

Inférence sur variables réelles par un modèle d'Ising

Victorin MARTIN

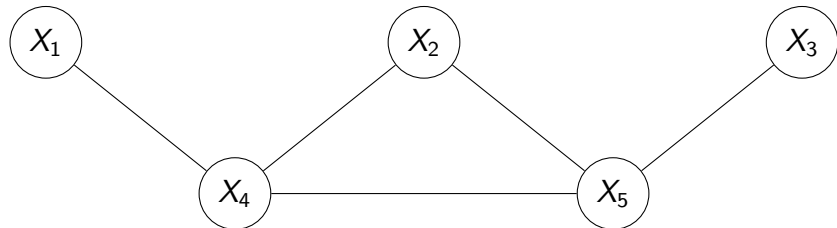


Plan de présentation

- 1 Introduction/Problématique.
- 2 Construction du modèle d'Ising.
- 3 Algorithme d'inférence.
- 4 Expérimentations numériques.

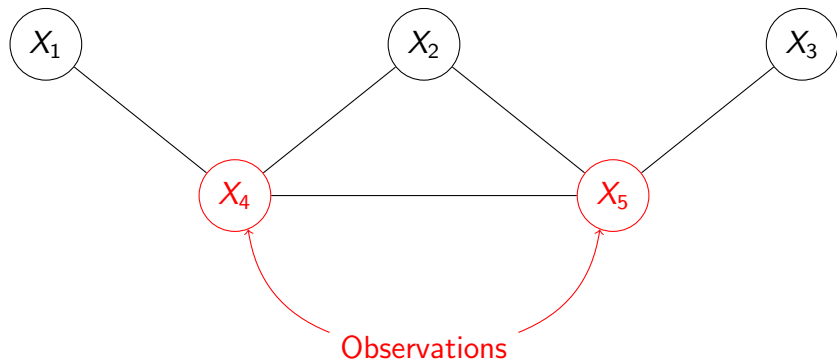
Champ Markovien de v.a.r

- Indépendance conditionnelle
- Densité jointe : $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$ (Théorème de Hammersley-Clifford)



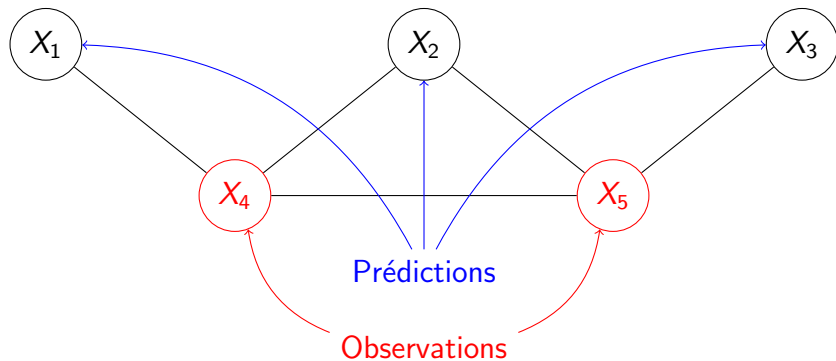
Champ Markovien de v.a.r

- Indépendance conditionnelle
- Densité jointe : $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$ (Théorème de Hammersley-Clifford)



Champ Markovien de v.a.r

- Indépendance conditionnelle
- Densité jointe : $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$ (Théorème de Hammersley-Clifford)



Deux problèmes à résoudre

- A partir d'un historique $\{\mathbf{X}^k\}_{k \in \{1..M\}}$, calibrer le modèle i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.
- Réaliser l'inférence conditionnelle aux observations.

Problématique

Deux problèmes à résoudre

- A partir d'un historique $\{\mathbf{X}^k\}_{k \in \{1..M\}}$, calibrer le modèle i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.
- Réaliser l'inférence conditionnelle aux observations.

L'inférence : un problème théoriquement trivial

- Densité $f_k(X_k) = \int_{\mathbf{X} \setminus X_k} \prod_{(ij)} \varphi_{ij}(X_i, X_j) d\mathbf{X}$.
- Supposons \mathbf{X} contenant n variables binaires $\rightarrow 2^{n-1}$ termes dans la somme...

Problématique

Deux problèmes à résoudre

- A partir d'un historique $\{\mathbf{X}^k\}_{k \in \{1..M\}}$, calibrer le modèle i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.
- Réaliser l'inférence conditionnelle aux observations.

