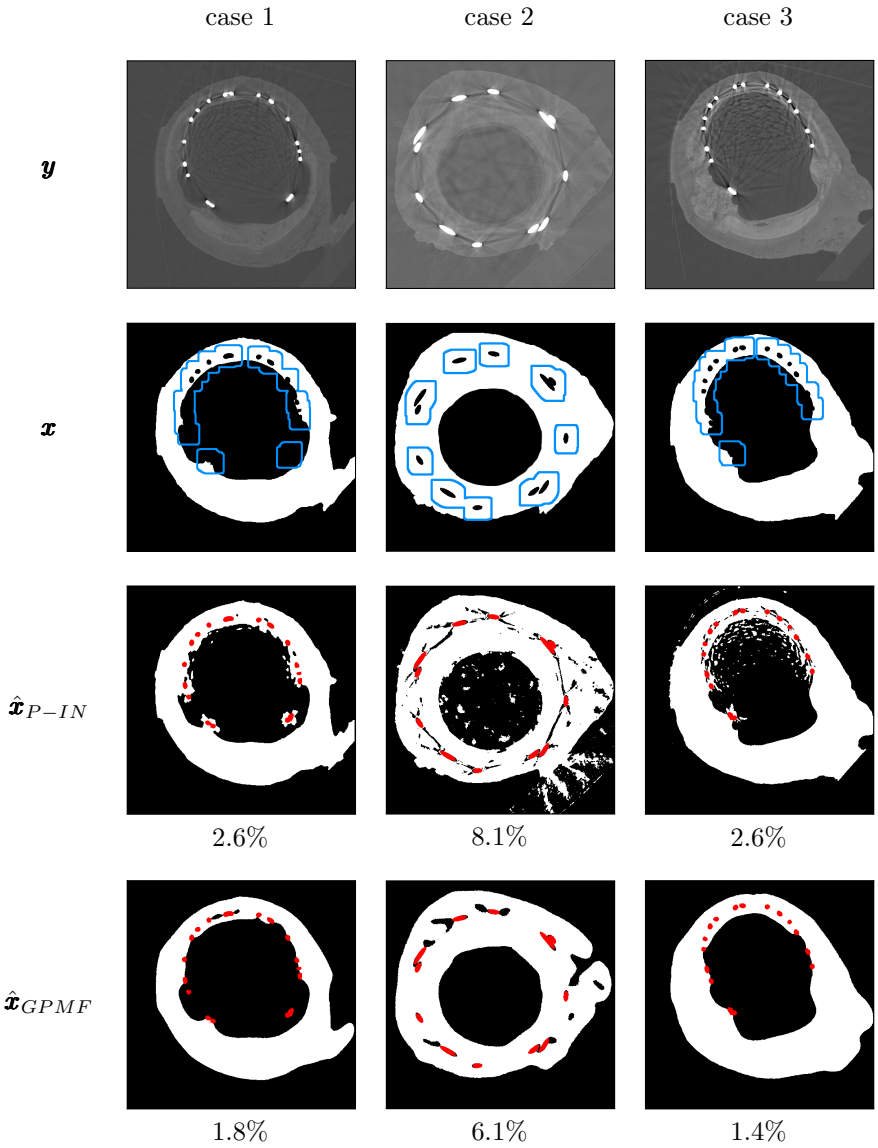


Figure 5.7.: P-IN and GPMF segmentations of organic material in mCT on more examples. The blue areas correspond to the areas of FN/FP computations. The indicated percentage represents the error rate in the segmentation computed over the whole image.

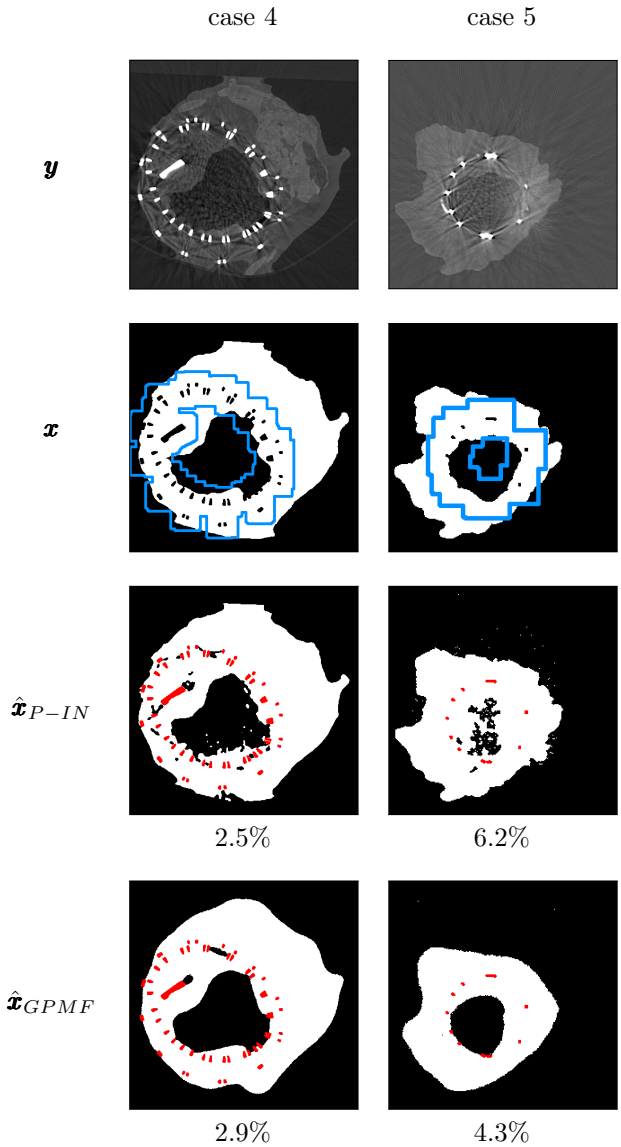


Figure 5.8.: P-IN and GPMF segmentations of organic material in mCT on more examples in the limiting cases for the GPMF model. The blue areas correspond to the areas of FN/FP computations. The indicated percentage represents the error rate in the segmentation computed over the whole image.

(a) Case A (b) Case B (c) Case C

Figure 5.9.: Ground truths for the experiment of unsupervised segmentation of calcifications.

5.3.2. Unsupervised calcification segmentation with STMTs

In this section, we work in the context of unsupervised image segmentation with STMTs. We want to compare, on real images, their relative performance with the HMF and HMT models. The goal of the experiment is to segment the atherosclerotic calcifications on a mCT image: the brightest parts of the mCT images (see Figure 5.9).

We also add additive independent Gaussian noise to form more diverse observations $\boldsymbol{Y} = \boldsymbol{y}$. Such a construction reflects situations where the mCT image would be of less good quality. Our objective is to evaluate, on real images, the capacity of STMTs to take the spatial context into account spatial context while offering fully deterministic inference computations. That is why the segmentation of atherosclerotic calcifications is the chosen objective: they are very homogeneous and rounded shapes which is expected to favor modelizations which can take into account the spatial context.

In this experiment we compare the HMF-IN model defined in Equations 3.29 and 3.30, the (quadtree) HMT-IN model, defined in Equation 1.24, and the (quadtree) STMT with Independent Noise, defined in Equations 3.5. The parameters are estimated unsupervisedly, with the IPET algorithm (Algorithm 3.2), as described in Section 2.3 for HMFs-IN and Section 3.3.2 for STMTs and HMTs-IN.

Figures 5.10, 5.11 and 5.12 depict such an experiment. The HMF-IN model seems to better segment the proposed shapes. However, for the higher noise levels, as stated in Chapter 3, we also encounter the issues arising from the stochasticity of the HMF-IN inference and parameter estimation procedures. Indeed, the results can differ a lot between two runs of the same experiment.

We also see that STMTs seem more prone to capture the spatial correlations than the HMT-IN model. As a result they offer better segmentations than the HMT-IN model as the noise level increases. At the lowest noise levels, STMTs and HMTs-IN perform similarly. When slightly better error rates are in favor of the HMT-IN model versus the STMT model, as in the first column of Figure 5.12, we find a contradiction with the results on synthetic data of Chapter 3. We might argue that this is due to the real world data we treat: the noise model might not be the best modelization for the image. The LLS estimation of the parameters might also be a limiting factor.

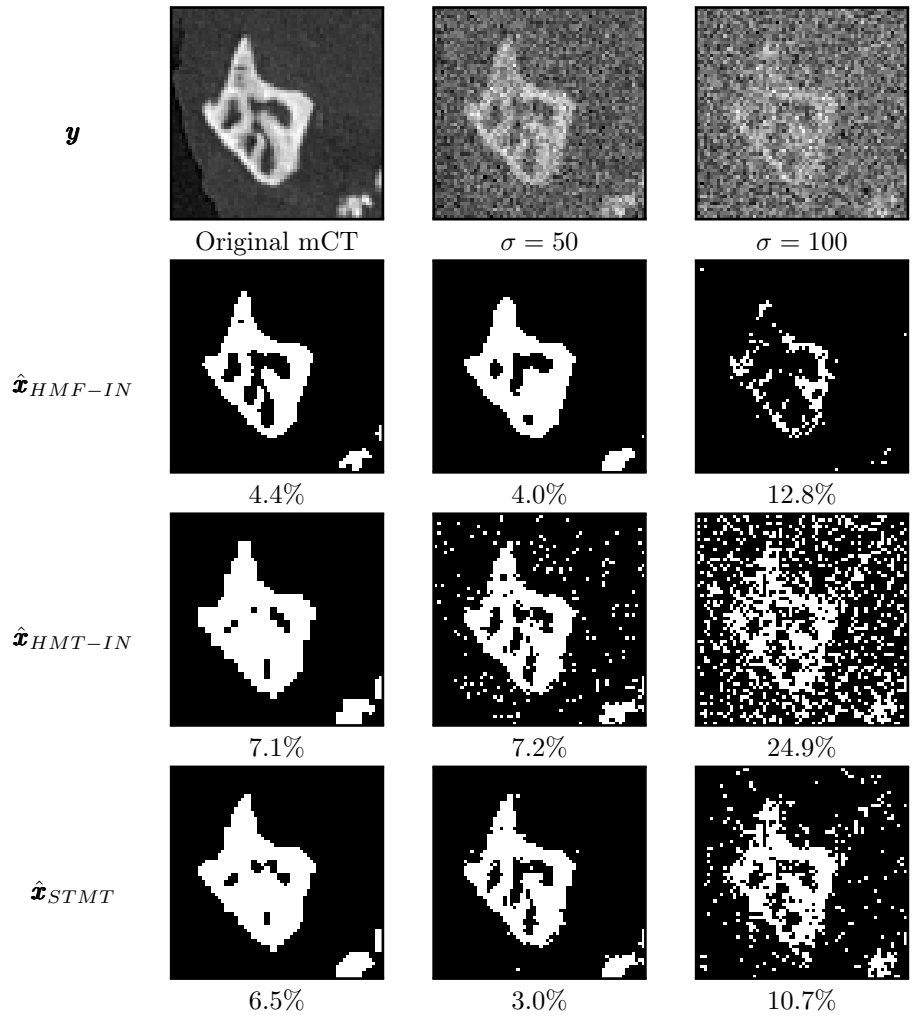


Figure 5.10.: Case A. Unsupervised segmentations of atherosclerotic calcifications. The percentages below the segmentations indicate the error rates with the ground truth. The ground truth is shown in Figure 5.9a.

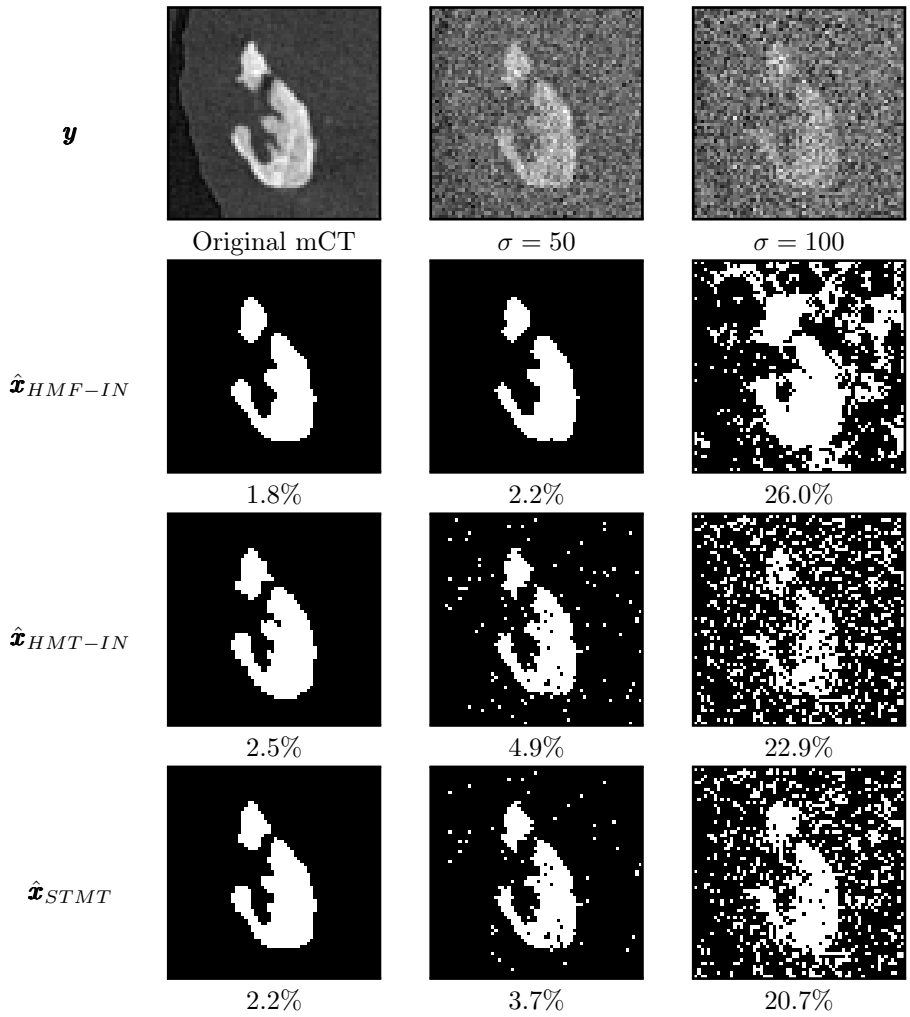


Figure 5.11.: Case B. Unsupervised segmentations of atherosclerotic calcifications. The percentages below the segmentations indicate the error rates with the ground truth. The ground truth is shown in Figure 5.9b.

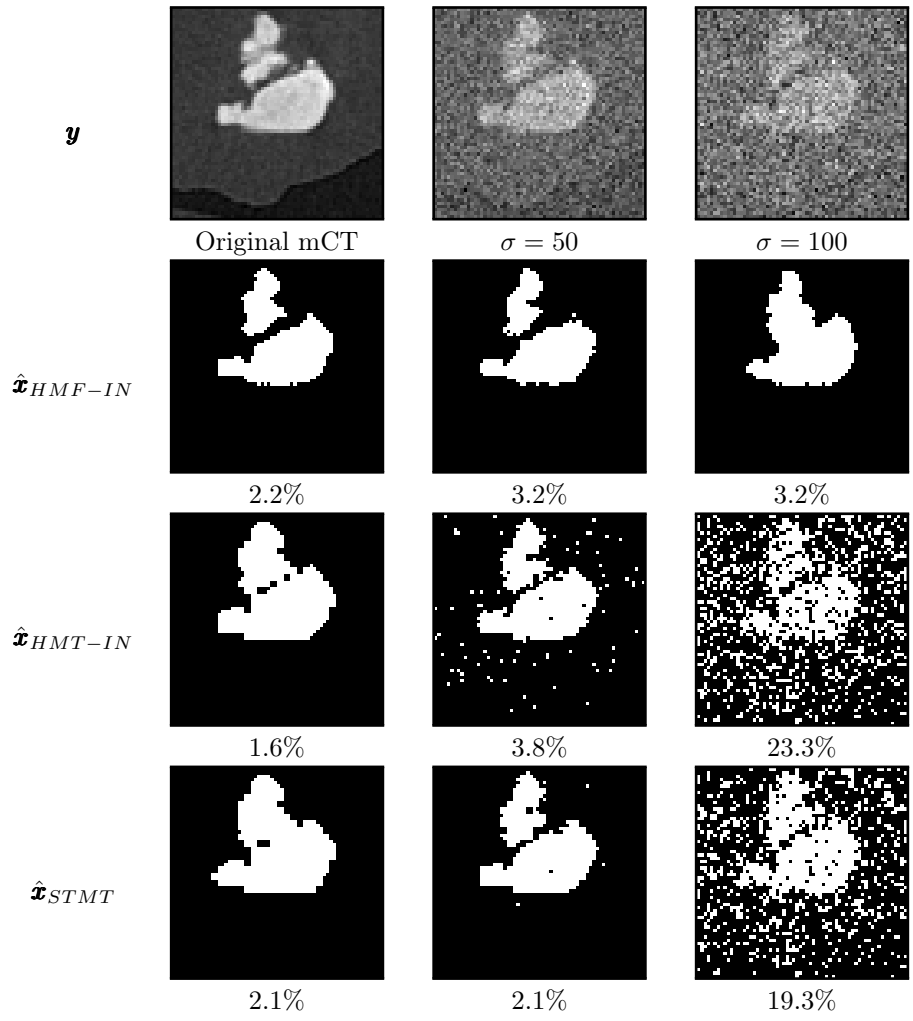


Figure 5.12.: Case C. Unsupervised segmentations of atherosclerotic calcifications. The percentages below the segmentations indicate the error rates with the ground truth. The ground truth is shown in Figure 5.9c.

5.4. Histological segmentation of microCT with Deep Learning

In this section we provide more context, details and results to the original project presented in Chapter 4.

5.4.1. Motivation

Our goal is to study peripheral obstructive artery diseases in femoropopliteal arteries. They are common affections of the blood vessels which are still not well understood (Yahagi et al., 2016) (Torii et al., 2019). On the one hand, microCT has been proven to be a valuable tool for the recognition of the different atherosclerotic components (Jinnouchi et al., 2018). On the other hand, histology, which is the gold standard to analyze and study the atherosclerotic process is a destructive technique and requires a skilled expert. Therefore, in this project we aim for an automatic histologic segmentation of microCT images based on artificial intelligence and, more precisely, Deep Learning (DL) which has become the state of the art approach for biomedical semantic image segmentation. With the algorithm we developed, the classic histologic process might be replaced by an automatic and non-invasive segmentation for a first analysis. Such an algorithm will save time and enable to process more data in a shorter time, since the histopathologists will be guided towards the most interesting biological materials. Eventually, the algorithm would help develop knowledge of the atherosclerotic process that would lead to better and more personalized treatments.

5.4.2. Protocol and database construction

In this section we give the main steps of the protocol that we set up for the construction of the database that was used as learning data for the DL algorithm. The steps are illustrated with images in Figure 5.14. In all the descriptions of this work, the colormap used for the histologic segmentation is given in Figure 5.15.

The steps of the proposed protocol are:

1. Retrieval of femoropopliteal arteries from amputated legs from patients who had undergone transfemoral amputations[3]. These explanted arteries were collected thanks to the Geprovas collaborative retrieval program. Figure 5.13 shows explanted arteries from the study.

2. The microCT 3D images of the arteries were acquired at the CVPath Institute, Inc, (Gaithersburg, MD, USA) using a Nikon X-Tek XT H 225ST[4].

[3] We started our study with six explanted arteries.
[4] https://www.nikonmetrology.com/en-gb/product/xt-h-255-st

3. Histology was then performed on the specimens as described in (Torii et al., 2019). Some histologic slices are depicted in Figure 5.14a.

4. Co-registration was subsequently performed manually between the microCT images and the histologic slices obtained during the two steps described above. This was performed with VGSTUDIOMAX 3.0[5] and ImageJ[6]. The result of this step consists in pairs of data: the microCT 2D image with its histologic ground truth. Some co-registered microCT images are illustrated in Figure 5.14b.

5. An expert then annotated the microCT images using the histologic ground truths in the GIMP software[7]. It was decided to segment the data into 11 classes. It has then been decided with the histopathologists that 6 classes were of interest to develop a first version of the algorithm. We now list the 6 classes, with the subclasses that possibly compose them:

   - *soft tissue* (ST): *soft tissue*, *formaldehyde*, *thrombus*, *fibrous plaque*.
   - *fatty tissue* (FT): *fatty tissue*, *lipid pool*.
   - *sheet calcification* (SC).
   - *nodular calcification* (NC).
   - *specimen holder* (SH).
   - *background* (Ba): *background*, *true lumen*.