L'inférence : un problème théoriquement trivial

- Densité $f_k(X_k) = \int_{\mathbf{X} \setminus X_k} \prod_{(ij)} \varphi_{ij}(X_i, X_j) d\mathbf{X}$.
- Supposons \mathbf{X} contenant n variables binaires $\rightarrow 2^{n-1}$ termes dans la somme...

Cas particulier : $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$

- Calibration simple ; estimation de μ et Σ .
- Inférence exacte tractable.

Problématique

Deux problèmes à résoudre

- A partir d'un historique $\{\mathbf{X}^k\}_{k \in \{1..M\}}$, calibrer le modèle i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.
- Réaliser l'inférence conditionnelle aux observations.

L'inférence : un problème théoriquement trivial

- Densité $f_k(X_k) = \int_{\mathbf{X} \setminus X_k} \prod_{(ij)} \varphi_{ij}(X_i, X_j) d\mathbf{X}$.
- Supposons \mathbf{X} contenant n variables binaires $\rightarrow 2^{n-1}$ termes dans la somme...

Cas particulier : $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$

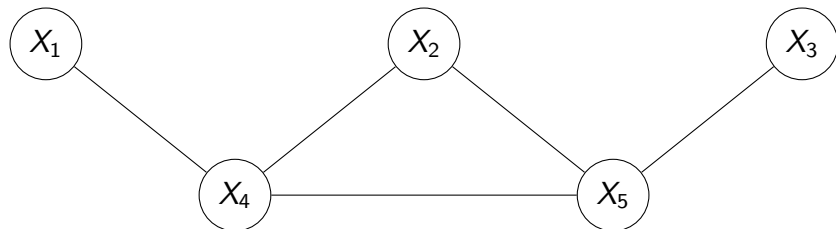
- Calibration simple ; estimation de μ et Σ .
- Inférence exacte tractable.

Cas général

- Calibration et inférence sont fait de manière approchée.
- Les 2 approximations s'influencent fortement...

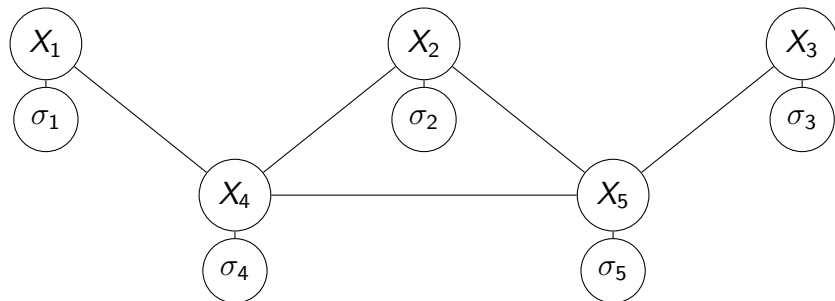
Notre approximation

MRF de variables réelles

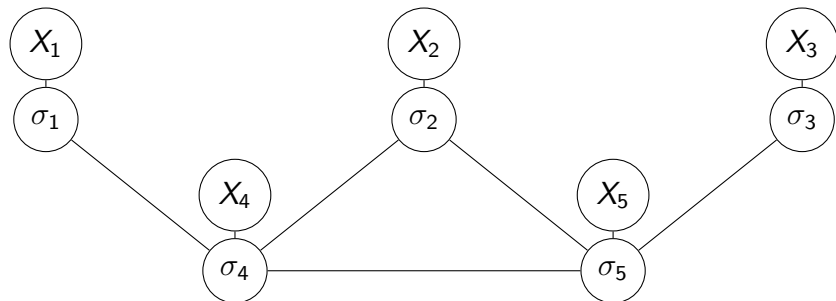


Notre approximation

MRF de variables réelles



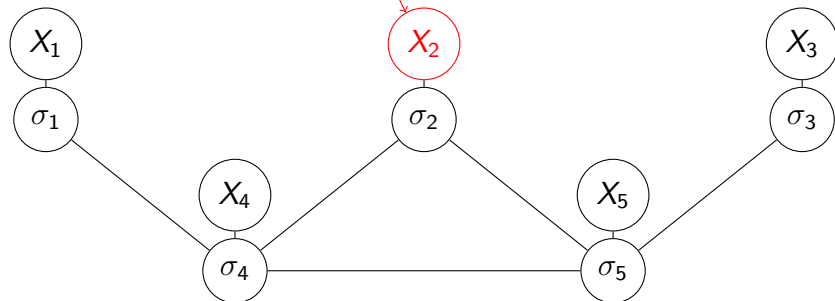
Notre approximation



MRF de variables binaires

Notre approximation

On observe $X_2 = x$.