   In the list above, we also specify the acronym that might be used to refer to the class. This step results in the creation of data pairs (microCT image and the corresponding class image) that are the data the DL algorithm will be trained on. Some annotated slices are illustrated in Figure 5.14c.

The subsequent steps of the project are described in Chapter 4. Those steps deal with the set-up and training of a Convolutional Neural Network (CNN) to achieve the automatic segmentation goal we described in Section 5.4.1. We now focus on the analysis of the resulting segmentations provided by the CNN.

Remark: The choice to reduce the number of classes from 11 to 6 has been made for various reasons. It is essentially a compromise between providing a segmentation meaningful from a histologic viewpoint but also to set up a solvable problem from a DL viewpoint. Indeed, this project represents exploratory work where data are rare, highly imbalanced and the some classes exhibit high similarity in images.

[5] https://www.volumegraphics.com/
[6] https://imagej.nih.gov/ij/
[7] https://www.gimp.org

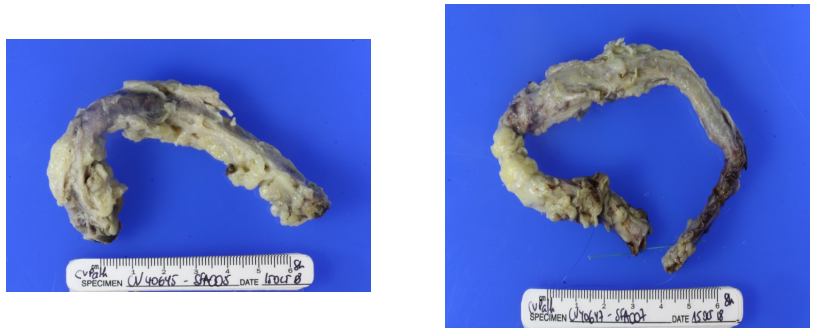


Figure 5.13.: Pictures of two explanted arteries after amputation, the first step of our protocol.

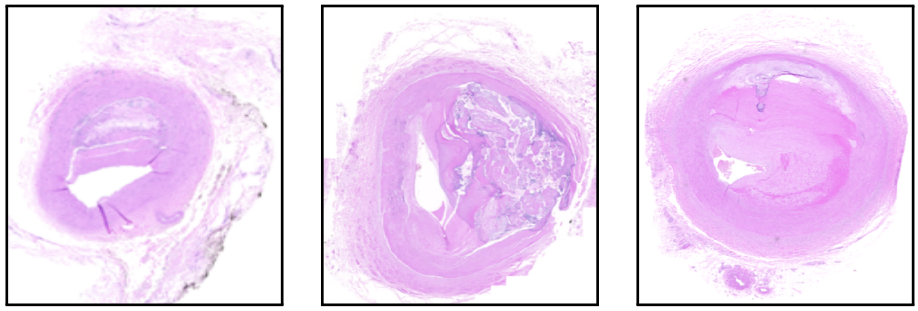
5.4.3. Results

2D segmentation on the test set

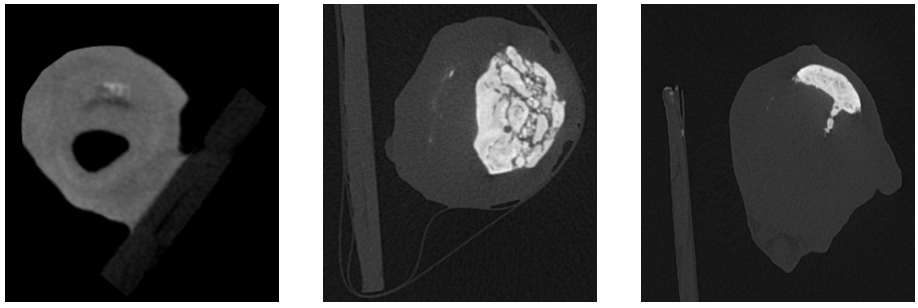
Table 4.1 references the Dice scores on the test set. These results indicate several degrees of difficulties in the automatic recognition of the classes. First of all, the *soft tissue*, *background* and *specimen holder* classes are well segmented (Dice scores > 0.86). Secondly, the CNN is capable of responding quite well to a crucial histologic analysis: the differenciation of the calcification classes, *sheet calcification* and *nodular calcification*, whose Dice scores are, respectively, 0.85 and 0.64. This ability is remarkable since the differenciation is almost impossible by the naked eye, even for a trained expert. Finally, we note that the *fatty tissue* class is the most complex to segment (Dice score of 0.41). This can be explained by the fact that this class is really similar in appearance to the *soft tissue* class. It is also under represented in the training data, which might be another cause for the high number of misclassifications.

To go further in the analysis, we give in Table 5.2 the multiclass confusion matrix associated with the segmentations of the CNN on the test set. We can then be more precise about the comments we made on the Dice scores. Looking at the SC and NC rows of the table we see that the main source of errors in the segmentation of the calcifications comes from a confusion between the two calcification types. Interestingly, by reading the FT row, we get that the reason for the bad FT score is that this class is most of the time misclassified as the ST class.

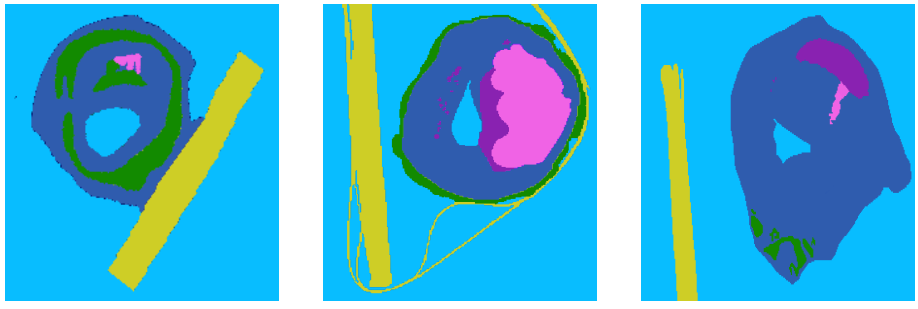
By considering the segmentation of each class as a binary *One versus the rest* classification problem we can compare the segmentation of each class using the Precision Recall (PR) curve and the Area Under Curve (AUC). A bigger AUC reveals a better classifier for a given class. PR curves and AUCs are given in Figure 5.16. This analysis exhibits the same hierarchy between classifiers as the Dice scores discussed above. Note that because of the high class imbalance the Receiver Operating Characteristic (ROC) curves are not exploitable, they are biased by the huge number of negatives from the Ba class and cannot be



(a) Three histologic slices resulting from Step 3.



(b) Three microCT images correlated with their histologic truth after Step 4.



(c) Three expert annotated microCT images obtained after Step 5.

Figure 5.14.: Illustrations of the different steps of the protocol on three examples.

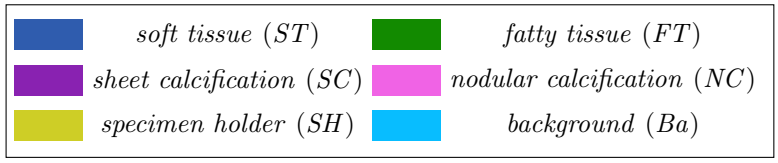


Figure 5.15.: Colormap for the histologic classes.

Truth \ Pred.	*ST*	*FT*	*SC*	*NC*	*SH*	*Ba*
ST	9.94	0.13	0.07	0.02	0.02	0.2
FT	0.50	0.31	0.002	0.0006	0.002	0.004
SC	0.06	0.007	0.78	0.11	0.0009	0.0005
NC	0.05	0	0.17	0.25	0.001	0.002
SH	0.15	0.34	0.008	0.003	2.73	0.32
Ba	0.17	0.06	0.003	0.001	0.19	83.4

Table 5.2.: Multiclass confusion matrix of the CNN results of the test set. Each cell has been normalized by the total number of pixels in the test set so that the sum of all the cells gives 1. The numbers in the cells must be multiplied by $\times 10^{-2}$.

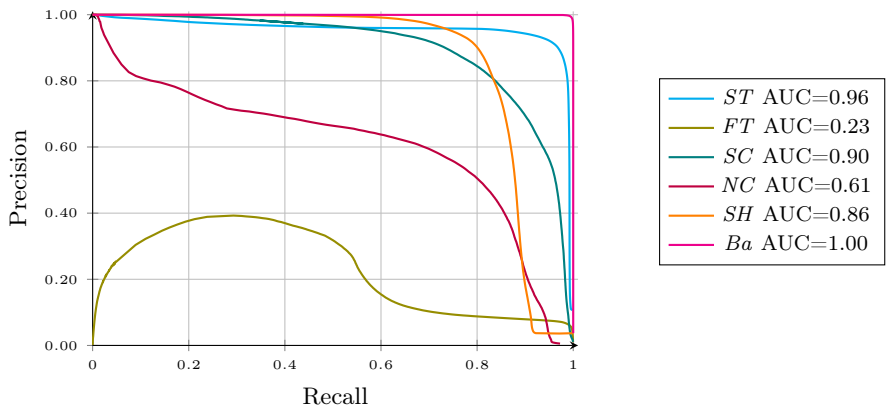


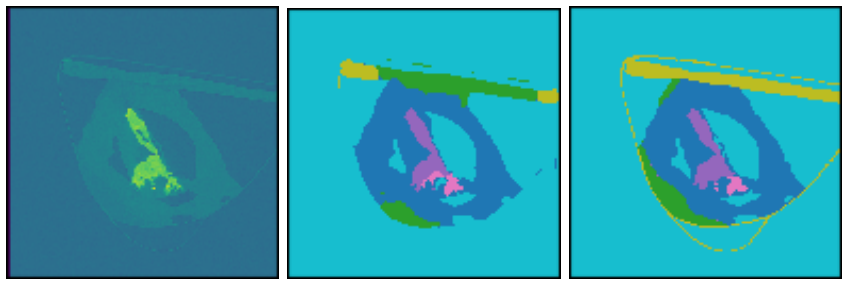
Figure 5.16.: Precision-Recall curves and their AUCs obtained by the CNN for each class on the test set.

properly compared. See (Murphy, 2012) for more details on PR and ROC curves.

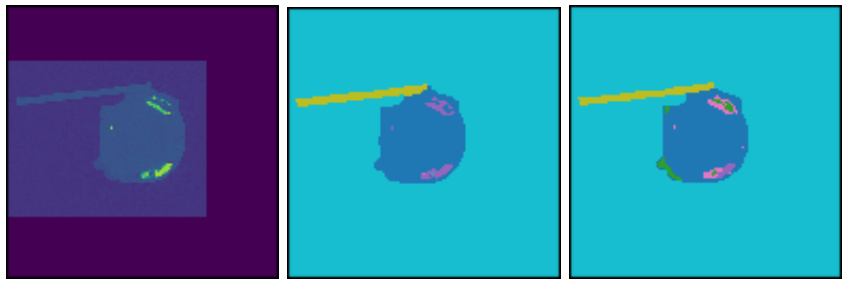
To conclude this section on the 2D segmentations, Figure 5.17 depicts some segmented slices by the CNN that illustrate the detailed discussion we give above. The illustrations reflect the complexity to correctly classify pixels from the *FT* class and the confusion that sometimes arises when classifying pixels into one of the calcification classes.

3D segmentation on test femoropopliteal arteries

We now provide in Figure 5.18, some examples of the complete segmentation of an artery by the CNN after fcCRF and MF VI post-processing (see Chapter 4). Such results are directly understandable by the histopathologists who can then analyze the explant at a histologic level with a microCT acquisition only. The expert can then decide how to go further in their analysis.

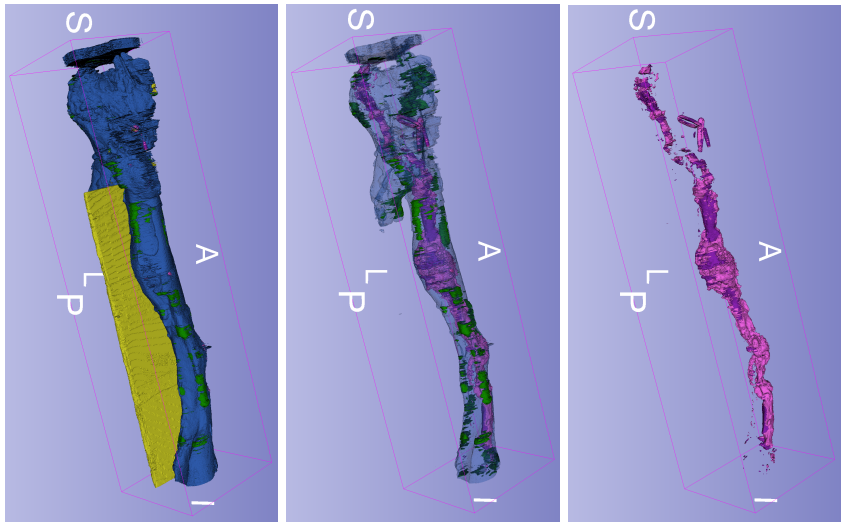


(a) MicroCT slice / CNN predictions / Ground truth

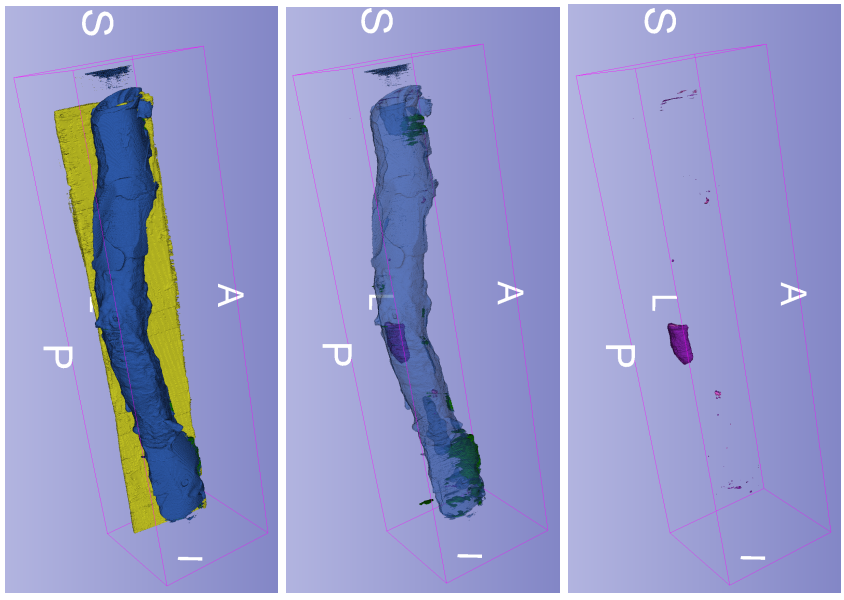


(b) MicroCT slice / CNN predictions / Ground truth

Figure 5.17.: Examples of 2D predictions by the CNN on samples from the test set. Colormap is given in Figure 5.15.



(a) All classes (except Ba) / $ST + FT + SC + NC$ / $SC + NC$



(b) All classes (except Ba) / $ST + FT + SC + NC$ / $SC + NC$

Figure 5.18.: 3D histological reconstructions of two arteries by combining the results of the CNN with a post-processing based on fcCRFs. The colormap is given in Figure 5.15.

5.5. Conclusion

In this chapter we offered an applicative perspective on the probabilistic models described along the thesis. Our goal was to show to what extent they could be the foundation of new tools used in the field of vascular surgery to develop and enhance the medical research as well as the clinical workflow. The results in this chapter are founded on CT, mCT and histological images. This highlights the importance of the imagery and the processing of the produced images in vascular surgery. The key role of images is expected to grow with time.

Conclusions and Perspectives

Gaussian Pairwise Markov Fields

Family of the probabilistic model studied:
Hidden Markov Fields
Model, innovation:
Gaussian Pairwise Markov Fields
Issue that has been addressed:
Unsupervised segmentation of images corrupted with spatially correlated noise
Perspective that have been opened in medical research:
Fine automated analysis of the biological environment of explanted stents

---

The GPMF model introduced in Chapter 2 is a way to confidently segment organic biomaterial in a scan corrupted by long-spatially correlated noise. We showed that GPMFs handle the stent artifacts and can offer the best compromise to limit the number of False Positive and False Negative pixels, as compared to other classical approach of segmentation.

We can give two perspectives based on this work. The first would be to introduce a spatial non-stationarity for the range and the variance, *i.e.*, to allow these parameters to vary according to the location on the images. This would indeed better reflect the correlated noise observed on the images since its strenght is not the same everywhere. One could then expect further improvements in the segmentations by, for example, reducing the undesired smoothing effect that can happen. One way to study this issue is to complexify the correlation function so that it integrates non-stationarities (Fouedjio et al., 2015) (Kleiber, 2016) (Nychka et al., 2018). Another approach might consider building a probabilistic graphical model which can switch between submodels with different parameters (Courbot, Monfrini, et al., 2018) (Vacar and Giovannelli, 2019). Another perspective is based on the construction of an annotated dataset (described in Section 5.4.2) of the images we want to segment[8]. This would enable us to construct a discriminative version of the GPMF model related to other models of CRFs recently developed (Radosavljevic et al., 2010) (Krähenbühl and Koltun, 2011) (Vemulapalli et al., 2016) (Petrović et al., 2019). Such a discriminative formulation of the model would potentially improve the resulting segmentations (Ng and Michael I Jordan, 2002) and offer a computational advantage for the inference process.

[8]This dataset was not available at the time of the work on the GPMF model.

Stent segmentation in X-ray images

Family of the probabilistic model studied:
Hidden Markov Chains
Model, innovation:
Hidden Markov Chain with specific noise model and path
Issue that has been addressed:
Unsupervised segmentation of deformed elements without prior information
Perspective that have been opened in medical research:
Automated analysis of stent deformations and fractures

The fine segmentation of stents in CT and mCT images developed in Chapter 5 faces the problem of artifacts which are a common type of corruption in image processing of biomaterials. We proposed a statistical model dedicated to handling these artifacts and restore the original smooth and tubular stent shape.

On the signal processing aspect, the model we developed could be improved with a noise model which explicitly takes into account the correlated noise such as Pairwise or Triplet Markov models (Gorynin, Gangloff, et al., 2018). One could also algorithmically improve the Markov Chain path, which is currently based on a slice-by-slice processing, by extending the path computation algorithm so that it works in full 3D. The problem of stent segmentation could also be approached in a supervised fashion if a database of annotated images were developed. A first Deep Learning algorithm for stent segmentation has been proposed recently for X-ray images in (Breininger et al., 2018).

On the medical aspect, being able to confidently segment stents in artifacts paves the way to further automated stent analyses. The main challenge would consist in using the extracted 3D stent as a basis for numerical biomechanical analyses carried to understand the cause of the stent failures and improve the treatments (Langs et al., 2011) (Koenrades et al., 2019). Other types of automated processing over the 3D segmented stent could consist in an automated stent model recognition or stent fracture localization. Such algorithms would enhance the Geprovas workflow of explant analysis.

Spatial Triplet Markov Tree

Family of the probabilistic model studied:
Hidden Markov Trees
Model, innovation:
Spatial Triplet Markov Tree
Issues that has been addressed:
Enhancing the correlations in Hidden Markov Trees in a local and spatial manner, increasing the knowledge on the relation between the Markov tree and Markov field models
Perspective that have been opened in medical research:
More efficient treatments of large medical images

The STMT model from Chapter 3 illustrates the interest of auxiliary random variables in probabilistic models. We explored to what extent this model could be a close alternative to MRF models with the capability of performing exact inference. This could be an asset over an approximate inference algorithm when dealing with large data such as medical images. For a better understanding of STMTs, we showed its closeness to the SBN model in the context of auxiliary random variable VI.

Further research on the proximity of STMTs and SBNs might follow the direction of deriving theoretical results quantifying the relations between the two models. One might also consider extending our auxiliary random variable VI algorithm to quadtrees to process images. Such an algorithm is expected to be very efficient based on the preliminary results that we obtained in the case of dyadic trees.