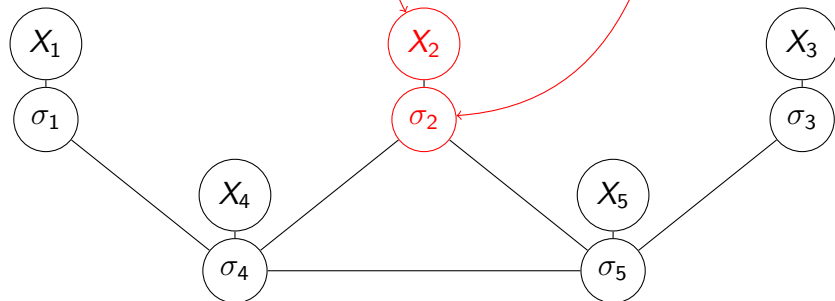


Notre approximation

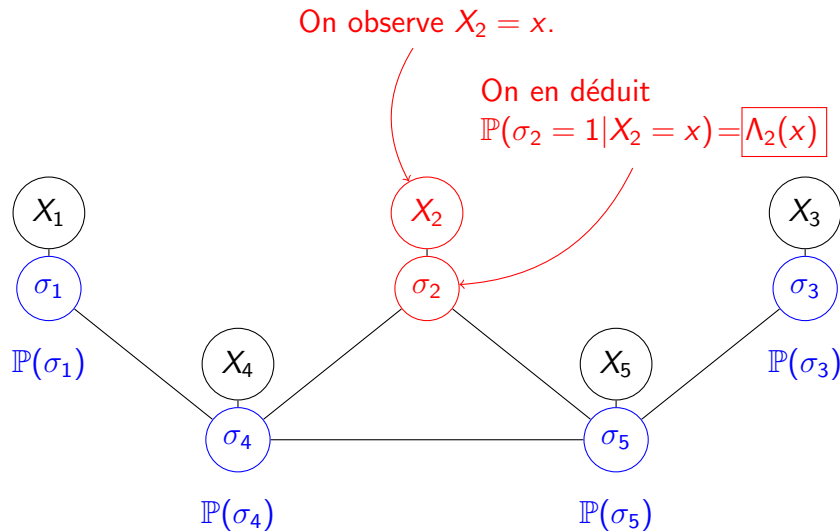
On observe $X_2 = x$.

On en déduit

$$\mathbb{P}(\sigma_2 = 1 | X_2 = x) = \Lambda_2(x)$$



Notre approximation



Notre approximation

Prédiction

$$\hat{X}_1 = \Gamma_1(\mathbb{P}(\sigma_1))$$

On observe $X_2 = x$.

On en déduit

$$\mathbb{P}(\sigma_2 = 1 | X_2 = x) = \Lambda_2(x)$$

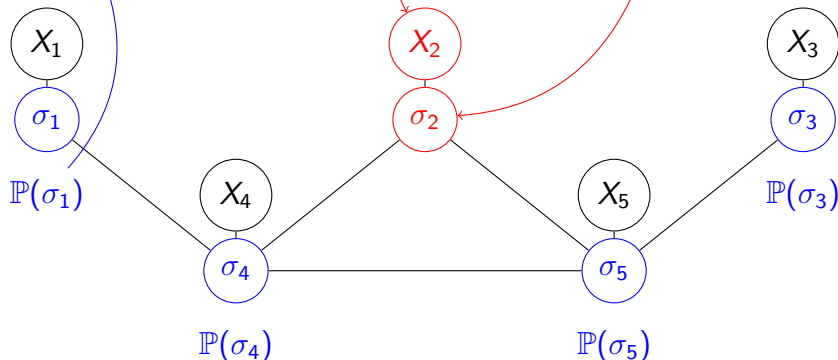


Schéma Global

$$\begin{array}{ccc} X_i = x_i \in \mathbb{R} & \xrightarrow{\Lambda_i} & \mathbb{P}(\sigma_i = 1 | X_i = x_i) \\ & & \downarrow \text{inférence} \\ X_j = x_j \in \mathbb{R} & \xleftarrow{\Gamma_j} & \mathbb{P}(\sigma_j = 1) \in [0, 1] \end{array}$$