Histologic segmentation of atherosclerotic arteries with Deep Learning

Family of the probabilistic model studied:
Conditional Random Fields
Model, innovation:
Improved inference in Conditional Markov Random Fields
Issue that has been addressed:
Improving the inference in a probabilistic model for post-processing the bidimensional segmentations of a Convolutional Neural Network to reconstruct a tridimensional segmentation
Perspective that have been opened in medical research:
A first histological analysis of the vascular segment using the X-ray image only

---

Thanks to the U-Net approach and the fcCRF post-processing described in Chapter 4, we reconstructed tridimensional arteries with annotated histologic elements. This study confirms the important role that microCT images could have in vascular research.

The main perspectives in this work is first to increase the score in the segmentation of each of the classes (particularly, in the current implementation of our model, the *fatty tissue* class) but also to integrate more classes and offer a more detailed histologic segmentation. As an example, it would be interesting to treat arteries containing a stent. The neural network approach, combined with a pre- or post-processing approach, could additionally segment the stent in the images as well as to handle the artifacts such images would contain. To this end, one could consider exploring recent advances in CNN, such as more recent versions of the U-Net network (Alom et al., 2018) (Oktay, Schlemper, et al., 2018). One could also study the interest of including shape priors within the CNN (Nosrati and Hamarneh, 2016) (Oktay, Ferrante, et al., 2017) (Ravishankar et al., 2017). Shape priors could be elaborated jointly with the experts collaborating on the project.

Appendix A.

Main algorithms in probabilistic modeling

Algorithm A.1: Gibbs sampler (S. Geman and D. Geman, 1984) to sample from an UGM using the *full conditionals*.

Data: $p(x_s|\boldsymbol{x}_{\mathcal{N}_s}), \forall s \in \mathcal{S}$, the full conditional equations of an UGM distribution $p(\boldsymbol{x})$.
Result: Series of samples $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^n$.
$n \leftarrow 1$
Initialize $\boldsymbol{x}^0$
while *convergence is not attained* **do**
  /* Initialize the new samples */
  $\boldsymbol{x}^n \leftarrow \boldsymbol{x}^{n-1}$
  /* Update at each site */
  for $s \in \mathcal{S}$ **do**
    $x_s^n \sim p(x_s^n|\boldsymbol{x}_{\mathcal{N}_s}^n)$
  end
  $n \leftarrow n + 1$
end

Algorithm A.2: Forward Backward (FB) algorithm, rescaled version from (Devijver, 1985). The original unscaled algorithm from (Baum, Petrie, et al., 1970) is prone to underflows; it lacks the rescaling by the term κ_s. Note that in the original unscaled version of FB we have $\alpha(x_s) = p(x_s, y_1, \ldots, y_s)$ and $\beta(x_s) = p(y_{s+1}, \ldots, y_N | x_s)$, $\forall s \in \mathcal{S}$.

Data: $\boldsymbol{y}$, a realization of the observed process,
$p(x_s|x_{s^-}), \forall s \in \bar{\mathcal{S}}$, transitions of the hidden process,
$p(x_s|x_{s^-}), \forall s \in \bar{\mathcal{S}}$,
$p(x_r)$, distribution at the root vertice.
Result: $p(x_s|\boldsymbol{y}), \forall s \in \mathcal{S}$, the posterior marginals.
/* Compute the rescaled *forward* probabilities $\alpha^*(x_s) = p(x_s|\mathbf{y}), \forall s \in \mathcal{S}$, the recursion is defined by */

$$\begin{aligned} \alpha^*(x_1) &= \frac{p(x_1)p(y_1|x_1)}{\sum_{x_1'} p(x_1')p(y_1|x_1')}, \\ \alpha^*(x_{s+1}) &= \frac{1}{\kappa_{s+1}} \sum_{x_s \in \Omega} \alpha^*(x_s)p(x_{s+1}|x_s)p(y_{s+1}|x_{s+1}), \\ &\forall s \in \{2, \ldots, N\}. \end{aligned} \tag{A.1}$$

/* Compute the rescaled *backward* probabilities $\beta^*(x_s) = \frac{p(y_{s+1},\ldots,y_N|x_s)}{p(x_{s+1},\ldots,x_N|y_1,\ldots,y_s)}$. The recursion is defined by */

$$\begin{aligned} \beta^*(x_1) &= 1, \\ \beta^*(x_s) &= \frac{1}{\kappa_{s+1}} \sum_{x_{s+1} \in \Omega} \beta^*(x_{s+1})p(x_{s+1}|x_s)p(y_{s+1}|x_{s+1}), \\ &\forall s \in \{1, \ldots, N-1\}. \end{aligned} \tag{A.2}$$

/* In Equations A.1 and A.2, the rescaling factor is */

$$\kappa_{s+1} = \sum_{x_{s+1}} \sum_{x_s \in \Omega} \alpha^*(x_s)p(x_{s+1}|x_s)p(y_{s+1}|x_{s+1}). \tag{A.3}$$

/* Compute the posterior marginals */

$$p(x_s|\boldsymbol{y}) = \alpha^*(x_s)\beta^*(x_s), \forall s \in \mathcal{S}. \tag{A.4}$$

Algorithm A.3: Simulated Annealing via serial Gibbs sampling (S. Geman and D. Geman, 1984). This algorithm allows only changes in the states towards a state of lower energy. We then easily get stuck in local minima. A practical implementation usually relies on several runs of the algorithm starting from different initial configurations. In Equation 1.4, we make the temperature parameter explicit so that: $p_T(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z}\exp(-E(\boldsymbol{x}|\boldsymbol{y})/T)$. The parameter schedule may vary (Delahaye et al., 2019).

Data: $\boldsymbol{x}^0$, an initial configuration of the Gibbs sampler,
T^0, an initial temperature,
$\boldsymbol{y}$, the observations.
Result: $\hat{\boldsymbol{x}}^{MAP} = \text{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$.
$t \leftarrow 1$
while *convergence is not attained* **do**
/* Sample a realization at temperature T^t */

$$\boldsymbol{x}^t \sim p_{T^t}(\boldsymbol{x}|\boldsymbol{y}). \tag{A.5}$$

/* Update the temperature according to the schedule. C is a constant that may be determined by trial and errors */

$$T^t = \frac{C}{\log(1+t)}. \tag{A.6}$$

end

Algorithm A.4: Expectation Maximization (Dempster et al., 1977).

Data: $\boldsymbol{\theta}^0$, an initial set of parameters,
$\boldsymbol{y}$, the observations.
Result: $\hat{\boldsymbol{\theta}}$, the set of estimated parameters.
$t \leftarrow 1$
while *convergence is not attained* **do**
/* E-step
Define $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1})$ by */

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1}) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}^{t-1})}[\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})]. \tag{A.7}$$

/* M-step
Estimate the new set of parameters */

$$\boldsymbol{\theta}^t = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1}). \tag{A.8}$$

$t \leftarrow t+1$
end

Algorithm A.5: Stochastic Expectation Maximization (Celeux, 1985).

Data: $\boldsymbol{\theta}^0$, an initial set of parameters,
$\boldsymbol{y}$, the observations.
Result: $\hat{\boldsymbol{\theta}}$, the set of estimated parameters.
$t \leftarrow 1$
while *convergence is not attained* **do**
  /* Stochastic E-step */
  Compute $p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}^{t-1})$.
  Draw samples to complete the data: $\boldsymbol{x}^t \sim p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}^{t-1})$.
  /* M-step */
  Maximum Likelihood estimation $\boldsymbol{\theta}^t$ on the completed data $(\boldsymbol{x}^t, \boldsymbol{y})$.
  $t \leftarrow t+1$
end

Algorithm A.6: Marroquin algorithm for the MPM computation (Marroquin et al., 1987) using the Gibbs sampler.

Data: $p(x_s|\boldsymbol{x}_{\mathcal{N}_s}), \forall s \in \mathcal{S}$ the full conditional equations of an UGM distribution $p(\boldsymbol{x})$,
K, the number of Gibbs samples.
Result: $\hat{\boldsymbol{x}}$ the MPM estimator of $p(\boldsymbol{x})$.
while *convergence is not attained* **do**
  for $k \in \{1, \ldots, K\}$ **do**
    $\boldsymbol{x}^k \leftarrow$ last sample from Gibbs sampling with $p(\boldsymbol{x})$
  end
  Frequency estimator of $\hat{p}(x_s = \omega), \forall s \in \mathcal{S}, \forall \omega \in \Omega$ using the realizations $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^K)$
  $\hat{x}_s^{MPM} = \text{argmax}_{\omega \in \Omega} \hat{p}(x_s = \omega)$
end

Appendix B.

Complements on GMRFs

B.1. Stationary GRFs and GMRFs

In this section we introduce GRFs and GMRFs, the latter is derived from GRFs with an important assumption that is made about the covariance matrix.

First, we need to define the notion of *positive-definite function*. A positive-definite function of a real variable is a complex-valued function $f : \mathbb{R} \to \mathbb{C}$ such that for any numbers $y_1, \dots, y_n$, the $n \times n$ matrix $A = (a_{i,j})_{i,j=1}^{n}, a_{i,j} = f(y_i - y_j)$ is positive semi-definite. In our application, $f : \mathbb{R} \to \mathbb{R}$ and the resulting A will be a real SPD.

A GRF or GMRF is called stationary if the following conditions hold:

- the mean vector is constant (does not depend on the vertice $s, \forall s \in \mathcal{S}$).
- the covariance between two random variables in the field only depends on the norm ($\ell 1$, Euclidian...) between those two points, such that, for some positive-definite function f:

$$\Sigma_{(i,j),(i',j')} = Cov(x_{i,j}, x_{i',j'}) = f(\|y_{i,j} - y_{i',j'}\|), \tag{B.1}$$

In Equation B.1, f is also called the *covariance function*. In the case of stationary GRFs or GMRFs, the covariance function is directly linked to a *correlation function* c such that:

$$Cov(a, b) = \sigma^2 c(a, b). \tag{B.2}$$

where $\sigma^2 \in \mathbb{R}_+$ is a constant variance (imposed for stationary GRFs or GMRFs (Abrahamsem, 1997)). Note that for unitary variance, the correlation function and the covariance function are equal.

By definition, we require the covariance function to be a positive-definite function, then Equation B.2 shows that a correlation function c is also a positive-definite function. Let $\|h\|$ be a distance function (Euclidean or $\ell 1$ for example[1]). Some classical correlation functions are, $\forall (s, s') \in \mathcal{S}^2$:

- the exponential correlation function:

$$c\,(s, s'; r) = \exp\left(-\frac{\|s - s'\|}{r}\right), \forall r > 0, \sigma^2 \geq 0. \tag{B.3}$$

[1] Potentially taken on the torus, see Section B.2.3.

- the Gaussian correlation function:

$$c(s, s'; r) = \exp\left(-2\left(\frac{\|s - s'\|}{r}\right)^2\right), \forall r > 0, \sigma^2 \geq 0. \tag{B.4}$$

The parameter r is called the *correlation length*, *decay* or *range*. Note that the Gaussian and exponential correlation functions can both be derived from the Matérn correlation function (Abrahamsem, 1997), a more general correlation function. Moreover, these correlation functions are given under their *isotropic* form. An example of *anisotropic* Gaussian or exponential correlation function would consist in setting different correlation lengths according to the direction. This is considered for example in (Dietrich and Newsam, 1997).

In this thesis, to perform segmentation with the GPMF model, we consider isotropic stationary covariance functions but with non-stationary mean vector. This is a case of non-stationarity much easier to handle than a non-stationary covariance matrix (Fuglstad et al., 2015).

Remark: The modeling of non-stationary GRFs and GMRFs is an active research area out of scope of this thesis. One can see (Gelfand et al., 2010)[Chapter 10] or (Risser, 2016) for an overview.

B.2. Spectral methods for GMRFs

This section is a condensed view of the development made in (Rue and Held, 2005) to establish computationally efficient formulas for GMRF manipulation, which is essential for the tractability of the algorithms developed in Chapter 2. These formulas are based on the Fourier transform and on a *periodic boundary* assumption, also called *torus* assumption. The latter assumption leads to a special form of the covariance matrix. We start by reviewing key elements of algebra over these matrices.

B.2.1. Circulant matrices

We here introduce *circulant matrices* and their related spectral properties. A $n \times n$ matrix C is circulant if, and only if, it has the form:

$$C = \begin{pmatrix} c_0 & c_1 & c_2 & \dots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \dots & c_{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ c_1 & c_2 & c_3 & \dots & c_0 \end{pmatrix}, \tag{B.5}$$

for some vector $\boldsymbol{c} = (c_0, c_1, \dots, c_{n-1})^T$. $\boldsymbol{c}$ is called the basis of C. A circulant matrix is fully specified by its basis. A $Nn \times Nn$ matrix C is block-circulant

with $N \times N$ blocks, if, and only if, it can be written as:

$$C = \begin{pmatrix} C_0 & C_1 & C_2 & \dots & C_{N-1} \\ C_{N-1} & C_0 & C_1 & \dots & C_{N-2} \\ \vdots & \vdots & \vdots & & \vdots \\ C_1 & C_2 & C_3 & \dots & C_0 \end{pmatrix}, \tag{B.6}$$

where C_i is a circulant $n \times n$ matrix with base $\boldsymbol{c}_i$. The base of C is the $n \times N$ matrix

$$C^b = (\boldsymbol{c}_0, \boldsymbol{c}_1, \dots, \boldsymbol{c}_{N-1}). \tag{B.7}$$

A block-circulant matrix is fully specified by its base or one block column or one block row.

B.2.2. Spectral properties

The following properties linked with the Fourier transform are available for circulant and block-circulant matrices. We will recall the ones for block-circulant matrices, hence we will use the bidimensional Fourier transform.

Let C be a block-circulant matrix with base C^b, then we have that C^{-1} has base $(C^{-1})^b$ with:

$$(C^{-1})^b = \text{IDFT2}(\text{DFT2}(C^b) \bullet (-1)). \tag{B.8}$$

Let C and D be two block-circulant matrices with bases C^b and D^b, then we have

$$CD = \text{IDFT2}(\text{DFT2}(C^b) \odot \text{DFT2}(D^b)). \tag{B.9}$$

Let C be a block-circulant matrix with base C^b, then

$$\Lambda = \text{DFT2}(C^b) \tag{B.10}$$

is the matrix of the eigenvalues of C.

B.2.3. Torus assumption for GRFs and GMRFs

The *torus assumption*, also known as *periodic boundary assumption* supposes that the boundaries of the grid (formed by the image) are joined together as on a torus. Figure B.1 illustrates the process. The main consequence[2] of the torus assumption for a stationary GRF/GMRF is that its covariance matrix is then block-circulant and each block is circulant itself. As a consequence, all the spectral properties for circulant matrices presented in Section B.2.2 can be used. The most used property in our study is the computationally efficient matrix inversion given in Equation B.8.

On toruses, one has to modify the distance used. For example, on a two dimensional torus, the Euclidean distance becomes, for $\boldsymbol{a} = (x_1, y_1)$ and $\boldsymbol{b} =$

[2] Under an appropriate ordering of the indices, see Section B.2.4.

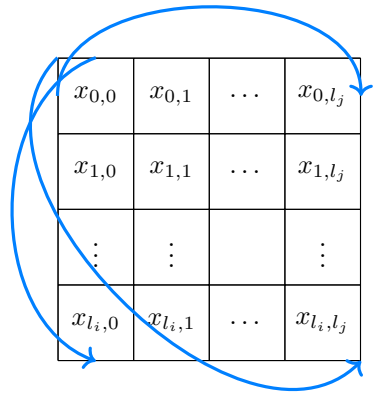


Figure B.1.: The torus assumption on an image: the three new neighbors of the top right corner are indicated by an arrow. The torus can be thought as a wrapping of a 2D image.

(x_2, y_2) on the surface of this torus formed by wrapping a grid of dimension $l_i \times l_j$:

$$d_{torus}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\min\left(|x_1 - x_2|, l_i - |x_1 - x_2|\right)^2 + \min\left(|y_1 - y_2|, l_j - |y_1 - y_2|\right)^2}. \tag{B.11}$$

All the MRF models of Chapter 2 are developed under the torus assumption.

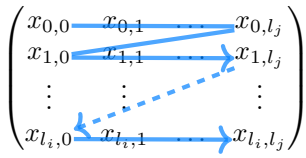
Remark: Another approach to using the spectral properties of circulant matrices without making the torus assumption is via *circulant embedding*. The method is based on the observation that, in the example of the 2D case, the covariance matrix of the GMRF will be block-Toeplitz with Toeplitz blocks[3]. This matrix is then *embedded* in a block-circulant with circulant block matrices and the spectral methods can be used. A presentation of this method in the 1D case is available in (Dietrich and Newsam, 1997) where the method was introduced and in (Kroese and Botev, 2013) for the 2D case.

B.2.4. Indices ordering

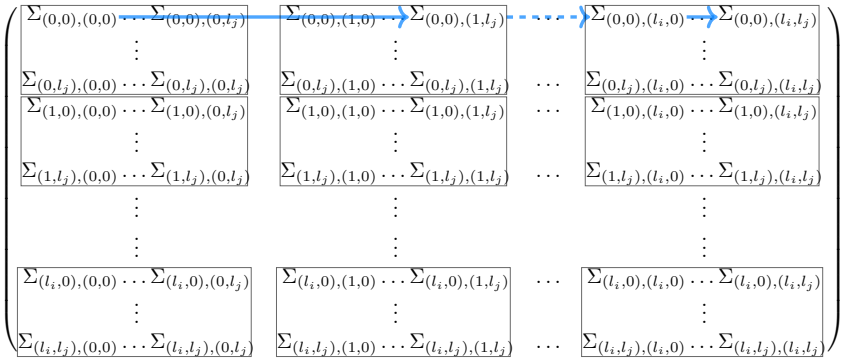
As noted before, for a Toeplitz or circulant matrix structure to appear for the covariance matrix, one needs to use an appropriate indexing of the matrix.

Focusing on covariance matrices of two dimensional GRF/GMRF, an entry in the covariance matrix will be stored in row major order, the mathematical

[3] Under an appropriate ordering of the indices, see Section B.2.4.



(a) The image representation with a matrix



(b) The associated covariance matrix

Figure B.2.: Building the covariance matrix of an image: the blue arrow highlights the order in which we will consider the neighbors of $x_{0,0}$ in the image and their order of appearance in the covariance matrix.

indexing relation is then, for a $l_i \times l_j$ two dimensional grid:

$$Cov(x_{i,j}, x_{i',j'}) = \Sigma_{(i,j),(i',j')}, \forall (i,i') \in \{0,\dots,l_i\}^2, \forall (j,j') \in \{0,\dots,l_j\}^2. \tag{B.12}$$

B.3. Application to GMRF simulation

In this section we illustrate GMRFs by drawing samples from the models in the 2D case using the Fourier based formulas of Section B.2.2.

We start by constructing a valid covariance matrix for the field. This can be done by specifying the variance of the field, choosing a correlation function c, parametrized by a range r, and following Algorithm B.1. In Algorithm B.2, we give the algorithm that can be followed to draw samples from the zero-mean Gaussian random field with base precision matrix $\boldsymbol{q}$.

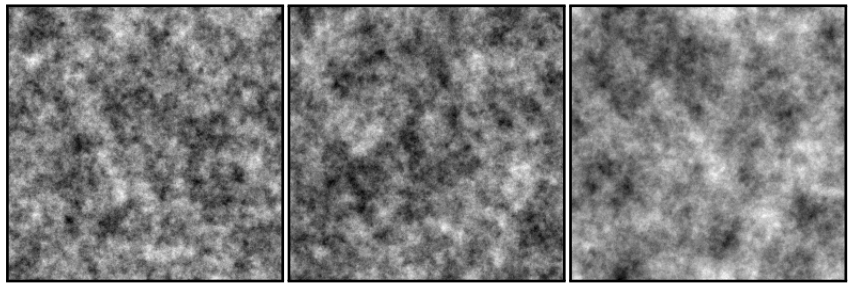
As an illustration, Figure B.3 depicts samples from different stationary GMRFs. For complete introductions to GRF simulation, one can read (Pichot, 2016), (Brown et al., 2019) or (Kroese and Botev, 2013).