Schéma Global

$$\begin{array}{ccc} X_i = x_i \in \mathbb{R} & \xrightarrow{\Lambda_i} & \mathbb{P}(\sigma_i = 1 | X_i = x_i) \\ & & \downarrow \text{inférence} \\ X_j = x_j \in \mathbb{R} & \xleftarrow{\Gamma_j} & \mathbb{P}(\sigma_j = 1) \in [0, 1] \end{array}$$

- Oublions le problème d'inférence (pour l'instant).

Schéma Global

$$\begin{array}{ccc} X_i = x_i \in \mathbb{R} & \xrightarrow{\Lambda_i} & \mathbb{P}(\sigma_i = 1 | X_i = x_i) \\ & & \downarrow \text{inférence} \\ X_j = x_j \in \mathbb{R} & \xleftarrow{\Gamma_j} & \mathbb{P}(\sigma_j = 1) \in [0, 1] \end{array}$$

- Oublions le problème d'inférence (pour l'instant).
- Comment choisir Λ (et Γ) ?

Choisir Λ est équivalent à définir σ

- $\Lambda(x) = \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x)$, avec $F_X(x) = \mathbb{P}(X \leq x)$

Choisir Λ est équivalent à définir σ

- $\Lambda(x) = \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x)$, avec $F_X(x) = \mathbb{P}(X \leq x)$

Différents critères

- Information mutuelle.
- Entropie.

Choix de Λ

Choisir Λ est équivalent à définir σ

- $\Lambda(x) = \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x)$, avec $F_X(x) = \mathbb{P}(X \leq x)$

Différents critères

- Information mutuelle.
- Entropie.

Contraintes sur Λ

- Fonction croissante (de 0 à 1), càdlàg.

Une interprétation stochastique de Λ

Λ est la distribution d'une variable aléatoire.

- Càdlàg, croissante de 0 à 1.
- $\Rightarrow \exists Y$ v.a.r $|\Lambda(x) = \mathbb{P}(Y \leq x) = F_Y(x)$.

Une interprétation stochastique de Λ

Λ est la distribution d'une variable aléatoire.

- Càdlàg, croissante de 0 à 1.
- $\Rightarrow \exists Y$ v.a.r $|\Lambda(x) = \mathbb{P}(Y \leq x) = F_Y(x)$.

$$\sigma \stackrel{\text{def}}{=} \mathbb{1}_{\{Y \leq X\}}.$$

Une interprétation stochastique de Λ

Λ est la distribution d'une variable aléatoire.

- Càdlàg, croissante de 0 à 1.
- $\Rightarrow \exists Y$ v.a.r $|\Lambda(x) = \mathbb{P}(Y \leq x) = F_Y(x)$.

$$\sigma \stackrel{\text{def}}{=} \mathbb{1}_{\{Y \leq X\}}.$$

Exemple

- $\Lambda = F_X \Rightarrow (X|\sigma = 1) \sim \max(X_1, X_2), (X|\sigma = 0) \sim \min(X_1, X_2)$.

Si Λ est inversible

- On peut choisir $\Gamma = \Lambda^{-1}$.
- $\Lambda^{-1}(b)$ correspond à la seule valeur de X telle que $(\sigma|X = x)$ soit distribué selon b .

Fonction de décodage Γ

Si Λ est inversible

- On peut choisir $\Gamma = \Lambda^{-1}$.
- $\Lambda^{-1}(b)$ correspond à la seule valeur de X telle que $(\sigma|X = x)$ soit distribué selon b .

Plus généralement, Λ n'est "que" càdlàg.

- En déconditionnant par σ on obtient une distribution \hat{F} :

$$\hat{F}(x) = bF^1(x) + (1 - b)F^0(x).$$

avec $F^s(x) = \mathbb{P}(X \leq x | \sigma = s)$.

- On calcule la statistique souhaitée de \hat{F} (moyenne, médiane, ...).
- On peut faire cela que Λ soit inversible ou non.