Algorithm B.1: Construction of $\boldsymbol{q}$

Result: $\boldsymbol{q}$, the base of the precision matrix for the GMRF.
for $i = \{1 \dots l_i\}$ **do**
  for $j = \{1 \dots l_j\}$ **do**
    $\boldsymbol{q}_{i,j} \leftarrow \sigma c\left(d_{torus}((0,0),(i,j)), r\right)$
  end
end

Algorithm B.2: Sampling a GMRF with via the Fourier transform properties

Data: $\boldsymbol{q}$, the base of the precision matrix of the model.
Result: $\boldsymbol{x}$, a sample of the GMRF.
Sample $\boldsymbol{z}$, a 0-mean complex field with independent elements distributed according to: $\Re(z_i) \sim \mathcal{N}(0,1), \Im(z_i) \sim \mathcal{N}(0,1)$
/* Matrix of the eigenvalues */
$\Lambda \leftarrow \text{DFT2}(\boldsymbol{q})$
/* Perform the sampling */
$\boldsymbol{x} \leftarrow \Re\left(\text{DFT2}\left(\left(\Lambda \bullet -\frac{1}{2}\right) \odot \boldsymbol{z}\right)\right)$



(a) $\sigma = 0.5$, $r = 9$ (b) $\sigma = 5$, $r = 9$ (c) $\sigma = 0.5$, $r = 18$

Figure B.3.: Samples from stationary GMRFs with exponential correlation function and zero mean.

Appendix C.

Complements on GPMF

C.1. Derivation of the single site equations

In this section we give details on the derivation of the single site equations for the GPMF model, starting from the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ given under its energy form in Equation 2.9. The process can be applied for all the processes with a similar energy definition, in particular the other models from the GPMF family.

We start using by the definition the conditional probabilities, for a fixed site $s \in \mathcal{S}$ we have:

$$\begin{aligned} p(x_s, y_s | x_{\mathcal{N}_s^X}, y_{\mathcal{N}_s^Y}) &= p(x_s, y_s | x_{\mathcal{S}\setminus\{s\}}, y_{\mathcal{S}\setminus\{s\}}), \ ((\boldsymbol{X}, \boldsymbol{Y}) \text{ Markovian}), \\ &= \frac{\tilde{p}(\boldsymbol{x}, \boldsymbol{y})}{\sum_{x_s \in \Omega} \int_{\mathbb{R}} p(x_s, y_s, x_{\mathcal{S}\setminus\{s\}}, y_{\mathcal{S}\setminus\{s\}}) \mathrm{d}y_s}, \\ &= \frac{\tilde{p}(\boldsymbol{x}, \boldsymbol{y})}{\sum_{x_s \in \Omega} \int_{\mathbb{R}} \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}y_s} \text{ since } \begin{cases} \boldsymbol{x} = (x_s, x_{\mathcal{S}\setminus\{s\}}), \\ \boldsymbol{y} = (y_s, y_{\mathcal{S}\setminus\{s\}}). \end{cases} \end{aligned} \tag{C.1}$$

Note that in the denominator, we sum over all the realizations of X_s and integrate over the domain of Y_s, s is fixed; this summation and integration do not happen at other sites. Now, using Equation 2.9 and developing all the terms, we have

$$\begin{aligned} p(x_s, y_s | x_{\mathcal{N}_s^X}, y_{\mathcal{N}_s^Y}) = & \frac{\exp\Big(-\Big(\sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{N}_s^X} V(x_s, x'_s) \dots}{\sum_{x_s} \int_{\mathbb{R}} \exp\Big(-\Big(\sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{N}_s^X} V(x_s, x_{s'}) \dots} \\ & \frac{\cdots + \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} Q_{s,s'}(y_s - \mu_{x_s})(y_{s'} - \mu_{x_{s'}})\Big)\Big)}{\cdots + \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} Q_{s,s'}(y_s - \mu_{x_s})(y_{s'} - \mu_{x_{s'}})\Big)\Big) \mathrm{d}y_s} \\ & \triangleq \frac{N}{D} \end{aligned} \tag{C.2}$$

Focus on the denominator D, we can extract all the terms that do not imply

the fixed site s from the sum and the integral:

$$\begin{aligned} D = \exp\Bigg(-\Bigg(&\sum_{\substack{s'\in\mathcal{S}\\ s'\neq s}}\sum_{\substack{s''\in\mathcal{N}_{s'}^X\\ s''\neq s}} V(x_{s'},x_{s''}) \\ &+\frac{1}{2}\sum_{\substack{(s',s'')\in\mathcal{S}^2\\ s'\neq s\wedge s''\neq s}} Q_{s',s''}(y_{s'}-\mu_{x_{s'}})(y_{s''}-\mu_{x_{s''}})\Bigg)\Bigg) \\ \times\sum_{x_s}\int_{\mathbb{R}}\exp\Bigg(-\Bigg(&\sum_{\substack{(s',s'')\in\mathcal{S}\times\mathcal{N}_{s'}^X\\ s'=s\vee s''=s}} V(x_{s'},x_{s''}) \\ &+\frac{1}{2}\sum_{\substack{(s',s'')\in\mathcal{S}^2\\ s'=s\vee s''=s}} Q_{s',s''}(y_{s'}-\mu_{x_{s'}})(y_{s''}-\mu_{x_{s''}})\Bigg)\Bigg)\mathrm{d}y_s. \end{aligned} \tag{C.3}$$

The first exponential term in D now gets simplified with terms from the numerator N. What remains then in N is the same terms as in the second exponential in D. So we can write:

$$\begin{aligned} p(x_s,y_s|x_{\mathcal{N}_s^X},y_{\mathcal{N}_s^Y}) = \frac{\exp\Big(-\Big(\sum_{\substack{(s',s'')\in\mathcal{S}\times\mathcal{N}_{s'}^X\\ s'=s\vee s''=s}} V(x_{s'},x_{s''})\dots}{\sum_{x_s}\int_{\mathbb{R}}\exp\Big(-\Big(\sum_{\substack{(s',s'')\in\mathcal{S}\times\mathcal{N}_{s'}^X\\ s'=s\vee s''=s}} V(x_{s'},x_{s''})\dots} \\ \frac{\cdots+\frac{1}{2}\sum_{\substack{(s',s'')\in\mathcal{S}^2\\ s'=s\vee s''=s}} Q_{s',s''}(y_{s'}-\mu_{x_{s'}})(y_{s''}-\mu_{x_{s''}})\Big)\Big)}{\cdots+\frac{1}{2}\sum_{\substack{(s',s'')\in\mathcal{S}^2\\ s'=s\vee s''=s}} Q_{s',s''}(y_{s'}-\mu_{x_{s'}})(y_{s''}-\mu_{x_{s''}})\Big)\Big)\mathrm{d}y_s}. \end{aligned} \tag{C.4}$$

The denominator is now a constant and can be discarded. We focus on the expression:

$$\begin{aligned} p(x_s,y_s|x_{\mathcal{N}_s^X},y_{\mathcal{N}_s^Y}) \propto \exp\Bigg(-\Bigg(&\sum_{\substack{(s',s'')\in\mathcal{S}\times\mathcal{N}_{s'}^X\\ s'=s\vee s''=s}} V(x_{s'},x_{s''}) \\ &+\frac{1}{2}\sum_{\substack{(s',s'')\in\mathcal{S}^2\\ s'=s\vee s''=s}} Q_{s',s''}(y_{s'}-\mu_{x_{s'}})(y_{s''}-\mu_{x_{s''}})\Bigg)\Bigg). \end{aligned} \tag{C.5}$$

Before further simplifications, note the fact that to get Equations C.3, C.4 and C.5 we used the following equality on sets, for a fixed $s\in\mathcal{S}$:

$$\begin{aligned} \{(s',s''):s'\in\mathcal{S}\wedge s''\in\mathcal{S}\} =&\{(s',s''):s'\in\mathcal{S}\wedge s''\in\mathcal{S}\wedge(s'\neq s\wedge s''\neq s)\} \\ &\cup\{(s',s''):s'\in\mathcal{S}\wedge s''\in\mathcal{S}\wedge(s'=s\vee s''=s)\}. \end{aligned} \tag{C.6}$$

For a fixed $s \in \mathcal{S}$, it can also be rewritten:

$$\begin{aligned}\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{S}\} = &\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{S} \wedge (s' \neq s \wedge s'' \neq s)\} \\ &\cup \{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{S} \wedge (s' = s \vee s'' = s) \wedge \neg(s' = s \wedge s'' = s)\} \\ &\cup \{(s, s)\}.\end{aligned} \tag{C.7}$$

Now it is important to note the following relation on cardinality, caused by the symmetry of the elements. For a fixed $s \in \mathcal{S}$:

$$\begin{aligned}|\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{S} \wedge (s' = s \vee s'' = s) \wedge \neg(s' = s \wedge s'' = s)\}| = \\ 2\,|\{(s', s'') : s' = s \wedge s'' \in \mathcal{S} \wedge s'' \neq s\}|\,.\end{aligned} \tag{C.8}$$

The same equalities slightly modified can be written for the set involving the set of the neighbors (recall that, by definition, the set of neighbors of the site s does not contain s), for a fixed $s \in \mathcal{S}$:

$$\begin{aligned}\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{N}_{s'}\} = &\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{N}_{s'} \wedge (s' \neq s \wedge s'' \neq s)\} \\ &\cup \{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{N}_{s'} \wedge (s' = s \vee s'' = s)\},\end{aligned} \tag{C.9}$$

with the following relation on cardinality:

$$\begin{aligned}|\{(s', s'') : s' \in \mathcal{S} \wedge s'' \in \mathcal{N}_{s'} \wedge (s' = s \vee s'' = s)\}| = \\ 2\,|\{(s', s'') : s' = s \wedge s'' \in \mathcal{N}_{s'} = \mathcal{N}_s\}|\,.\end{aligned} \tag{C.10}$$

By definition, in Equations C.9 and C.10, both s' and s'' can not simultaneously be equal to s. Finally, using the preceding remarks on sets, we can rewrite Equation C.5 in a simplified form, always for $s \in \mathcal{S}$ fixed:

$$\begin{aligned}p(x_s, y_s | x_{\mathcal{N}_s^X}, y_{\mathcal{N}_s^Y}) \propto \exp\Big(-\Big(2 \sum_{s' \in \mathcal{N}_s^X} V(x_s, x_{s'}) + \frac{1}{2} Q_{s,s}(y_s - \mu_{x_s})^2 \dots \\ \dots + \frac{1}{2} 2 \sum_{\substack{s' \in \mathcal{S} \\ s' \neq s}} Q_{s,s'}(y_s - \mu_{x_s})(y_{s'} - \mu_{x_{s'}})\Big)\Big).\end{aligned} \tag{C.11}$$

Remark: Note that Equations C.10 and C.11 only consider the current case where all the cliques are of size 2.

C.2. T-Gibbs sampler complementary experiment

In this section, we consider the P-GMRF model which tends to produce, for particular initializations, samples that worsen. This can be stabilized using the T-Gibbs sampler and the parametrization described in Section 2.5.1.

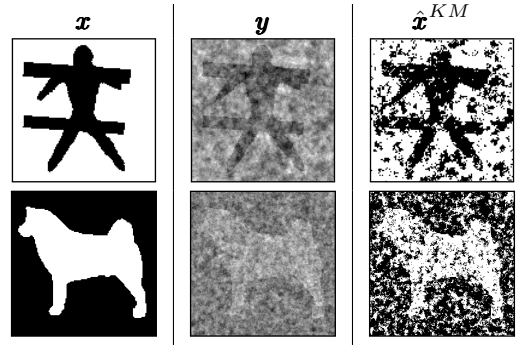
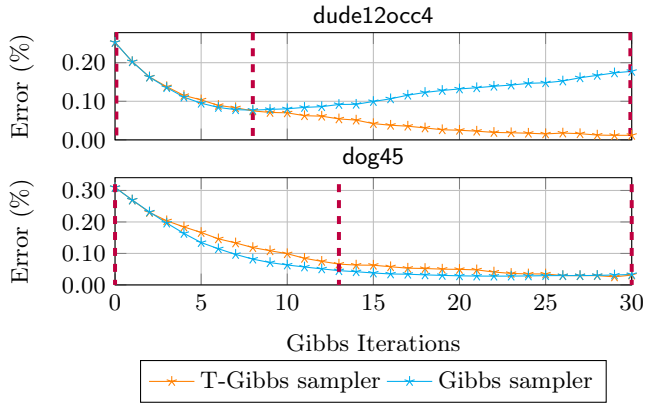


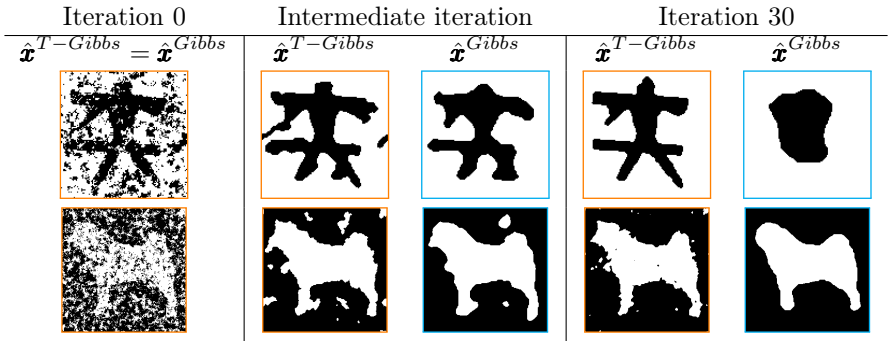
Figure C.1.: Initial data for the T-Gibbs experiment: the ground truth $\boldsymbol{x}$, the observations $\boldsymbol{y}$ and the Kmeans segmentation $\hat{\boldsymbol{x}}^{KM}$. In both cases the parameters of the simulated additive GMRF are $\mu_0 = 0, \mu_1 = 1, \sigma = 1$ and $r = 3$.

We consider the supervised segmentation of the dude12occ4 and dog45 images from the dataset with the P-GMRF model. Figure C.1 depicts the ground truth, the observed image and the KMeans segmentation which serves as an initialization for the Gibbs and T-Gibbs samplers. In this supervised context the parameters were estimated with the complete data $(\boldsymbol{x}, \boldsymbol{y})$.

The results presented in Figure C.1 are largely in favour of the T-Gibbs sampler. They illustrate two typical practical behaviors of the algorithms. The first is a total loss of the image details which occurs with the Gibbs sampler and not with the T-Gibbs sampler. The second is a faster convergence of the T-Gibbs sampler over the Gibbs sampler.



(a) Error rate as a function the iterations of Gibbs and T-Gibbs sampler in the segmentation of the observed images of Figure C.1 with the P-GMRF model. The purple vertical lines indicate iterations explored in (b), where we represent, from left to right, *Iteration 0*, *Intermediate iteration* and *Iteration 30*.



(b) Illustration of some iterations given in (a) to highlight the different behaviors of the Gibbs and T-Gibbs samplers.

Figure C.2.: Supervised segmentation experiment with the P-GMRF model, to compare the T-Gibbs and Gibbs samplers in typical cases. The final segmentation after 30 iterations of both samplers is better in the case of the T-Gibbs sampler. In a first typical scenario (found in the dude12occ4 case) the T-Gibbs sampler avoids a total loss of the image details and ends up with a much better error rate. The second case (dog45 case) illustrates, in the case of the T-Gibbs, a faster convergence and a final result which conserves more details despite a similar error rate to the Gibbs sampler.

Appendix D.

Complements on Variational Inference

D.1. Mean Field Variational Inference in SBNs

In the section we derive the parameter update equations that are needed to solve the maximization of the opposite of KL divergence of Equation 3.40. We first isolate terms involving q_j from the other:

$$-\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})) = \sum_{x_j} q_j(x_j) \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \log p(\boldsymbol{x}) \\ - \sum_{x_j} q_j(x_j) \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \left[\sum_{k\neq j} \log q_k(x_s) + \log q_j(x_j) \right]. \tag{D.1}$$

To maximize $-\mathbb{KL}(q||p)$ with the constraints $\forall i \in \mathcal{S}, \sum_{x_i} q_i(x_i) = 1$, we introduce Lagrangian multipliers $\lambda_i, \forall i \in \mathcal{S}$, (for each of the constraints) and we define the new target to maximize:

$$-\widetilde{\mathbb{KL}}(q(\boldsymbol{x})||p(\boldsymbol{x})) = -\mathbb{KL}(q(\boldsymbol{x})||p(\boldsymbol{x})) + \sum_i \lambda_i \left(\sum_{x_i} q_i(x_i) - 1 \right). \tag{D.2}$$

We now consider the functional derivative of $-\widetilde{\mathbb{KL}}(q(\boldsymbol{x})||p(\boldsymbol{x}))$ with respect to q_j, and then the derivative with respect to λ_j, both derivatives $\forall j \in \mathcal{S}$. Term by term we have:

$$\frac{\partial}{\partial q_j} \left\{ \sum_{x_j} q_j(x_j) \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \log p(\boldsymbol{x}) \right\} = \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \log p(\boldsymbol{x}), \tag{D.3}$$

$$\frac{\partial}{\partial q_j} \left\{ \sum_{x_j} q_j(x_j) \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \sum_{k\neq j} \log q_k(x_s) \right\} = \sum_{\boldsymbol{x}\setminus x_j} \prod_{i\neq j} q_i(x_i) \sum_{k\neq j} \log q_k(x_s), \tag{D.4}$$

$$\frac{\partial}{\partial q_j}\left\{\sum_{x_j} q_j(x_j)\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\log q_j(x_j)\right\} = \sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left(\log q_j(x_j)+1\right), \\ = \log q_j(x_j)+1, \tag{D.5}$$

$$\frac{\partial}{\partial q_j}\left\{\sum_i \lambda_i\left(\sum_{x_i} q_i(x_i)-1\right)\right\} = \sum_i \lambda_i, \tag{D.6}$$

where Equation D.3 depends on q_j through $p(\boldsymbol{x})$, Equation D.4 depends on q_j but Equations D.5 and D.6 do not depend on q_j. The second derivative gives:

$$\frac{\partial}{\partial \lambda_j}\left\{-\widetilde{\mathbb{KL}}(q(\boldsymbol{x})||p(\boldsymbol{x}))\right\} = \sum_{x_j} q_j(x_j) - 1. \tag{D.7}$$

We set the derivatives to 0, and we then need to solve the following system for q_j, $\forall j \in \mathcal{S}$:

$$\begin{cases} \sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\Big[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\Big] - \\ \qquad\qquad \log q_j(x_j) - 1 + \sum_i \lambda_i = 0, \\ \sum_{x_j} q_j(x_j) - 1 = 0. \end{cases} \tag{D.8}$$

With elementary manipulations we get:

$$\begin{cases} q_j(x_j) = \exp\left(\sum_i \lambda_i - 1\right)\exp\Big(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\Big[\log p(\boldsymbol{x}) - \\ \qquad\qquad\qquad\qquad \sum_{k\neq j}\log q_k(x_s)\Big]\Big), \\ \sum_{x_j} q_j(x_j) = 1. \end{cases} \tag{D.9}$$

By integrating the first equation (*i.e.* summing on all the $x_j \in \Omega$) and injecting the second in the first we get:

$$\exp\left(\sum_i \lambda_i - 1\right)\sum_{x_i}\exp\left(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\right]\right) = 1, \tag{D.10}$$

which becomes:

$$\exp\left(-\sum_i \lambda_i + 1\right) = \sum_{x_i}\exp\left(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\right]\right), \tag{D.11}$$

and finally:

$$\sum_i \lambda_i = 1 - \log\left(\sum_{x_i} \exp\left(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\right]\right)\right). \tag{D.12}$$

Plugging Equation D.12 in the first line of Equation D.9, $\forall j \in \mathcal{S}, \forall x_j \in \Omega$:

$$q_j(x_j) = \frac{1}{Z}\exp\left(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\right]\right), \tag{D.13}$$

where

$$Z = \sum_{x_j}\exp\left(\sum_{\boldsymbol{x}\setminus x_j}\prod_{i\neq j} q_i(x_i)\left[\log p(\boldsymbol{x}) - \sum_{k\neq j}\log q_k(x_s)\right]\right). \tag{D.14}$$

Now we remark that all the terms that do not involve x_j get simplified. We can start by simplifying the second term of the exponential:

$$q_j = \frac{1}{Z_j}\exp\left(\mathbb{E}_{\{x_i\}_{i\neq j}\sim\prod_{i\neq j} q_i(x_i)}\left[\log p(\boldsymbol{x})\right]\right), \tag{D.15}$$

with Z_j a normalization constant.

We now turn to the precise case of our study: p is the distribution of a SBN (using Equation 3.38). Equation D.15 can be further simplified, $\forall j \in \mathcal{S}$:

$$\begin{aligned} q_j(x_j) &= \frac{1}{Z_j}\exp\left(\mathbb{E}_{\{x_i\}_{i\neq j}\sim\prod_{i\neq j} q_i(x_i)}\left[\log p(\boldsymbol{x})\right]\right), \\ &= \frac{1}{Z_j}\exp\left(\mathbb{E}_{\{x_i\}_{i\neq j}\sim\prod_{i\neq j} q_i(x_i)}\left[\sum_{d_s}\log p_{d_s}\right]\right), \text{ (reparametrization)} \\ &= \frac{1}{Z'_j}\exp\left(\mathbb{E}_{\{x_i\}_{i\neq j}\sim\prod_{i\neq j} q_i(x_i)}\left[\sum_{d_j\in\mathcal{D}_j}\log p_{d_j}\right]\right), \text{ (simplifications)} \end{aligned} \tag{D.16}$$

where $\mathcal{D}_j$ is the set of clusters of variables containing variable x_j. The expectation distributes over the terms of $\mathcal{D}_j$, let us now make explicit one of these expectation terms. For a particular $d_j = (x_{s^1}, x_{s^2}, x_{s^3})$, we have that one of the three variables is x_j. Without loss of generality, if $x_{s^1} = x_j$:

$$\mathbb{E}_{\{x_i\}_{i\neq j}\sim\prod_{i\neq j} q_i(x_i)}\left[\log p_{d_j}\right] = \sum_{x_{s^2},x_{s^3}} q_{s^2}(x_{s^2})q_{s^3}(x_{s^3})\log p_{(x_j,x_{s^2},x_{s^3})}. \tag{D.17}$$

Using this result in the general expression of the variational parameter, $\forall j \in$

$\mathcal{S}, \forall x_j \in \Omega$:

$$q_j(x_j) = \frac{1}{Z'_j} \exp\left(\sum_{(x_j, x_{s^2}, x_{s^3}) = d_j \in \mathcal{D}_j} \mathbb{E}_{\{x_i\}_{i \neq j} \sim \prod_{i \neq j} q_i(x_i)} \left[\log p_{d_j} \right] \right),$$
$$= \frac{1}{Z'_j} \exp\left(\sum_{(x_j, x_{s^2}, x_{s^3}) = d_j \in \mathcal{D}_j} \sum_{x_{s^2}, x_{s^3}} q_{s^2}(x_{s^2}) q_{s^3}(x_{s^3}) \log p_{(x_j, x_{s^2}, x_{s^3})} \right). \tag{D.18}$$

We emphasize that the choice of $x_{s^1} = x_j$ is purely arbitrary and made to illustrate the fact that one of the variables in d_j is x_j, $\forall d_j \in \mathcal{D}_j$.

Remark: Due to the sparsity of the connections in SBNs, the summations induced by the expectations can be computed exactly. However, in denser graphical models, some additional approximations must be introduced for the tractability of these expectations. These approximations consider a modified objective function which constitutes an upper bound to the KL divergence to minimize. Such procedures to obtain a coarser but more tractable objective function often consist in the simplification (*e.g.* linearization) of the term $\log p(\boldsymbol{x})$. See for example (Wiegerinck, 2000) or (Depraetere and Vandebroek, 2017).

Finally the MF inference in SBN is done according to the following steps:

1. Initialize q for all factors.
2. $\forall j \in \mathcal{S}$, update q_j with the expression that maximizes Equation 3.40, *i.e.*, Equation D.18.
3. Check convergence and repeat step 2 if needed.

Following this procedure, the divergence is proved to decrease at each update of $q_j, \forall j \in \mathcal{S}$ (Wiegerinck, 2000).

D.2. Markov Tree Variational Inference in SBNs

This section gives details to solve the optimization problem given in Equation 3.42.

We first isolate one of the factors, r_{c_j} of the variational distribution r. This leads to splitting sums and products according to the different clusters $c_i, \forall i \in \mathcal{S}$. It follows:

$$
\begin{aligned}
-\mathbb{KL}(r(\boldsymbol{x})||p(\boldsymbol{x})) &= \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \left(\log p(\boldsymbol{x}) - \sum_{c_k} \log r_{c_k} \right), \\
&= \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \log p(\boldsymbol{x}) \\
&\quad - \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \log r_{c_j} \\
&\quad - \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \sum_{c_k \neq c_j} \log r_{c_k},
\end{aligned}
\tag{D.19}
$$

where we used the fact that $\sum_{c_k} r_{c_k} = \sum_{c_k \neq c_j} r_{c_k} + r_{c_j}$. We want to maximize this quantity with the constraints $\sum_{x_i} = r(x_i|x_{i^-}) = \sum_{x_i} r_{c_i} = 1, \forall i \in \mathcal{S}$. Therefore, similarly to Section D.1, we introduce Lagrangian multipliers $\lambda_i, \forall i \in \mathcal{S}$, and the quantity to maximize becomes:

$$
-\widetilde{\mathbb{KL}}(r(\boldsymbol{x})||p(\boldsymbol{x})) = -\mathbb{KL}(r(\boldsymbol{x})||p(\boldsymbol{x})) + \sum_i \lambda_i \left(\sum_{x_i} r_{c_i} - 1 \right). \tag{D.20}
$$

We now need to find the functional derivative of Equation D.20. Plugging Equation D.19 in Equation D.20, we exhibit the functional derivatives with respect to each r_{c_j} for each of the terms:

$$
\frac{\partial}{\partial r_{c_j}} \left\{ \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \log p(\boldsymbol{x}) \right\} = \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \log p(\boldsymbol{x}), \tag{D.21}
$$

$$
\begin{aligned}
\frac{\partial}{\partial r_{c_j}} \left\{ \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \log r_{c_j} \right\} &= \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} (\log r_{c_j} + 1), \\
&= \log r_{c_j} + 1,
\end{aligned}
\tag{D.22}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial r_{c_j}} \left\{ \sum_{x_j, x_{j^-}} r_{c_j} \sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \sum_{c_k \neq c_j} \log r_{c_k} \right\} = \\
&\sum_{\boldsymbol{x}\setminus\{x_j, x_{j^-}\}} \prod_{c_i \neq c_j} r_{c_i} \sum_{c_k \neq c_j} \log r_{c_k},
\end{aligned}
\tag{D.23}
$$

$$
\frac{\partial}{\partial r_{c_j}} \left\{ \sum_i \lambda_i \left(\sum_{x_i} r_{c_i} - 1 \right) \right\} = \sum_i \lambda_i. \tag{D.24}
$$

It follows that the functional derivative of Equation D.20 is, $\forall j \in \mathcal{S}$:

$$\frac{\partial}{\partial r_{c_j}} \left\{ -\widetilde{\mathbb{KL}}(r(\boldsymbol{x})||p(\boldsymbol{x})) \right\} = \mathbb{E}_{\{r_{c_i}\}_{c_i \neq c_j} \sim \prod_{c_i \neq c_j} r_{c_i}} \left[\log p(\boldsymbol{x}) - \log r_{c_j} - 1 - \sum_{c_k \neq c_j} \log r_{c_k} + \sum_i \lambda_i \right]. \tag{D.25}$$

And the derivative of Equation D.20, with respect to λ_j, is, $\forall j \in \mathcal{S}$:

$$\frac{\partial}{\partial \lambda_j} \left\{ -\widetilde{\mathbb{KL}}(r(\boldsymbol{x})||p(\boldsymbol{x})) \right\} = \sum_{x_j} r_{c_j} - 1. \tag{D.26}$$

We set Equations D.25 and D.26 equal to 0 and, following the same manipulations as done from Equation D.8 to Equation D.13, we can find the expression of the variational factor r_{c_j} minimizing the divergence, $\forall j \in \mathcal{S}$:

$$\log r_{c_j} = \frac{1}{Z} \exp \left(\mathbb{E}_{\{r_{c_i}\}_{c_i \neq c_j} \sim \prod_{c_i \neq c_j} r_{c_i}} \left[\log p(\boldsymbol{x}) - \sum_{c_k \neq c_j} \log r_{c_k} \right] \right), \tag{D.27}$$

where Z is a normalization constant. This expression can be simplified because many of the terms do not depend on c_j. Let $\mathcal{D}_j$ be a set whose elements are clusters of variables d_j containing x_j, $\mathcal{B}_j$ a set of clusters of variables b_j also containing x_j but with the condition $b_j \neq c_j$. Equation D.27 simplifies in:

$$\begin{aligned} r_{c_j} &= r(x_j | x_{j^-}), \\ &= \frac{1}{Z_j} \exp \left(\mathbb{E}_{\{r_{c_i}\}_{c_i \neq c_j} \sim \prod_{c_i \neq c_j} r_{c_i}} \left[\sum_{d_j \in \mathcal{D}_j} \log p_{d_j} - \sum_{b_j \in \mathcal{B}_j} \log r_{b_j} \right] \right). \end{aligned} \tag{D.28}$$

with Z_j a normalization constant.

The expectation appearing in Equation D.28 requires to sample from the joint law $\boldsymbol{x} \setminus \{x_j, x_{j^-}\}$ given x_j and x_{j^-}. Such a sampling can be done when r is a MT because of the straightforward and sparse decomposition. Note also that, in fact, we only need to sample the variables in each cluster (upon which the expectation is computed, *i.e.* d_j or b_j). Then $\boldsymbol{X}$ does not need to be fully sampled if we recompute the local marginal distributions and the local transition distributions between the MT VI procedure. Note that further developments would consider rewriting Equation D.28 similarly to Equation D.17.

To summarize the steps of the MT-structured variational inference in SBNs:

1. Initialize r for all factors.

2. $\forall j \in \mathcal{S}$, update r_{c_j} with the expression that maximizes Equation 3.42, *i.e.*, Equation D.28.

3. Check convergence and repeat step 2 if needed.

Remark: Because r is not fully factorized, we observe, in the update Equation D.28, as opposed to the MF variational inference, a dependency on variables of r itself. Moreover, for a given update of r_{c_j}, the variables of p and r which play a role are all in the Markov blanket of x_j.

D.3. STMT auxiliary variable Variational Inference in SBNs

As opposed to the previous sections, we now need to derive the update equations of three variational parameters $t_{c_i}, t_{c'_i}$ and $t_{c''_i}, \forall i \in \mathcal{S}$, to solve the maximization of Equation 3.45.

Update for t_{c_j}

We have, by isolating a particular t_{c_j}:

$$
\begin{aligned}
-\mathbb{KL}(t||p) = & \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \Bigg(\log p(\boldsymbol{x}, \boldsymbol{v}) - \\
& \qquad \left(\sum_{c_k} \log t_{c_k} + \sum_{c'_k} \log t_{c'_k} + \sum_{c''_k} \log t_{c''_k} \right) \Bigg), \\
= & \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \log p(\boldsymbol{x}, \boldsymbol{v}) \qquad \text{(A)} \\
& - \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \sum_{c_k \neq c_j} \log t_{c_k} \qquad \text{(B)} \\
& - \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \sum_{c'_k} \log t_{c'_k} \qquad \text{(C)} \\
& - \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \sum_{c''_k} \log t_{c''_k} \qquad \text{(D)} \\
& - \sum_{x_j, x_{j^-}, v_{n(j^-)}} t_{c_j} \sum_{\substack{\boldsymbol{x}\setminus\{x_j, x_{j^-}\} \\ \boldsymbol{v}\setminus\{v_{n(j^-)}\}}} \prod_{c_i \neq c_j} t_{c_i} \prod_{c'_s} t_{c'_s} \prod_{c''_s} t_{c''_s} \log t_{c_j} \qquad \text{(E)}.
\end{aligned}
\tag{D.29}
$$

As previously done, we rewrite this objective to maximize introducing a Lagrange multipliers $\lambda_i, \forall i \in \mathcal{S}$, to fulfill the constraints, $\forall i \in \mathcal{S}$:

$$\sum_{x_i, v_i^{\leftarrow}, v_i^{\rightarrow}} t(x_i|x_{i^-}, \boldsymbol{v}_{i^-}) t(v_i^{\leftarrow}|x_{i^-}, \boldsymbol{v}_{i^-}) t(v_i^{\rightarrow}|x_{i^-}, \boldsymbol{v}_{i^-}) = \sum_{x_i, v_i^{\leftarrow}, v_i^{\rightarrow}} t_{c_i} t_{c_i'} t_{c_i''} = 1. \tag{D.30}$$

We then want to maximize:

$$\begin{aligned} -\widetilde{\mathbb{KL}}(t(\boldsymbol{x}, \boldsymbol{v})||p(\boldsymbol{x}, \boldsymbol{v})) &= -\mathbb{KL}(t(\boldsymbol{x}, \boldsymbol{v})||p(\boldsymbol{x}, \boldsymbol{v})) + \underbrace{\sum_i \lambda_i \left(\sum_{x_i, v_i^{\leftarrow}, v_i^{\rightarrow}} t_{c_i} t_{c_i'} t_{c_i''} - 1 \right)}_{F}, \\ &= A - B - C - D - E + F. \end{aligned} \tag{D.31}$$

Plugging Equation D.29 into Equation D.31, the functional derivatives of each term, with respect to t_{c_j}, are:

$$\begin{aligned} \frac{\partial}{\partial t_{c_j}} \{A\} &= \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c_i'} t_{c_i'} \prod_{c_i''} t_{c_i''}} [\log p(\boldsymbol{x}, \boldsymbol{v})], \\ \frac{\partial}{\partial t_{c_j}} \{B\} &= \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c_i'} t_{c_i'} \prod_{c_i''} t_{c_i''}} \left[\sum_{c_k \neq c_j} \log t_{c_k} \right], \\ \frac{\partial}{\partial t_{c_j}} \{C\} &= \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c_i'} t_{c_i'} \prod_{c_i''} t_{c_i''}} \left[\sum_{c_k'} \log t_{c_k'} \right], \\ \frac{\partial}{\partial t_{c_j}} \{D\} &= \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c_i'} t_{c_i'} \prod_{c_i''} t_{c_i''}} \left[\sum_{c_k''} \log t_{c_k''} \right], \\ \frac{\partial}{\partial t_{c_j}} \{E\} &= \mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c_i'} t_{c_i'} \prod_{c_i''} t_{c_i''}} [\log t_{c_j} + 1], \\ &= \log t_{c_j} + 1, \\ \frac{\partial}{\partial t_{c_j}} \{F\} &= \sum_i \lambda_i. \end{aligned} \tag{D.32}$$

The derivative of Equation D.31 with respect to $\lambda_j, \forall j \in \mathcal{S}$, is:

$$\frac{\partial}{\partial \lambda_j} \left\{ -\widetilde{\mathbb{KL}}(t(\boldsymbol{x}, \boldsymbol{v})||p(\boldsymbol{x}, \boldsymbol{v})) \right\} = \sum_{x_j, v_j^{\leftarrow}, v_j^{\rightarrow}} t_{c_j} t_{c_j'} t_{c_j''}. \tag{D.33}$$

We set Equations D.32 and D.33 equal to 0 and, following the same manipulations as done from Equation D.8 to Equation D.13, we can find the expression of the variational factor t_{c_j} minimizing the divergence, $\forall j \in \mathcal{S}$:

$$
\begin{aligned}
t_{c_j} &= \frac{1}{Z} \exp\left(\mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c'_i} t_{c'_i} \prod_{c''_s} t_{c''_s}} \left[\log p(\boldsymbol{x}, \boldsymbol{v}) - \sum_{c_k \neq c_j} \log t_{c_k} - \right. \right. \\
&\qquad \left. \left. \sum_{c'_k} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \right] \right), \\
&= \frac{1}{Z'} \exp\left(\mathbb{E}_{\prod_{c_i \neq c_j} t_{c_i} \prod_{c'_i} t_{c'_i} \prod_{c''_s} t_{c''_s}} \left[\log p(\boldsymbol{x}) - \sum_{c_k \neq c_j} \log t_{c_k} - \right. \right. \\
&\qquad \left. \left. \sum_{c'_k} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \right] \right),
\end{aligned}
\tag{D.34}
$$

where Z and Z' are different normalization constants. Indeed some simplifications were made on on the terms of the sums which do not involve c_j.

Following the same developments as for other VI procedures, we can exhibit easily the fact that many more terms which do not involve c_j can be simplified. Indeed, since STMTs still have a relatively sparse and highly structured associated graph, by carefully selecting the subsets of the variables involved in the expectations, we can straightforwardly simplify the last expression as previously done.

Update for $t_{c'_j}$

Similarly we find:

$$
\begin{aligned}
t_{c'_j} &= \frac{1}{Z} \exp\left(\mathbb{E}_{\prod_{c_i} t_{c_i} \prod_{c'_i \neq c_j} t_{c'_i} \prod_{c''_i} t_{c''_i}} \left[\log p(\boldsymbol{x}, \boldsymbol{v}) - \sum_{c_k} \log t_{c_k} - \right. \right. \\
&\qquad \left. \left. \sum_{c'_k \neq c'_j} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \right] \right), \\
&= \frac{1}{Z'} \exp\left(\mathbb{E}_{\prod_{c_i} t_{c_i} \prod_{c'_i \neq c_j} t_{c'_i} \prod_{c''_i} t_{c''_i}} \left[\log p(\boldsymbol{v}|\boldsymbol{x}) - \sum_{c_k} \log t_{c_k} - \right. \right. \\
&\qquad \left. \left. \sum_{c'_k \neq c'_j} \log t_{c'_k} - \sum_{c''_k} \log t_{c''_k} \right] \right),
\end{aligned}
\tag{D.35}
$$

where Z and Z' are different normalization constants. Again, further simplifications can be made straightforwardly.

Update for $t_{c''_j}$

Similarly we find:

$$t_{c''_j} = \frac{1}{Z} \exp\left(\mathbb{E}_{\prod_{c_i} t_{c_i} \prod_{c'_i} t_{c'_i} \prod_{c''_i \neq c_j} t_{c''_i}}\left[\log p(\boldsymbol{v}|\boldsymbol{x}) - \sum_{c_k} \log t_{c_k} - \sum_{c'_k} \log t_{c'_k} - \sum_{c''_k \neq c''_j} \log t_{c''_k}\right]\right) \tag{D.36}$$

where Z is a normalization constant. Again, further simplifications can be made straightforwardly.

Remark: We can conclude this section recalling the remark at the end of the section on MT VI: in the final update equations for STMT VI we see more dependencies on the distribution t itself than in Equation D.28. Indeed this comes from the fact that t, which follows a STMT distribution, has more direct dependencies between its random variables than r which follows a MT distribution.

Appendix E.

Complements on fcCRFs

E.1. Augmented space for the bilateral kernel computations

In this section we demonstrate how the term $\sum_{i\in\mathcal{S}} k_1(\mathbf{f}_j,\mathbf{f}_i)q_i(x')$, $\forall x' \in \Omega$, $\forall j \in \mathcal{S}$ (from Equation 4.8) can be interpreted as a convolution so that it can be approximated with the approach seen in Section 4.3.2. Recall that k_1 is defined in Equation 4.5 and is called the bilateral kernel. It is the product of a Gaussian on the spatial distance and a Gaussian on the pixel value difference (Paris, Kornprobst, et al., 2009). The concept of augmented space has been defined in (Chen et al., 2007) based on (Paris and F. Durand, 2006). We now study how the derivations from the latter adapt to our case.