Choix de Λ

Information Mutuelle, Λ_{MI}

- $\operatorname{argmax}_{\Lambda} I(\sigma, X) = \mathbb{1}_{\{x \geq q_X^{0.5}\}}$.

Démonstration.

$$I(X, \sigma) = H(\mathbb{P}(\sigma = 1)) - \int_x H(\Lambda(x)) dF_X(x),$$

avec $H(x) = -x \log x - (1 - x) \log(1 - x)$. Terme de droite nul pour $\Lambda(x) \in \{0, 1\}$, terme de gauche maximal pour $\mathbb{P}(\sigma = 1) = 0.5$. □

σ est une fonction déterministe de X & Λ non inversible.

Entropie (relative à la mesure uniforme)

Entropie de $U = \Lambda(X)$

Entropie, Λ_S

- $\operatorname{argmax}_{\Lambda} S(\Lambda) = F_X(x)$ (fonction de repartition de X).

Démonstration.

L'entropie est maximale pour une variable uniforme sur $[0, 1]$ et la fonction de repartition transforme X en une v.a. uniforme. □

Entropie (relative à la mesure uniforme)

$$\text{Entropie de } U = \Lambda(X)$$

Entropie, Λ_S

- $\operatorname{argmax}_{\Lambda} S(\Lambda) = F_X(x)$ (fonction de repartition de X).

Démonstration.

L'entropie est maximale pour une variable uniforme sur $[0, 1]$ et la fonction de repartition transforme X en une v.a. uniforme. □

Sans observations, quelle prédiction ?

- $\mathbb{P}(\sigma = 1) = \frac{1}{2}$.
- $F_X^{-1}(\mathbb{P}(\sigma = 1)) = q_X^{0.5} \Rightarrow$ optimal seulement pour une erreur L^1 .

σ est une variable aléatoire à X fixé.

Loi des paires de variables binaires

- Loi marginales fixées par le choix de Λ

$$\mathbb{P}(\sigma_i = 1) = \int_x \Lambda_i(x) dF_x(x).$$

- Reste un paramètre de corrélation $\mathbb{P}(\sigma_i \sigma_j = 1)$.

Correlations des variables binaires

Loi des paires de variables binaires

- Loi marginales fixées par le choix de Λ

$$\mathbb{P}(\sigma_i = 1) = \int_x \Lambda_i(x) dF_x(x).$$

- Reste un paramètre de corrélation $\mathbb{P}(\sigma_i \sigma_j = 1)$.

Deux méthodes

- Méthode des moments : $\mathbb{E}[\Lambda_1(X_1)\Lambda_2(X_2)] = \langle \Lambda_1(X_1)\Lambda_2(X_2) \rangle$:

$$\text{cov}(\sigma_1, \sigma_2) = \widehat{\text{cov}}(\Lambda_1(X_1), \Lambda_2(X_2)) \prod_{i \in \{1,2\}} \frac{\text{var}(\sigma_i)}{\text{var}(\Lambda_i(X_i))}.$$

- Maximum de vraisemblance (via l'algorithme EM).

Objectif : calcul approché des lois marginales

- A partir d'une forme produit.

$$\mathbb{P}(\boldsymbol{\sigma}) = \prod_{(ij)} \varphi_{ij}(\sigma_i, \sigma_j) \prod_i \theta_i(\sigma_i)$$

Objectif : calcul approché des lois marginales

- A partir d'une forme produit.

$$\mathbb{P}(\boldsymbol{\sigma}) = \prod_{(ij)} \varphi_{ij}(\sigma_i, \sigma_j) \prod_i \theta_i(\sigma_i)$$

Belief Propagation (BP)

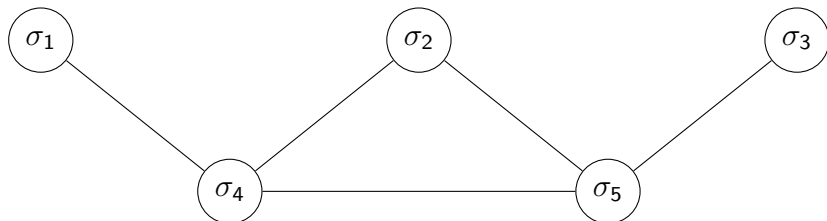
- Méthode itérative par passage de messages.
- Marginales exactes si le graphe ne comporte pas de cycle.
- Minimisation d'une pseudo-distance aux vraies marginales en général.
- S'il y a convergence...

Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

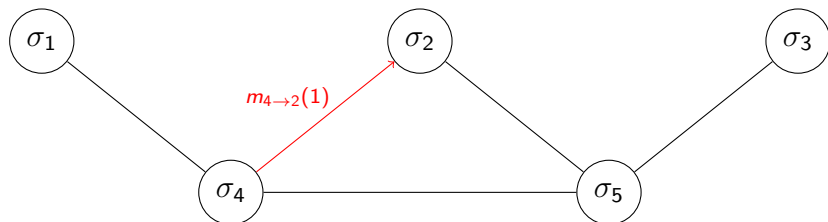


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

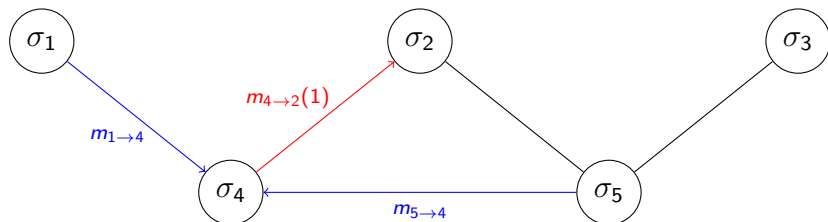


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

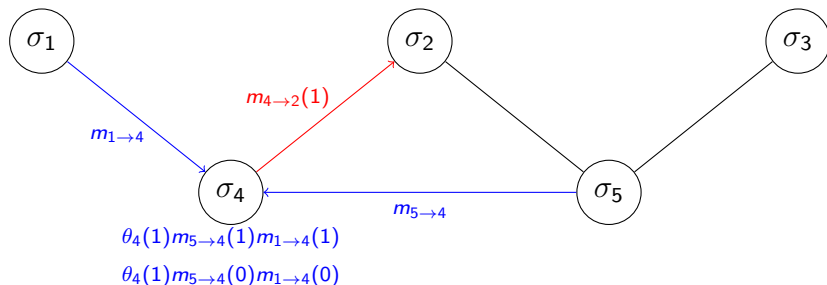


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

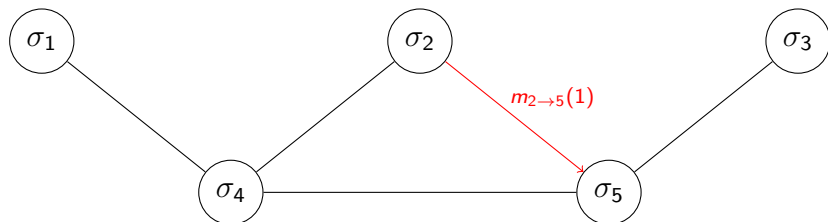


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

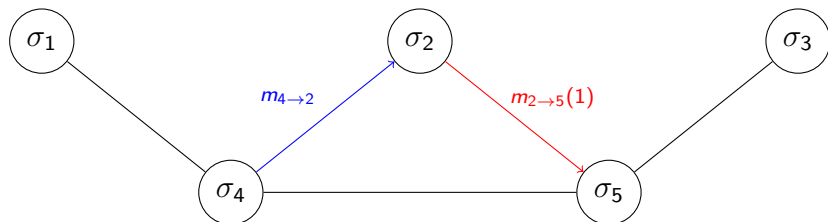


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

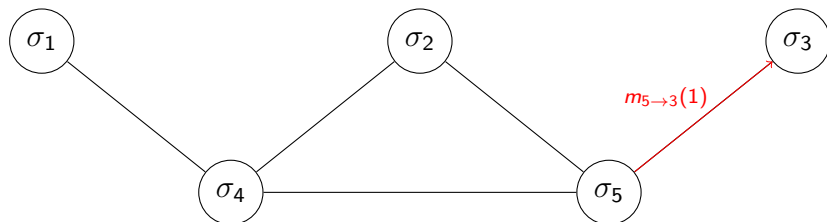


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

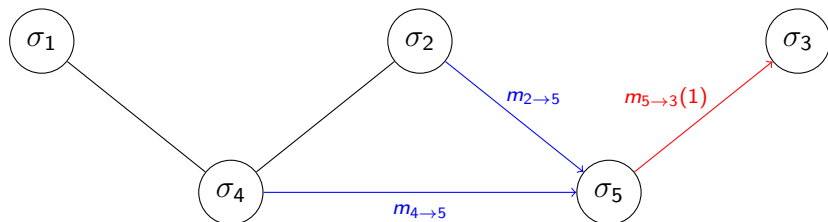


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$

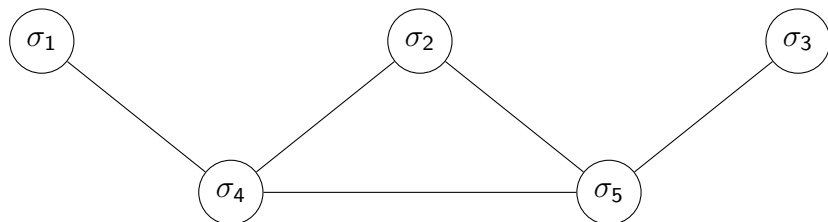


Belief Propagation

Formule de mise à jour des messages

- Message envoyé par un noeud i vers un noeud j

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i \in \{0,1\}} \varphi_{ij}(\sigma_i, \sigma_j) \theta_i(\sigma_i) \prod_{k \in \partial j \setminus i} m_{k \rightarrow i}(\sigma_i).$$



$$b_i(\sigma_i) \propto \theta_i(\sigma_i) \prod_{j \in \partial i} m_{j \rightarrow i}(\sigma_i)$$

$$b_{ij}(\sigma_i, \sigma_j) \propto \varphi_{ij}(\sigma_i, \sigma_j) \frac{b_i(\sigma_i) b_j(\sigma_j)}{m_{i \rightarrow j}(\sigma_i) m_{j \rightarrow i}(\sigma_j)}$$

Belief Propagation

Comment introduire nos observations ?

- Une observation de X_i se traduit en probabilité sur σ_i .
- Fixer la valeur de σ_i est naturel avec BP pas sa probabilité.
- Comment fixer une probabilité ?

Belief Propagation

Comment introduire nos observations ?

- Une observation de X_i se traduit en probabilité sur σ_i .
- Fixer la valeur de σ_i est naturel avec BP pas sa probabilité.
- Comment fixer une probabilité ?

Repartons du problème de minimisation

{Points fixes stables de BP} \subset {Minima locaux d'une approximation de $\text{KL}(b||\mathbb{P})$ }

$$\min_b \sum_{\sigma} b(\sigma) \log \frac{b(\sigma)}{\mathbb{P}(\sigma)}$$

sous contraintes

$$b(\sigma) = \prod_{(ij)} \frac{b_{ij}(\sigma_i, \sigma_j)}{b_i(\sigma_i) b_j(\sigma_j)} \prod_i b_i(\sigma_i), \quad \sum_{\sigma_j} b_{ij}(\sigma_i, \sigma_j) = b_i(\sigma_i), \quad \sum_{\sigma_j} b_j(\sigma_j) = 1.$$

L'équation de mise à jour est obtenue à partir des points stationnaires du Lagrangien de ce problème.

Mirror Belief Propagation

Nouvelles contraintes

Pour chaque observation $X_i = x_i$ on ajoute une contrainte du type

$$b_i(\sigma_i = 1) = \Lambda(x_i) \stackrel{\text{def}}{=} b_i^*(1), \quad b_i(\sigma_i = 0) = 1 - \Lambda(x_i) \stackrel{\text{def}}{=} b_i^*(0)$$

Mirror Belief Propagation

Nouvelles contraintes

Pour chaque observation $X_i = x_i$ on ajoute une contrainte du type

$$b_i(\sigma_i = 1) = \Lambda(x_i) \stackrel{\text{def}}{=} b_i^*(1), \quad b_i(\sigma_i = 0) = 1 - \Lambda(x_i) \stackrel{\text{def}}{=} b_i^*(0)$$

Version modifiée de Belief Propagation :

- $m_{i \rightarrow j}(\sigma_j)$ inchangé si X_i n'est pas observé
- sinon :

$$m_{i \rightarrow j}(\sigma_j) \propto \sum_{\sigma_i} \varphi_{ij}(\sigma_i, \sigma_j) \frac{b_i^*(\sigma_i)}{m_{j \rightarrow i}(\sigma_i)}$$

Le noeud i ne laisse plus traverser d'information.