We have, $\forall x' \in \Omega$, $\forall j \in \mathcal{S}$:

$$\begin{aligned}\tilde{q}_{j,1}(x') &= \sum_{i\in\mathcal{S}} k_1(\mathbf{f}_j,\mathbf{f}_i)q_i(x') - q_j(x'),\\ &= \sum_{i\in\mathcal{S}} \exp\left(-\frac{|j-i|^2}{2\theta_\alpha^2} - \frac{|y_j-y_i|^2}{2\theta_\beta^2}\right) q_i(x') - q_j(x'),\\ &= \sum_{i\in\mathcal{S}} G_{\theta_\alpha}(|j-i|)G_{\theta_\beta}(|y_j-y_i|)q_i(x') - q_j(x'),\end{aligned} \tag{E.1}$$

where $G_{\theta_\alpha}(|s-s'|)$ and $G_{\theta_\beta}(|y_s-y_{s'}|)$ are the two Gaussian kernels respectively on the spatial distance and pixel value difference. The *augmented space* technique now appears by summing over an auxiliary variable defined in the space of the pixel intensities $\mathcal{I}$[1]. The so-called augmented space is the product space $\mathcal{S}\times\mathcal{I}$. It follows, with δ the Kronecker function:

$$\tilde{q}_{j,1}(x') = \sum_{i\in\mathcal{S}}\sum_{\xi\in\mathcal{I}} G_{\theta_\alpha}(|j-i|)G_{\theta_\beta}(|y_j-\xi|)\delta_\xi^{y_i} q_i(x') - q_j(x'). \tag{E.2}$$

We can see that $\dot{G}_{\theta_\alpha\theta_\beta} = G_{\theta_\alpha}(|j-i|)G_{\theta_\beta}(|y_j-\xi|)$ defines a Gaussian kernel on $\mathcal{S}\times\mathcal{I}$. Then, $\forall x' \in \Omega$, define the augmented space variational distribution $\dot{q}(x')$ with, $\forall (i,\xi) \in \mathcal{S}\times\mathcal{I}$:

$$\dot{q}_{i,\xi}(x') = \delta_\xi^{y_i} q_i(x'), \tag{E.3}$$

[1] Typically $\mathcal{I} = [0, 255]$ for grayscale images, or another integer interval depending on the image format. $\mathcal{I}$ can also be vectorial for color images.

It follows that we can finally write Equation E.1 as a convolution in the augmented space:

$$\tilde{q}_{j,1}(x') = [\dot{G}_{\theta_\alpha \theta_\beta} * \dot{q}(x')](\mathbf{f}_j) - q_j(x'). \tag{E.4}$$

As an illustration, if $\mathcal{I}$ is an unidimensional space (grayscale image) and $\mathcal{S}$ the set of vertices of a d-dimensional image, the augmented space variational distribution $\bar{q}(x')$,$\forall x' \in \Omega$ has $d+1$ finite dimensions. The data cube it forms is sparse, it is null everywhere except at one point of the augmented axis $\mathcal{I}$, for each site i of the image in $\mathcal{S}$, where it takes the value $q_i(x')$. We then have:

$$\dot{q}_{i,\xi}(x') = \begin{cases} 0, & \forall i \in \mathcal{S}, \forall \xi \in \mathcal{I} \setminus \{y_i\}, \\ q_i(x'), & \forall i \in \mathcal{S}, \text{ if } \xi = y_i. \end{cases} \tag{E.5}$$

E.2. Markov Chain Variational Inference in fcCRFs

In this section we derive the equations to solve the optimization problem of Equation 4.11. In what follows, N is the number of parallel MCs and M is their length. The initial steps of the derivation are similar to the Markov Tree VI update equations in Appendix D.2. However, we need to consider the visible random variables. We get, $\forall n \in \{1, \dots, N\}$, $\forall m \in \{2, \dots, M\}$[2] and $\forall (x_m^n, x_{m-1}^n) \in \Omega^2$:

$$\begin{aligned}
l_m^n(x_m^n | x_{m-1}^n) &= \frac{1}{Z} \exp\left(\mathbb{E}_{\boldsymbol{x}_{-m} \sim l_1^n(x_1^n) \prod_{m' \neq m} l_{m'}^n(x_{m'}^n | x_{m'-1}^n)} \left[\log p(\boldsymbol{x}|\boldsymbol{y}) - \right. \right. \\
&\left. \left. \sum_{m'' \neq m} \log l_{m''}^n(x_{m''}^n | x_{m''-1}^n) + l_1^n(x_1^n) \right] \right), \\
&= \frac{1}{Z'} \exp\left(\mathbb{E}_{\boldsymbol{x}_{-m} \sim l_1^n(x_1^n) \prod_{m' \neq m} l_{m'}^n(x_{m'}^n | x_{m'-1}^n)} \left[- \sum_{s \in \mathcal{S}} \psi_u(x_s) - \right. \right. \\
&\left. \left. \sum_{(s,s') \in \mathcal{S}^2} \psi_p(x_s, x_{s'}) - \sum_{m'' \neq m} \log l_{m''}^n(x_{m''}^n | x_{m''-1}^n) + l_1^n(x_1^n) \right] \right), \\
&= \frac{1}{Z''} \exp\left(- \psi_u(x_m^n) + \mathbb{E}_{\boldsymbol{x}_{-m} \sim l_1^n(x_1^n) \prod_{m' \neq m} l_{m'}^n(x_{m'}^n | x_{m'-1}^n)} \left[- \right. \right. \\
&\left. \left. \sum_{\substack{m', n' \\ (m',n') \neq (m,n)}} \psi_p(x_m^n, x_{m'}^{n'}) - \sum_{m'' \neq m} \log l_{m''}^n(x_{m''}^n | x_{m''-1}^n) + l_1^n(x_1^n) \right] \right),
\end{aligned} \tag{E.6}$$

where Z, Z' and Z'' are normalizing constants. $\boldsymbol{x}_{-m}$ is a short notation for the vector $(x_1, \dots, x_{m-1}, x_{m+1}, \dots, x_M)$. We used Equations 4.3 and 4.4 which

[2] The case $m = 1$ can be straightforwardly derived.

were successively simplified by discarding the constant terms with respect to l_m^n.

Further simplifications involve the last summation term inside the exponential. There is only one $l_{m''}^n$ term such that $m'' \in \{2, \dots, M\}$ and $m'' \neq m$ and that plays a role with l_m^n: $l_{m-1}^n(x_{m-1}^n | x_{m-2}^n)$. That is, in the last line of Equation E.6, $\sum_{m'' \neq m} \log l_{m''}^n(x_{m''}^n | x_{m''-1}^n)$ can be simplified in $\log l_{m-1}^n(x_{m-1}^n | x_{m-2}^n)$. Note also that $l_1^n(x_1^n)$ gets simplified since, for brevity, we also consider $m'' \neq 1$. Thus, $\forall n \in \{1, \dots, N\}$, $\forall m \in \{2, \dots, M\}$ and $\forall (x_m^n, x_{m-1}^n) \in \Omega^2$:

$$
\begin{aligned}
l_m^n(x_m^n | x_{m-1}^n) = \frac{1}{Z} \exp \Bigg(& - \psi_u(x_m^n) - \\
& \underbrace{\sum_{\substack{m', n' \\ (m', n') \neq (m, n)} } \sum_{(x_{m'}^{n'}, x_{m'-1}^{n'}) \in \Omega^2} l_{m'}^{n'}(x_{m'}^{n'} | l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) \psi_p(x_m^n, x_{m'}^{n'}) }_{A} - \\
& \sum_{x_{m-2}^n \in \Omega} l_{m-1}^n(x_{m-1}^n | x_{m-2}^n) l_{m-2}^n(x_{m-2}^n) \log l_{m-1}^n(x_{m-1}^n | x_{m-2}^n) \Bigg),
\end{aligned}
\tag{E.7}
$$

where Z is a new normalization constant. Let $\boldsymbol{m}^{--}$ denotes the set $\{m-1, m-2, \dots, 1\}$ The last equation is established by taking into account the chain structure: we only sample the variables of interest using the local marginal and transition distributions. Indeed we know that:

$$
\begin{aligned}
\sum_{(x_{m'}^{n'}, x_{m'-1}^{n'}) \in \Omega^2} & l_{m'}^{n'}(x_{m'}^{n'} | l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) \approx \\
& \sum_{\substack{x_m^n \in \Omega \\ \boldsymbol{x}_{\boldsymbol{m}^{--}}^n \in \Omega^{|\boldsymbol{m}^{--}|}}} l_m^n(x_m^n | x_{m-1}^n) \dots l_2(x_2^n | x_1^n) l_1^n(x_1^n),
\end{aligned}
\tag{E.8}
$$

and, for a fixed x_{m-1}^n:

$$
\begin{aligned}
\sum_{x_{m-2}^n \in \Omega} & l_{m-1}^n(x_{m-1}^n | x_{m-2}^n) l_{m-2}^n(x_{m-2}^n) \approx \\
& \sum_{\boldsymbol{x}_{(m-1)^{--}} \in \Omega^{|(m-1)^{--}|}} l_{m-1}^n(x_{m-1}^n | x_{m-2}^n) \dots l_2^n(x_2^n | x_1^n) l_1^n(x_1^n).
\end{aligned}
\tag{E.9}
$$

These approximations can be done at each iteration of the update algorithm using the Forward Backward (FB) algorithm (Algorithm A.2)[3] to compute the

[3] Note that in this case, we use the FB algorithm without any observations, this is equivalent to taking all the conditional likelihood terms of the equations of Algorithm A.2 as equal to 1.

marginal distributions that are used. These approximations are obvious for computational reasons.

We now focus on the term A of Equation E.7 which requires a summation on all the sites:

$$
\begin{aligned}
A = & \sum_{\substack{m',n' \\ (m',n')\neq(m,n)}} \sum_{(x_{m'}^{n'},x_{m'-1}^{n'})\in\Omega^2} l_{m'}^{n'}(x_{m'}^{n'}|l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) \Big[\\
& (1-\delta_{x_m^n}^{x_{m'}^{n'}}) \sum_{r=1}^{2} w_r k_r(\mathbf{f}_m^n, \mathbf{f}_{m'}^{n'}) \Big], \\
= & \sum_{r=1}^{2} w_r \sum_{i\neq j} \sum_{\substack{m',n' \\ (m',n')\neq(m,n)}} \Bigg[\\
& \sum_{(x_{m'}^{n'},x_{m'-1}^{n'})\in\Omega^2} l_{m'}^{n'}(x_{m'}^{n'}|l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) (1-\delta_{x_m^n}^{x_{m'}^{n'}}) k_r(\mathbf{f}_m^n, \mathbf{f}_{m'}^{n'}) \Bigg], \\
= & \sum_{(x_{m'}^{n'},x_{m'-1}^{n'})\in\Omega^2} (1-\delta_{x_m^n}^{x_{m'}^{n'}}) \sum_{r=1}^{2} w_r \Bigg[\\
& \sum_{\substack{m',n' \\ (m',n')\neq(m,n)}} k_r(\mathbf{f}_m^n, \mathbf{f}_{m'}^{n'}) l_{m'}^{n'}(x_{m'}^{n'}|l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) \Bigg], \\
= & \sum_{x_{m'}^{n'}\in\Omega} (1-\delta_{x_m^n}^{x_{m'}^{n'}}) \sum_{r=1}^{2} w_r \sum_{\substack{m',n' \\ (m',n')\neq(m,n)}} \Bigg[\\
& k_r(\mathbf{f}_m^n, \mathbf{f}_{m'}^{n'}) \sum_{x_{m'-1}^{n'}\in\Omega} l_{m'}^{n'}(x_{m'}^{n'}|l_{m'-1}^{n'}) l_{m'-1}^{n'}(x_{m'-1}^{n'}) \Bigg].
\end{aligned}
\tag{E.10}
$$

The manipulation performed in Equation E.10 was already seen in Section 4.3.2. It enables the reinterpretation of the equation as a convolution.

Remark: For completeness, let us give Equation E.6 when formulated in terms of clusters of variables as we did in all the other VI derivations of this thesis. We have, $\forall j \in \mathcal{S}$:

$$
\begin{aligned}
l_{c_j} &= \frac{1}{Z} \exp\left(\mathbb{E}_{\prod_{c_i \neq c_j} l_{c_i}} \left[\log p(\boldsymbol{x}|\boldsymbol{y}) - \sum_{c_k \neq c_j} \log l_{c_k} \right] \right), \\
&= \frac{1}{Z'} \exp\left(\mathbb{E}_{\prod_{c_i \neq c_j} l_{c_i}} \left[-E(\boldsymbol{x}|\boldsymbol{y}) - \sum_{c_k \neq c_j} \log l_{c_k} \right] \right), \\
&= \frac{1}{Z'} \exp\left(\mathbb{E}_{\prod_{c_i \neq c_j} l_{c_i}} \left[-\sum_{s \in \mathcal{S}} \psi_u(x_s) - \sum_{(s,s') \in \mathcal{S}^2} \psi_p(x_s, x_{s'}) - \right. \right. \\
&\qquad \left. \left. \sum_{c_k \neq c_j} \log l_{c_k} \right] \right), \\
&= \frac{1}{Z''} \exp\left(-\psi_u(x_j) + \mathbb{E}_{\prod_{c_i \neq c_j} l_{c_i}} \left[-\sum_{i \neq j} \psi_p(x_j, x_i) - \sum_{c_k \neq c_j} \log l_{c_k} \right] \right), \\
&\quad \text{with } c_k \text{ s.t. } c_k \cap c_j \neq \emptyset \text{ and } c_k \neq c_j,
\end{aligned}
\tag{E.11}
$$

Constant terms with respect to c_j are successively simplified. We have, from the condition on c_k, that the only possibility for c_k to have a non-empty intersection with c_j is that $c_k = (c_{j^-}, c_{(j^-)^-})$. This also leads to Equation E.7. Note that the last line of Equation E.11, without the last summation on $\log l_{c_k}$, $\forall c_k \neq c_j$, falls back on the derivation of the MF updates given in Equation 4.6.

Appendix F.

Complements on the applications to vascular surgery

F.1. Fine stent segmentation: exponential mixture models

In this section we show that a mixture of exponential (MoE) distributions fits better the observed realizations of random variables (the image) that we want to segment. Works such as (Delignon et al., 1997) (Pieczynski, Bouvrais, et al., 2000) and (Monfrini and Pieczynski, 2005) involve the use of the exponential distribution in Hidden Markov Models and motivated the work we present here.

Remark: Our work does not make directly use of the notion of *generalized* mixture models introduced in the articles cited above. Within generalized models the distributions that compose the mixture have to be decided statistically according to given rules. In our case, since the observed realizations vary weakly between images to segment, it is reasonable to fix a given mixture beforehand. This way we avoid the additional complexity of estimating the mixture type before estimating its parameters.

Recall the definition of the likelihood given in Equation 1.6. It becomes, for $\boldsymbol{y} = (y_1, \dots, y_n)$ the observed independent realizations of the random process that represents the image:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} p(y_i; \boldsymbol{\theta}), \tag{F.1}$$

where:

$$\boldsymbol{\theta} = \begin{cases} (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*, \text{ for Gaussian distributions,} \\ (\lambda, \delta) \in \mathbb{R}_+^* \times \mathbb{R}, \text{ for exponential distributions.} \end{cases} \tag{F.2}$$

We will use the BIC score (Wit et al., 2012) as a way to find the best fitting mixture to our data. The BIC score for a model is defined as:

$$\text{BIC} = -2 \ln L(\boldsymbol{\theta}; \boldsymbol{y}) + p \ln n, \tag{F.3}$$

where p is the number of parameters estimated in the model and n is the size of the sample. The BIC values are computed with the resulting MLE parameters

Class \ Mixture model	MoG	MoE-G	MoE
stent and *rest*	$77.39 \cdot 10^7$	$77.37 \cdot 10^7$	$\mathbf{63.97 \cdot 10^7}$

Table F.1.: BIC values for different mixtures of distributions on a typical 3D image. In the case of the mixture of exponential-Gaussian (MoE-G), the exponential distribution is attributed to the stent. It appears that the MoE fits better the data (lower BIC value).

found for each distribution. The latter were computed with the *stent* and *rest* segmentations given by the preprocessing Frangi-based step. We know that a lower BIC value indicates a better fitting model (more likely to be the true model). The results illustrated in Table F.1 show that a lower BIC value is attained for the MoE. Figure F.1 gives the histograms of the pixel values according to the two classes of the segmentation problem. It is notable that the histogram associated with each class has an exponential shape.

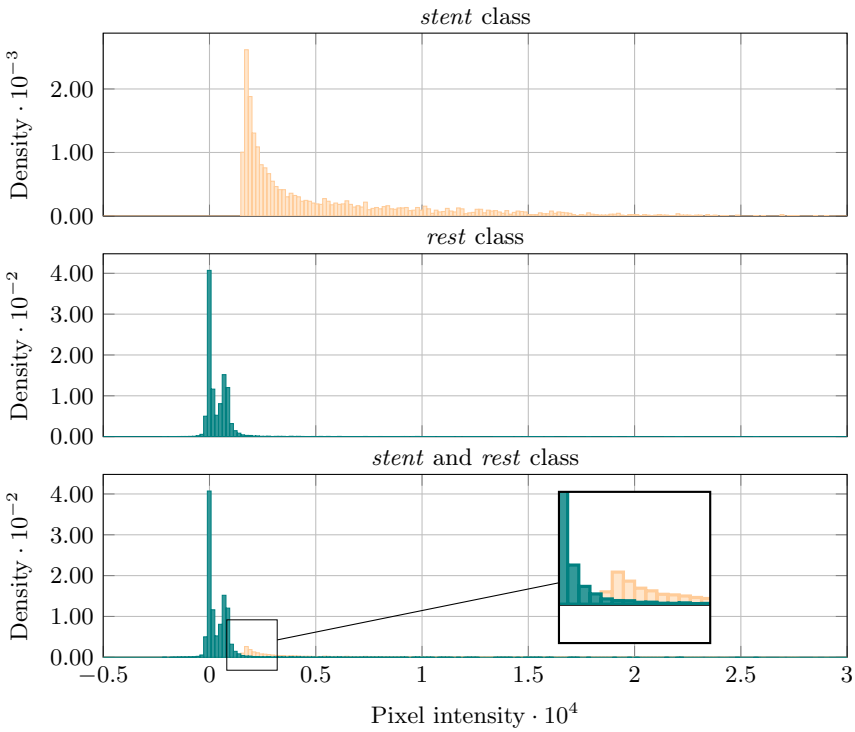


Figure F.1.: Histograms for the realizations of the random variables for the *stent* and *rest* classes, on a typical data slice, after the preprocessing segmentation. The density values are divided by the total number of pixel in the slice on the three histograms.

Bibliography

Petter Abrahamsem. *A review of Gaussian random fields and correlation functions*. Technical Report No. 917. Norwegian Computing Center, 1997 (cit. on pp. 157, 158).

Felix V Agakov and David Barber. "An auxiliary variational method". In: *International Conference on Neural Information Processing*. Springer. 2004, pp. 561–566 (cit. on pp. 68, 88, 91).

Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation". In: *arXiv preprint arXiv:1802.06955* (2018) (cit. on p. 152).

David Arthur and Sergei Vassilvitskii. "K-means++: The advantages of careful seeding". In: *Proceedings of the 18th ACM-SIAM symposium on Discrete algorithms*. 2007, pp. 1027–1035 (cit. on p. 60).

Leonard E Baum and Ted Petrie. "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6 (1966), pp. 1554–1563 (cit. on pp. 2, 19, 35, 36).

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* 41.1 (1970), pp. 164–171 (cit. on pp. 29, 154).

Avi Ben-Cohen, Idit Diamant, Eyal Klang, Michal Amitai, and Hayit Greenspan. "Fully convolutional network for liver segmentation and lesions detection". In: *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 77–85 (cit. on pp. 3, 100).

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. "Greedy layer-wise training of deep networks". In: *Advances in neural information processing systems*. 2007, pp. 153–160 (cit. on p. 101).

Julian Besag. "On the statistical analysis of dirty pictures". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 48.3 (1986), pp. 259–279 (cit. on pp. 38, 48, 75, 80).

Julian Besag. "Statistical analysis of non-lattice data". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3 (1975), pp. 179–195 (cit. on p. 33).

Julian E Besag. "Nearest-neighbour systems and the auto-logistic model for binary data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.1 (1972), pp. 75–83 (cit. on p. 38).

Christopher Bishop and John Winn. "Structured variational distributions in VIBES". In: *Proceedings Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003. ISBN: 0-9727358-0-1 (cit. on p. 87).

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006 (cit. on p. 24).

Ottar N Bjørnstad and Wilhelm Falck. "Nonparametric spatial covariance functions: estimation and testing". In: *Environmental and Ecological Statistics* 8.1 (2001), pp. 53–70 (cit. on p. 52).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877 (cit. on p. 86).

Jiří Borovec, Jan Švihlík, Jan Kybic, and David Habart. "Supervised and unsupervised segmentation using superpixels, model estimation, and graph cut". In: *Journal of Electronic Imaging* 26.6 (2017), pp. 1–17 (cit. on p. 60).

Yuri Boykov and Vladimir Kolmogorov. "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision." In: *IEEE transactions on Pattern Analysis and Machine Intelligence* 26.9 (2004), pp. 1124–1137 (cit. on p. 60).