Détail de l'expérience

Détail de l'expérience

- On observe les variables dans un ordre aléatoire.

Détail de l'expérience

- On observe les variables dans un ordre aléatoire.
- On prédit les variables non observées.

Détail de l'expérience

- On observe les variables dans un ordre aléatoire.
- On prédit les variables non observées.
- On calcule l'erreur L^1 moyenne de prédiction.

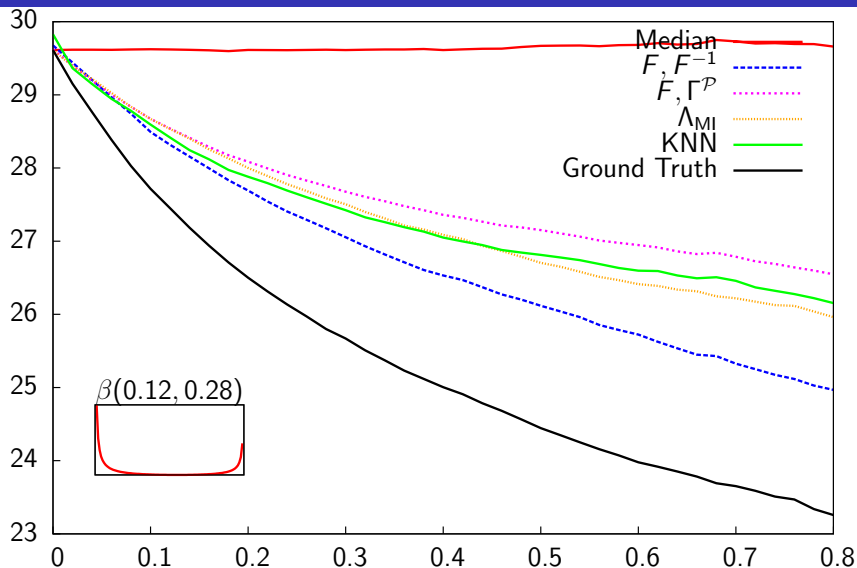
Détail de l'expérience

- On observe les variables dans un ordre aléatoire.
- On prédit les variables non observées.
- On calcule l'erreur L^1 moyenne de prédiction.

Données synthétiques

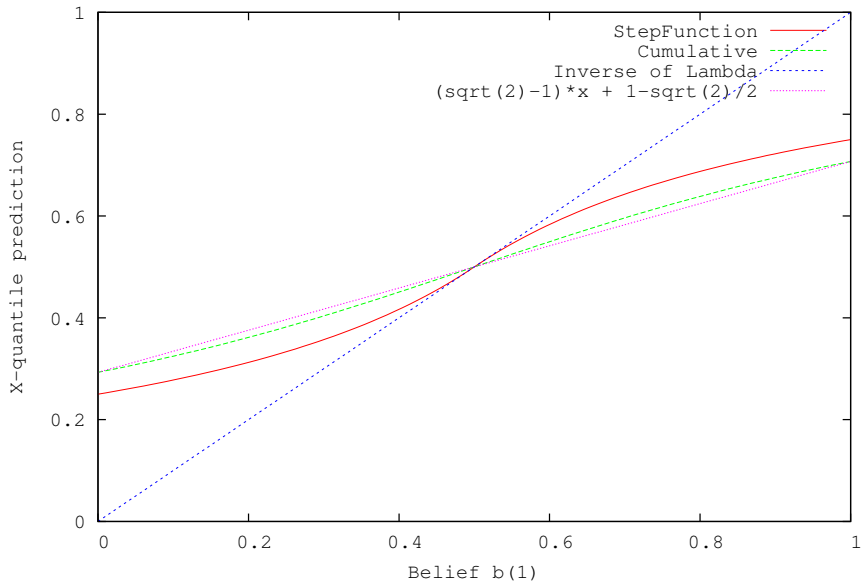
Variables $\beta(a, b)$ distribuées sur un arbre.

Expérimentations numériques



Merci de votre attention !

Quantile of X prediction for $p=1/2$ and different Decoding Methods



Expérimentations numériques