Katharina Breininger, Shadi Albarqouni, Tanja Kurzendorfer, Marcus Pfister, Markus Kowarschik, and Andreas Maier. "Intraoperative stent segmentation in X-ray fluoroscopy for endovascular aortic repair". In: *International journal of computer assisted radiology and surgery* 13.8 (2018), pp. 1221–1231 (cit. on p. 150).

Pierre Brémaud. *Discrete Probability Models and Methods: Probability on Graphs and Trees, Markov Chains and Random Fields, Entropy and Coding*. Springer, 2017 (cit. on p. 36).

Stephanie Bricq, Ch Collet, and Jean-Paul Armspach. "Unifying framework for multimodal brain MRI segmentation based on Hidden Markov Chains". In: *Medical image analysis* 12.6 (2008), pp. 639–652 (cit. on pp. 4, 108).

Clayton J Brinster, G Thomas Escousse, Hernan A Bazan, Charles C Leithead, and W Charles Sternbergh III. "Financial viability of endovascular aortic repair in the modern era: a single center experience". In: *Journal of Vascular Surgery* (2020). In press (cit. on p. 11).

D Andrew Brown, Christopher S McMahan, and Stella Watson Self. "Sampling strategies for fast updating of Gaussian Markov random fields". In: *The American Statistician* (2019), pp. 1–24 (cit. on pp. 28, 48, 49, 161).

Amandine Brun, R Alcaraz Mor, M Bourrelly, G Dalivoust, G Gazazian, R Boufercha, MP Lehucher-Michel, and I Sari-Minodier. "Radiation protection for surgeons and anesthetists: practices and knowledge before and after training". In: *Journal of Radiological Protection* 38.1 (2018), p. 175 (cit. on p. 16).

Jinzheng Cai, Le Lu, Yuanpu Xie, Fuyong Xing, and Lin Yang. "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function". In: *arXiv preprint arXiv:1707.04912* (2017) (cit. on p. 100).

Miguel A Carreira-Perpiñan and Geoffrey E Hinton. "On contrastive divergence learning." In: *AISTATS*. Vol. 10. Society for Artificial Intelligence and Statistics, 2005, pp. 33–40 (cit. on p. 53).

Gilles Celeux. "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem". In: *Computational statistics quarterly* 2 (1985), pp. 73–82 (cit. on pp. 33, 50, 53, 77, 156, 213).

Gilles Celeux, Didier Chauveau, and Jean Diebolt. *On Stochastic Versions of the EM Algorithm*. Research Report RR-2514. INRIA, 1995 (cit. on pp. 3, 19, 33).

N Chakfé and F Heim. "What do we learn from explant analysis programs?" In: *European Journal of Vascular and Endovascular Surgery* 54.2 (2017), pp. 133–134 (cit. on p. 13).

Nabil Chakfé, Holger Diener, Anne Lejay, Ojan Assadian, Xavier Berard, Jocelyne Caillon, Inge Fourneau, Andor WJM Glaudemans, Igor Koncar, Jes Lindholt, Germano Melissano, Ben R Saleem, Eric Senneville, Riemer H J A Slart, Zoltan Szeberin, Maarit Venermo, Frank Vermassen, and Thomas R Wyss. "Editor's Choice–European Society for Vascular Surgery (ESVS) 2020 Clinical Practice Guidelines on the Management of Vascular Graft and Endograft Infections". In: *European Journal of Vascular and Endovascular Surgery* 59.3 (2020), pp. 339–384 (cit. on p. 12).

Jiawen Chen, Sylvain Paris, and Frédo Durand. "Real-time edge-aware image processing with the bilateral grid". In: *ACM Transactions on Graphics (TOG)* 26.3 (2007), pp. 103–112 (cit. on p. 179).

KyungHyun Cho, Tapani Raiko, and Alexander Ilin. "Parallel tempering is efficient for learning restricted Boltzmann machines". In: *The 2010 International Joint Conference on Neural Networks*. IEEE. 2010, pp. 1–8 (cit. on p. 54).

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 424–432 (cit. on p. 100).

OpenStax CNX. *OpenStax, Anatomy & Physiology*. http://cnx.org/contents/14fb4ad7-39a1-4eee-ab6e-3ef2482e3e22@8.24. 2016 (cit. on pp. 10, 11).

Jean-Baptiste Courbot, Vincent Mazet, Emmanuel Monfrini, and Christophe Collet. "Pairwise Markov fields for segmentation in astronomical hyperspectral images". In: *Signal Processing* 163 (2019), pp. 41–48 (cit. on pp. 39, 49, 57).

Jean-Baptiste Courbot, Emmanuel Monfrini, Vincent Mazet, and Christophe Collet. "Oriented triplet Markov fields". In: *Pattern Recognition Letters* 103 (2018), pp. 16–22 (cit. on pp. 20, 39).

Jean-Baptiste Courbot, Emmanuel Monfrini, Vincent Mazet, and Christophe Collet. "Triplet Markov trees for image segmentation". In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2018, pp. 233–237 (cit. on pp. 5, 70, 149).

Jean-Baptiste Courbot, Edmond Rust, Emmanuel Monfrini, and Christophe Collet. "2-step robust vertebra segmentation". In: *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2015, pp. 157–162 (cit. on pp. 4, 108).

Jean-Baptiste Courbot, Edmond Rust, Emmanuel Monfrini, and Christophe Collet. "Vertebra segmentation based on two-step refinement". In: *Journal of computational surgery* 4.1 (2016), p. 1 (cit. on pp. 126, 127).

Noel Cressie. *Statistics for spatial data*. Wiley Online Library, 1992 (cit. on p. 52).

Noel Cressie and Nicolas Verzelen. "Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields". In: *Computational Statistics & Data Analysis* 52.5 (2008), pp. 2794–2807 (cit. on p. 47).

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. "BM3D image denoising with shape-adaptive principal component analysis". In: 2009 (cit. on p. 60).

Tarun Dave, J Ezhilan, Hardik Vasnawala, and Vinod Somani. "Plaque regression and plaque stabilisation in cardiovascular diseases". In: *Indian Journal of Endocrinology and Metabolism* 17.6 (2013), p. 983 (cit. on p. 9).

Michael W Davidson and Mortimer Abramowitz. "Optical microscopy". In: *Encyclopedia of Imaging Science and Technology* (2002) (cit. on p. 16).

Daniel Delahaye, Supatcha Chaimatanan, and Marcel Mongeau. "Simulated annealing: From basics to applications". In: *Handbook of Metaheuristics*. Springer, 2019, pp. 1–35 (cit. on pp. 30, 155).

Yves Delignon, Abdelwaheb Marzouki, and Wojciech Pieczynski. "Estimation of generalized mixtures and its application in image segmentation". In: *IEEE Transactions on image processing* 6.10 (1997), pp. 1364–1375 (cit. on pp. 127, 185).

Stefanie Demirci, Ali Bigdelou, Lejing Wang, Christian Wachinger, Maximilian Baust, Radhika Tibrewal, Reza Ghotbi, Hans-Henning Eckstein, and Nassir Navab. "3d stent recovery from one x-ray projection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 178–185 (cit. on p. 125).

Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22 (cit. on pp. 3, 19, 32, 155, 213).

Nicolas Depraetere and Martina Vandebroek. "A comparison of variational approximations for fast inference in mixed logit models". In: *Computational Statistics* 32.1 (2017), pp. 93–125 (cit. on p. 172).

Haluk Derin and Howard Elliott. "Modeling and segmentation of noisy and textured images using Gibbs random fields". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (1987), pp. 39–55 (cit. on pp. 50, 51, 76).

Pierre A Devijver. "Baum's forward-backward algorithm revisited". In: *Pattern Recognition Letters* 3.6 (1985), pp. 369–373 (cit. on pp. 154, 213).

Claude R Dietrich and Garry N Newsam. In: *SIAM Journal on Scientific Computing* 18.4 (1997), pp. 1088–1107 (cit. on pp. 158, 160).

Tamara Dimitrova and Ljupco Kocarev. "Graphical Models Over Heterogeneous Domains and for Multilevel Networks". In: *IEEE Access* 6 (2018), pp. 69682–69701 (cit. on p. 43).

Nicolas Dobigeon, Alfred O Hero, and Jean-Yves Tourneret. "Hierarchical Bayesian sparse image reconstruction with application to MRFM". In: *IEEE transactions on Image Processing* 18.9 (2009), pp. 2059–2070 (cit. on p. 35).

Mathias Drton and Marloes H Maathuis. "Structure learning in graphical modeling". In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 365–393 (cit. on pp. 31, 43).

Jean-Baptiste Durand, Paulo Goncalves, and Yann Guédon. "Computational methods for hidden Markov tree models-An application to wavelet trees". In: *IEEE Transactions on Signal Processing* 52.9 (2004), pp. 2551–2560 (cit. on pp. 68, 69).

Jean-Baptiste Durand and Paulo Gonçalves. *Statistical inference for hidden Markov tree models and application to wavelet trees*. Research Report RR-4248. INRIA, 2001 (cit. on pp. 29, 37).

Jim C Elliott and SD Dover. "X-ray microtomography". In: *Journal of Microscopy* 126.2 (1982), pp. 211–213 (cit. on p. 16).

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger. "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nature methods* 16.1 (2019), pp. 67–70 (cit. on p. 101).

Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. "Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks". In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 64–76 (cit. on pp. 103, 118).

Paul Fieguth. *Statistical image processing and multidimensional modeling*. Springer, 2010 (cit. on p. 15).

Roger Fjortoft, Yves Delignon, Wojciech Pieczynski, Marc Sigelle, and Florence Tupin. "Unsupervised classification of radar images using hidden Markov chains and hidden Markov random fields". In: *IEEE Transactions on geoscience and remote sensing* 41.3 (2003), pp. 675–686 (cit. on p. 108).

Brian P Flannery, Harry W Deckman, Wayne G Roberge, and Kevin L D'Amico. "Three-dimensional X-ray microtomography". In: *Science* 237.4821 (1987), pp. 1439–1444 (cit. on p. 16).

Francky Fouedjio, Nicolas Desassis, and Thomas Romary. "Estimation of space deformation model for non-stationary random functions". In: *Spatial Statistics* 13 (2015), pp. 45–61 (cit. on p. 149).

Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. "Multiscale vessel enhancement filtering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 1998, pp. 130–137 (cit. on pp. 125, 128, 131).

Brendan J Frey. "Extending factor graphs so as to unify directed and undirected graphical models". In: *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*. 2003 (cit. on p. 28).

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer, 2001 (cit. on p. 50).

Steven Greg Friedman. *A history of vascular surgery*. Wiley Online Library, 2005 (cit. on p. 10).

Geir-Arne Fuglstad, Finn Lindgren, Daniel Simpson, and Håvard Rue. "Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy". In: *Statistica Sinica* (2015), pp. 115–133 (cit. on pp. 48, 158).

Kunihiko Fukushima and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285 (cit. on p. 101).

Hugo Gangloff, Jean-Baptiste Courbot, Emmanuel Monfrini, and Christophe Collet. "Segmentation non-supervisée dans les champs de Markov couples gaussiens". In: *Colloque GRETSI*. 2019 (cit. on p. 21).

Hugo Gangloff, Jean-Baptiste Courbot, Emmanuel Monfrini, and Christophe Collet. "Spatial Triplet Markov Trees for auxiliary variational inference in Spatial Bayes Networks". In: *Stochastic Modeling Techniques and Data Analysis international conference (SMTDA'20)*. 2020. In press (cit. on p. 21).

Hugo Gangloff, Emmanuel Monfrini, Christophe Collet, and Nabil Chakfé. "Segmentation de stents dans des données médicales à rayons-X corrompues par les artéfacts". In: *Colloque GRETSI*. 2019 (cit. on p. 21).

Hugo Gangloff, Emmanuel Monfrini, Christophe Collet, and Nabil Chakfé. "Unsupervised segmentation of stents corrupted by artifacts in medical X-ray images". In: *2020 International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2020. In press (cit. on p. 21).

Hugo Gangloff, Emmanuel Monfrini, Mohamed Zied Ghariani, Mickaël Ohana, Christophe Collet, and Nabil Chakfé. "Improved Centerline Tracking for new descriptors of atherosclerotic aortas". In: *2020 International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2020. In press (cit. on p. 21).

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010 (cit. on p. 158).

Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741 (cit. on pp. 4, 28, 30, 34, 38, 54, 153, 155, 213).

Z. Ghahramani and M. I. Jordan. "Factorial hidden Markov models". In: *Advances in Neural Information Processing Systems*. 1996, pp. 472–478 (cit. on pp. 87, 108).

Zoubin Ghahramani. "Graphical models: parameter learning". In: *Handbook of brain theory and neural networks* 2 (2002), pp. 486–490 (cit. on p. 35).

Nathalie Giordana and Wojciech Pieczynski. "Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.5 (1997), pp. 465–475 (cit. on p. 108).

Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. "Parallel gibbs sampling: From colored fields to thin junction trees". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 324–332 (cit. on p. 28).

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016 (cit. on pp. 24, 26, 39, 98, 100).

Matthew R Gormley and Jason Eisner. "Structured belief propagation for NLP". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Tutorial Abstracts*. 2015, pp. 5–6 (cit. on pp. 29, 30).

Ivan Gorynin, Laurent Crelier, Hugo Gangloff, Emmanuel Monfrini, and Wojciech Pieczynski. "Performance comparison across hidden, pairwise and triplet Markov models' estimators". In: *International Journal of Mathematical and Computational Methods* 1 (2016) (cit. on p. 21).

Ivan Gorynin, Hugo Gangloff, Emmanuel Monfrini, and Wojciech Pieczynski. "Assessing the segmentation performance of pairwise and triplet Markov models". In: *Signal Processing* 145 (2018), pp. 183–192 (cit. on pp. 2, 3, 19, 21, 39, 150).

Ivan Gorynin, Emmanuel Monfrini, and Wojciech Pieczynski. "Pairwise Markov models for stock index forecasting". In: *25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 2041–2045 (cit. on p. 39).

Aditya Gupta and Bhuwan Dhingra. "Stock market prediction using hidden markov models". In: *2012 Students Conference on Engineering and Systems*. IEEE. 2012, pp. 1–4 (cit. on p. 37).

Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. "Histopathological image analysis: A review". In: *IEEE reviews in biomedical engineering* 2 (2009), pp. 147–171 (cit. on p. 100).

Mark Haidekker. *Advanced biomedical image analysis*. John Wiley & Sons, 2011 (cit. on p. 128).

John Hammersley and Clifford Peter. "Markov field on finite graphs and lattices". Unpublished manuscript. 1971 (cit. on p. 27).

Houda Hanzouli, Jérôme Lapuyade-Lahorgue, Emmanuel Monfrini, Gaspar Delso, Wojciech Pieczynski, Dimitris Visvikis, and Mathieu Hatt. "PECT/CT image denoising and segmentation based on a multi observation and multi scale Markov tree model". In: *NSS/MIC 2013 : Nuclear Science Symposium and Medical Imaging Conference*. IEEE, 2013 (cit. on p. 37).

Houda Hanzouli-Ben Salah, Jerome Lapuyade-Lahorgue, Julien Bert, Didier Benoit, Philippe Lambin, Angela Van Baardwijk, Emmanuel Monfrini, Wojciech Pieczynski, Dimitris Visvikis, and Mathieu Hatt. "A framework based on hidden Markov trees for multimodal PET/CT image co-segmentation". In: *Medical physics* 44.11 (2017), pp. 5835–5848 (cit. on p. 39).

Intisar Rizwan I Haque and Jeremiah Neubert. "Deep learning approaches to biomedical image segmentation". In: *Informatics in Medicine Unlocked* 18 (2020), p. 100297 (cit. on p. 100).

Geoffrey E Hinton. "A practical guide to training restricted Boltzmann machines". In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619 (cit. on p. 101).

Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 101).

Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015) (cit. on p. 103).

Pejman Jabehdar Maralani, Nicola Schieda, Elizabeth M Hecht, Harold Litt, Nicole Hindman, Chinthaka Heyn, Matthew S Davenport, Greg Zaharchuk, Christopher P Hess, and Jeffrey Weinreb. "MRI safety and devices: An update and expert consensus". In: *Journal of Magnetic Resonance Imaging* 51.3 (2020), pp. 657–674 (cit. on p. 16).

Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 11–19 (cit. on p. 101).

Frederick Jelinek, Lalit Bahl, and Robert Mercer. "Design of a linguistic statistical decoder for the recognition of continuous speech". In: *IEEE Transactions on Information Theory* 21.3 (1975), pp. 250–256 (cit. on p. 36).

Zaiping Jing, Huajuan Mao, and Weihui Dai. *Endovascular Surgery and Devices*. Springer, 2018 (cit. on pp. 11, 12).

Hiroyuki Jinnouchi, Sho Torii, Matthew Kutyna, Atsushi Sakamoto, Frank D Kolodgie, Aloke V Finn, and Renu Virmani. "Micro–Computed Tomography Demonstration of Multiple Plaque Ruptures in a Single Individual Presenting With Sudden Cardiac Death". In: *Circulation: Cardiovascular Imaging* 11.10 (2018), e008331 (cit. on p. 141).

Michael I Jordan. "Graphical models". In: *Statistical science* 19.1 (2004), pp. 140–155 (cit. on p. 29).

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An introduction to variational methods for graphical models". In: *Machine learning* 37.2 (1999), pp. 183–233 (cit. on pp. 30, 33, 86, 87).

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36 (2017), pp. 61–78 (cit. on pp. 3, 100, 117, 120).

Jörg Hendrik Kappes, Paul Swoboda, Bogdan Savchynskyy, Tamir Hazan, and Christoph Schnörr. "Multicuts and perturb & MAP for probabilistic graph clustering". In: *Journal of Mathematical Imaging and Vision* 56.2 (2016), pp. 221–237 (cit. on p. 30).

Zoltan Kato and Josiane Zerubia. "Markov random fields in image segmentation". In: *Foundations and Trends in Signal Processing* 5.1–2 (2012), pp. 1–155 (cit. on pp. 2, 39, 48, 49, 60).

Kentaro Kinebuchi, D Darian Muresan, and Thomas W Parks. "Image interpolation using wavelet based hidden Markov trees". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. Vol. 3. IEEE. 2001, pp. 1957–1960 (cit. on p. 37).

Diederik P Kingma. "Variational inference & deep learning: A new synthesis". PhD thesis. 2017 (cit. on p. 89).

Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 101).

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. "Optimization by simulated annealing". In: *Science* 220.4598 (1983), pp. 671–680 (cit. on p. 30).

William Kleiber. "High resolution simulation of nonstationary Gaussian random fields". In: *Computational Statistics & Data Analysis* 101 (2016), pp. 277–288 (cit. on p. 149).

Almar Klein, J Adam van der Vliet, Luuk J Oostveen, Yvonne Hoogeveen, Leo J Schultze Kool, W Klaas Jan Renema, and Cornelis H Slump. "Automatic segmentation of the wire frame of stent grafts from CT data". In: *Medical image analysis* 16.1 (2012), pp. 127–139 (cit. on p. 125).

Maaike A Koenrades, Esmeralda M Struijs, Almar Klein, Hendrik Kuipers, Michel MPJ Reijnen, Cornelis H Slump, and Robert H Geelkerken. "Quantitative Stent Graft Motion in ECG Gated CT by Image Registration and Segmentation: In Vitro Validation and Preliminary Clinical Results". In: *European journal of vascular and endovascular surgery* 58.5 (2019), pp. 746–755 (cit. on p. 150).

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009 (cit. on p. 24).

Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. "MRF energy minimization and beyond via dual decomposition". In: *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2010), pp. 531–552 (cit. on p. 30).

Daisuke Komura and Shumpei Ishikawa. "Machine learning methods for histopathological image analysis". In: *Computational and structural biotechnology journal* 16 (2018), pp. 34–42 (cit. on p. 100).

Philipp Krähenbühl and Vladlen Koltun. "Efficient inference in fully connected crfs with gaussian edge potentials". In: *Advances in Neural Information Processing Systems*. 2011, pp. 109–117 (cit. on pp. 7, 100, 104–106, 120, 149, 213).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105 (cit. on pp. 98, 100, 101).

Dirk P Kroese and Zdravko I Botev. "Spatial process generation". In: *arXiv preprint arXiv:1308.0399* (2013) (cit. on pp. 160, 161).

Salomé H. Kuntz, Hiroyuki Jinnouchi, Sho Torii, Anne Cornelissen, Atsushi Sakamoto, Yu Sato, Maria E. Romero, Frank Kolodgie, Aloke V. Finn, Adeline Schwein, Mickael Ohana, Anne Lejay, Nabil Chakfé, and Renu Virmani. "Co-registration of peripheral atherosclerotic plaques assessed by conventional CT-angiography, micro-CT and histology in CLTI patients". In: *European Journal of Vascular and Endovascular Surgery* (2020). In press (cit. on p. 22).

J-M Laferté, Patrick Pérez, and Fabrice Heitz. "Discrete Markov image modeling and inference on the quadtree". In: *IEEE Transactions on image processing* 9.3 (2000), pp. 390–404 (cit. on pp. 2, 29, 37, 69).

Pierre Lanchantin, Jérôme Lapuyade-Lahorgue, and Wojciech Pieczynski. "Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise". In: *Signal Processing* 91.2 (2011), pp. 163–175 (cit. on pp. 2, 19, 39, 108).

Georg Langs, Nikos Paragios, Pascal Desgranges, Alain Rahmouni, and Hicham Kobeiter. "Learning deformation and structure simultaneously: In situ endograft deformation analysis". In: *Medical image analysis* 15.1 (2011), pp. 12–21 (cit. on pp. 125, 150).

Steffen L Lauritzen and David J Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2 (1988), pp. 157–194 (cit. on pp. 30, 68, 86, 91).

Anne Lejay, Benjamin Colvard, Louis Magnus, Delphine Dion, Yannick Georg, Julie Papillon, Fabien Thaveau, Bernard Geny, Lee Swanström, Fréderic Heim, and Nabil Chakfé. "Explanted vascular and endovascular graft analysis: where do we stand and what should we do?" In: *European Journal of Vascular and Endovascular Surgery* 55.4 (2018), pp. 567–576 (cit. on pp. 13, 124).

David Lesage, Elsa D Angelini, Isabelle Bloch, and Gareth Funka-Lea. "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes". In: *Medical image analysis* 13.6 (2009), pp. 819–845 (cit. on p. 125).

Stan Z Li. *Markov random field modeling in image analysis*. Springer, 2009 (cit. on pp. 37, 45, 49, 60).

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes". In: *IEEE transactions on medical imaging* 37.12 (2018), pp. 2663–2674 (cit. on p. 100).

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88 (cit. on p. 101).

Qiang Liu and Alexander Ihler. "Variational algorithms for marginal MAP". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3165–3200 (cit. on p. 35).

Michelle Livne, Jana Rieger, Orhun Utku Aydin, Abdel Aziz Taha, Ela Marie Akay, Tabea Kossen, Jan Sobesky, John D Kelleher, Kristian Hildebrand, Dietmar Frey, and Vince I Madai. "A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease". In: *Frontiers in neuroscience* 13 (2019) (cit. on p. 126).

H-A Loeliger. "An introduction to factor graphs". In: *IEEE Signal Processing Magazine* 21.1 (2004), pp. 28–41 (cit. on p. 29).

Jose Marroquin, Sanjoy Mitter, and Tomaso Poggio. "Probabilistic solution of ill-posed problems in computational vision". In: *Journal of the American statistical association* 82.397 (1987), pp. 76–89 (cit. on pp. 35, 38, 45, 56, 57, 156, 213).

Odyssée Merveille, Hugues Talbot, Laurent Najman, and Nicolas Passat. "Tubular Structure Filtering by Ranking Orientation Responses of Path Operators". In: *European Conference on Computer Vision (ECCV)*. Vol. 8690. Springer, 2014, pp. 203–218 (cit. on pp. 125, 131).

Max Mignotte, Christophe Collet, Patrick Pérez, and Patrick Bouthemy. "Three-class Markovian segmentation of high-resolution sonar images". In: *Computer Vision and Image Understanding* 76.3 (1999), pp. 191–204 (cit. on p. 39).

Tom Minka. *Discriminative models, not discriminative training*. Technical Report MSR-TR-2005-144. Microsoft Research, 2005 (cit. on p. 35).

Emmanuel Monfrini, J Lecomte, F Desbouvries, and W Pieczynski. "Image and signal restoration using pairwise Markov trees". In: *IEEE Workshop on Statistical Signal Processing*. IEEE. 2003, pp. 174–177 (cit. on pp. 68, 69).

Emmanuel Monfrini and Wojciech Pieczynski. "Estimation de mélanges généralisés dans les arbres de Markov cachés, application à la segmentation des images de cartons d'orgue de barbarie". In: *TS. Traitement du signal* 22.2 (2005) (cit. on pp. 127, 185).

Lia Morra, Silvia Delsanto, and Loredana Correale. *Artificial Intelligence in Medical Imaging: From Theory to Clinical Practice*. CRC Press, 2019 (cit. on p. 103).

Robin Morris, Xavier Descombes, and Josiane Zerubia. "Fully Bayesian image segmentation-an engineering perspective". In: *Proceedings of International Conference on Image Processing*. Vol. 3. IEEE. 1997, pp. 54–57 (cit. on p. 35).

Raphaël Mourad, Christine Sinoquet, N. Zhang, T. Liu, and Philippe Leray. "A survey on latent tree models and applications". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 157–203 (cit. on p. 37).

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012 (cit. on pp. 2, 24, 27, 86, 112, 145).

Radford M Neal. "Sampling from multimodal distributions using tempered transitions". In: *Statistics and computing* 6.4 (1996), pp. 353–366 (cit. on p. 54).

Andrew Y Ng and Michael I Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". In: *Advances in neural information processing systems*. 2002, pp. 841–848 (cit. on pp. 35, 149).

W Kirt Nichols and Wei Wei. "Has open vascular surgery disappeared?" In: *Missouri Medicine* 108.3 (2011), p. 182 (cit. on p. 11).

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer, 2006 (cit. on p. 31).

Masoud S Nosrati and Ghassan Hamarneh. "Incorporating prior knowledge in medical image segmentation: a survey". In: *arXiv preprint arXiv:1607.01092* (2016) (cit. on p. 152).

Alexey A Novikov, David Major, Maria Wimmer, Dimitrios Lenis, and Katja Bühler. "Deep sequential segmentation of organs in volumetric medical scans". In: *IEEE transactions on medical imaging* 38.5 (2018), pp. 1207–1215 (cit. on pp. 3, 100).

Chinedu Innocent Nwoye, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos". In: *International journal of computer assisted radiology and surgery* 14.6 (2019), pp. 1059–1067 (cit. on p. 19).

Douglas Nychka, Dorit Hammerling, Mitchell Krock, and Ashton Wiens. "Modeling and emulation of nonstationary Gaussian fields". In: *Spatial statistics* 28 (2018), pp. 21–38 (cit. on p. 149).

Mickaël Ohana, Soraya El Ghannudi, Elie Girsowicz, Anne Lejay, Yannick Georg, Fabien Thaveau, Nabil Chakfe, and Catherine Roy. "Detailed cross-sectional study of 60 superficial femoral artery occlusions: morphological quantitative analysis can lead to a new classification". In: *Cardiovascular diagnosis and therapy* 4.2 (2014), p. 71 (cit. on p. 13).

Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation". In: *IEEE transactions on medical imaging* 37.2 (2017), pp. 384–395 (cit. on p. 152).

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. "Attention u-net: Learning where to look for the pancreas". In: *arXiv preprint arXiv:1804.03999* (2018) (cit. on pp. 101, 152).

V. Olariu, D. Coca, Stephen A. Billings, Peter Tonge, Paul Gokhale, Peter W Andrews, and Visakan Kadirkamanathan. "Modified variational Bayes EM estimation of hidden Markov tree model of cell lineages". In: *Bioinformatics* 25.21 (2009), pp. 2824–2830 (cit. on p. 87).

Sylvain Paris and Frédo Durand. "A fast approximation of the bilateral filter using a signal processing approach". In: *European conference on computer vision*. Springer. 2006, pp. 568–580 (cit. on pp. 107, 179).

Sylvain Paris, Pierre Kornprobst, Jack Tumblin, and Frédo Durand. *Bilateral filtering: Theory and applications*. Now Publishers Inc, 2009 (cit. on pp. 104, 179).

Hyoung Suk Park, Yong Eun Chung, and Jin Keun Seo. "Computed tomographic beam-hardening artefacts: mathematical characterization and analysis". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2043 (2015) (cit. on p. 18).

Joon Park and Roderic S Lakes. *Biomaterials: an introduction*. Springer, 2007 (cit. on p. 11).

Youngsuk Park, David Hallac, Stephen Boyd, and Jure Leskovec. "Learning the network structure of heterogeneous data via pairwise exponential Markov random fields". In: *Proceedings of machine learning research* 54 (2017), p. 1302 (cit. on p. 43).

Mark S Pearce, Jane A Salotti, Mark P Little, Kieran McHugh, Choonsik Lee, Kwang Pyo Kim, Nicola L Howe, Cecile M Ronckers, Preetha Rajaraman, Alan W Craft, Louise Parker, and Amy Berrington de Gonzalez. "Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study". In: *The Lancet* 380.9840 (2012), pp. 499–505 (cit. on p. 16).

Judea Pearl. "Reverend Bayes on inference engines: A distributed hierarchical approach". In: *Proceedings of the National Conference on Artificial Intelligence*. 1982, pp. 133–136 (cit. on p. 29).

Marcelo Pereyra, Nicolas Dobigeon, Hadj Batatia, and Jean-Yves Tourneret. "Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm". In: *IEEE Transactions on Image Processing* 22.6 (2013), pp. 2385–2397 (cit. on pp. 45, 47).

David Perrin, Pierre Badel, Laurent Orgeas, Christian Geindreau, Sabine rolland du Roscoat, Jean-Noël Albertini, and Stéphane Avril. "Patient-specific simulation of endovascular repair surgery with tortuous aneurysms requiring flexible stent-grafts". In: *Journal of the mechanical behavior of biomedical materials* 63 (2016), pp. 86–99 (cit. on p. 125).

Andrija Petrović, Mladen Nikolić, Miloš Jovanović, and Boris Delibašić. "Gaussian Conditional Random Fields for Classification". In: *arXiv preprint arXiv:1902.00045* (2019) (cit. on pp. 42, 149).

Géraldine Pichot. *Algorithms for stationary Gaussian random field generation*. Technical Report RT-0484. INRIA Paris, Dec. 2016 (cit. on p. 161).

Wojciech Pieczynski. "Arbres de Markov couple". In: *Comptes Rendus Mathématique* 335.1 (2002), pp. 79–82 (cit. on p. 69).

Wojciech Pieczynski, Julien Bouvrais, and Christophe Michel. "Estimation of generalized mixture in the case of correlated sensors". In: *IEEE Transactions on Image Processing* 9.2 (2000), pp. 308–312 (cit. on pp. 127, 185).

Wojciech Pieczynski, Cédric Hulard, and Thomas Veit. "Triplet Markov chains in hidden signal restoration". In: *Image and Signal Processing for Remote Sensing VIII*. Vol. 4885. International Society for Optics and Photonics. 2003, pp. 58–68 (cit. on p. 39).

Wojciech Pieczynski and Abdel-Nasser Tebbache. "Pairwise Markov random fields and segmentation of textured images". In: *Machine graphics and vision* 9.3 (2000), pp. 705–718 (cit. on pp. 2, 19, 39).

V PW Scholtes, J PPM de Vries, L M Catanzariti, D PV de Kleijn, F L Moll, G J de Borst, and G Pasterkamp. "Biobanking in atherosclerotic disease, opportunities and pitfalls". In: *Current cardiology reviews* 7.1 (2011), pp. 9–14 (cit. on p. 19).

Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic. "Continuous Conditional Random Fields for Regression in Remote Sensing." In: *Proceedings of the 19th European Conference on Artificial Intelligence*. 2010, pp. 809–814 (cit. on pp. 42, 149).

Juliette Raffort, Cédric Adam, Marion Carrier, Ali Ballaith, Raphael Coscas, Elixène Jean-Baptiste, Réda Hassen-Khodja, Nabil Chakfé, and Fabien Lareyre. "Artificial intelligence in abdominal aortic aneurysm". In: *Journal of Vascular Surgery* (2020) (cit. on p. 13).

Mahmoud Rafieian-Kopaei, Mahbubeh Setorki, Monir Doudi, Azar Baradaran, and Hamid Nasri. "Atherosclerosis: process, indicators, risk factors and new hopes". In: *International journal of preventive medicine* 5.8 (2014), p. 927 (cit. on p. 9).

Hariharan Ravishankar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Prasad Sudhakar, and Vivek Vaidya. "Learning and incorporating shape models for semantic segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 203–211 (cit. on p. 152).

Mark D Risser. "Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches". In: *arXiv preprint arXiv:1610.02447* (2016) (cit. on p. 158).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on pp. 7, 101, 102).

Gina Delia Roque-Torres. "Application of Micro-CT in Soft Tissue Specimen Imaging". In: *Micro-computed Tomography (micro-CT) in Medicine and Engineering*. Springer, 2020, pp. 139–170 (cit. on p. 16).

Michael E Rosenfeld. "An overview of the evolution of the atherosclerotic plaque: from fatty streak to plaque rupture and thrombosis". In: *Zeitschrift für Kardiologie* 89.7 (2000), pp. VII2–VII6 (cit. on p. 9).

Oleksii Rubel, Vladimir Lukin, and Karen Egiazarian. "Additive Spatially Correlated Noise Suppression by Robust Block Matching and Adaptive 3D Filtering". In: *Journal of Imaging Science and Technology* 62.6 (2018), pp. 60401–1 (cit. on p. 60).

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005 (cit. on pp. 3, 4, 19, 28, 42, 47–49, 158).

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536 (cit. on p. 103).

Hans Sagan. *Space-filling curves*. Springer, 2012 (cit. on p. 108).

Ruslan R Salakhutdinov. "Learning in Markov random fields using tempered transitions". In: *Advances in neural information processing systems*. 2009, pp. 1598–1606 (cit. on pp. 28, 56).

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2017, pp. 379–387 (cit. on p. 103).

GD Saul and HM Gerard. "Physical fitness, dynamic extra-arterial pressures, and the pathogenesis and distribution of atherosclerosis". In: *Medical hypotheses* 36.3 (1991), pp. 228–237 (cit. on p. 10).

Lawrence K Saul, Tommi Jaakkola, and Michael I Jordan. "Mean field theory for sigmoid belief networks". In: *Journal of artificial intelligence research* 4 (1996), pp. 61–76 (cit. on p. 30).

Francoise Schmitt, Max Mignotte, Christophe Collet, and Pierre Thourel. "Estimation of noise parameters on sonar images". In: *Statistical and Stochastic Methods for Image Processing*. Vol. 2823. International Society for Optics and Photonics. 1996, pp. 2–12 (cit. on p. 39).

Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1 (2019), p. 60 (cit. on pp. 103, 111).

Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. "Deep neural network models for computational histopathology: A survey". In: *arXiv preprint arXiv:1912.12378* (2019) (cit. on pp. 3, 100).

Mario Stanke, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack. "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources". In: *BMC bioinformatics* 7.1 (2006), p. 62 (cit. on p. 36).

W Charles Sternbergh III and Samuel R Money. "Hospital cost of endovascular versus open repair of abdominal aortic aneurysms: a multicenter study". In: *Journal of vascular surgery* 31.2 (2000), pp. 237–244 (cit. on p. 11).

Julien Stoehr. "A review on statistical inference methods for discrete Markov random fields". In: *arXiv preprint arXiv:1704.03331* (2017) (cit. on p. 28).

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248 (cit. on p. 103).

Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017 (cit. on p. 15).

Charles Sutton and Andrew McCallum. "An introduction to conditional random fields". In: *Foundations and Trends in Machine Learning* 4.4 (2012), pp. 267–373 (cit. on pp. 32, 35, 98).

Tijmen Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1064–1071 (cit. on pp. 3, 19, 33).

Tomoki Toda. "Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis". In: *Hidden Markov Models, Theory and Applications*. InTech, 2011, pp. 131–150 (cit. on p. 36).

Sho Torii, Jihad A Mustapha, Jagat Narula, Hiroyoshi Mori, Fadi Saab, Hiroyuki Jinnouchi, Kazuyuki Yahagi, Atsushi Sakamoto, Maria E Romero, Navneet Narula, Frank D Kolodgie, Renu Virmani, and Aloke V Finn. "Histopathologic characterization of peripheral arteries in subjects with abundant risk factors: correlating imaging with pathology". In: *JACC: Cardiovascular Imaging* 12.8 Part 1 (2019), pp. 1501–1513 (cit. on pp. 141, 142).

Cornelia Vacar and Jean-François Giovannelli. "Unsupervised joint deconvolution and segmentation method for textured images: a Bayesian approach and an advanced sampling algorithm". In: *EURASIP Journal on Advances in Signal Processing* 2019.1 (2019), p. 17 (cit. on pp. 35, 149).

Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellapa. "Gaussian conditional random field network for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3224–3233 (cit. on pp. 42, 149).

Joost M Verburg and Joao Seco. "CT metal artifact reduction method correcting for beam hardening and missing projections". In: *Physics in Medicine & Biology* 57.9 (2012), p. 2803 (cit. on pp. 126, 129).

Andrew Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE transactions on Information Theory* 13.2 (1967), pp. 260–269 (cit. on p. 29).

Martin J Wainwright and Michael I Jordan. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1.1-2 (2008), pp. 1–305 (cit. on pp. 24, 25).

Chaohui Wang, Nikos Komodakis, and Nikos Paragios. "Markov random field modeling, inference & learning in computer vision & image understanding: A survey". In: *Computer Vision and Image Understanding* 117.11 (2013), pp. 1610–1627 (cit. on p. 39).

Wim Wiegerinck. "Variational Approximations between Mean Field Theory and the Junction Tree Algorithm". In: *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. UAI'00. Stanford, California: Morgan Kaufmann Publishers Inc., 2000, pp. 626–633. ISBN: 1558607099 (cit. on pp. 87, 172).

Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. "'All models are wrong...': an introduction to model uncertainty". In: *Statistica Neerlandica* 66.3 (2012), pp. 217–236 (cit. on pp. 127, 185).

Kazuyuki Yahagi, Frank D Kolodgie, Fumiyuki Otsuka, Aloke V Finn, Harry R Davis, Michael Joner, and Renu Virmani. "Pathophysiology of native coronary, vein graft, and in-stent atherosclerosis". In: *Nature Reviews Cardiology* 13.2 (2016), p. 79 (cit. on p. 141).

M. Yahiaoui, Emmanuel Monfrini, and Bernadette Dorizzi. "Implementation of unsupervised statistical methods for low-quality iris segmentation". In: *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*. IEEE. 2014, pp. 566–573 (cit. on p. 126).

Meriem Yahiaoui, Emmanuel Monfrini, and Bernadette Dorizzi. "Markov Chains for unsupervised segmentation of degraded NIR iris images for person recognition". In: *Pattern Recognition Letters* 82 (2016), pp. 116–123 (cit. on p. 108).

Dong Yang, Qiaoying Huang, Leon Axel, and Dimitris N. Metaxas. "Multi-component deformable models coupled with 2D-3D U-Net for automated probabilistic segmentation of cardiac walls and blood". In: *IEEE 15th International Symposium on Biomedical Imaging* (2018), pp. 479–483 (cit. on p. 100).

Laurent Younes. "On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates". In: *Stochastics: An International Journal of Probability and Stochastic Processes* 65.3-4 (1999), pp. 177–228 (cit. on p. 33).

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026 (cit. on pp. 4, 19, 30, 86, 89).

Nevin Lianwen Zhang and David Poole. "A simple approach to Bayesian network computations". In: *Proceedings of the biennial conference-Canadian society for computational studies of intelligence*. 1994, pp. 171–178 (cit. on p. 29).

Yanbo Zhang and Hengyong Yu. "Convolutional neural network based metal artifact reduction in x-ray computed tomography". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1370–1381 (cit. on pp. 126, 129).

Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. "3D brain tumor segmentation through integrating multiple 2D FCNNs". In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 191–203 (cit. on p. 100).

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. "Conditional random fields as recurrent neural networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1529–1537 (cit. on p. 101).

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A nested u-net architecture for medical image segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11 (cit. on p. 101).

Piotr Zwiernik. *Semialgebraic statistics and latent tree models*. CRC Press, 2015 (cit. on p. 37).

List of Figures

List of Tables

List of Algorithms