a latent Ising model for real-valued variables inference

Victorin MARTIN

Under the supervision of Cyril FURTLEHNER and Jean-Marc LASGOUTTES.

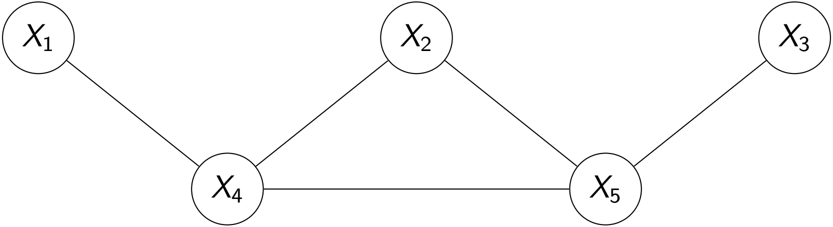
Outline

Real valued variables model

Pairwise Markov random field of real valued variables

- The joint pdf writes as a product (Hammersley-Clifford's theorem):

$$\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$$

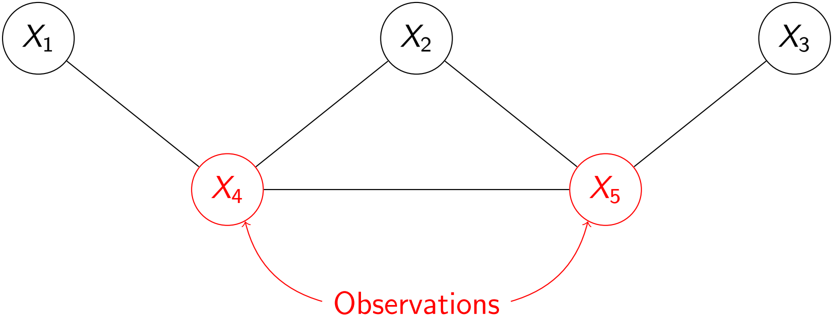


Real valued variables model

Pairwise Markov random field of real valued variables

- The joint pdf writes as a product (Hammersley-Clifford's theorem):

$$\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$$

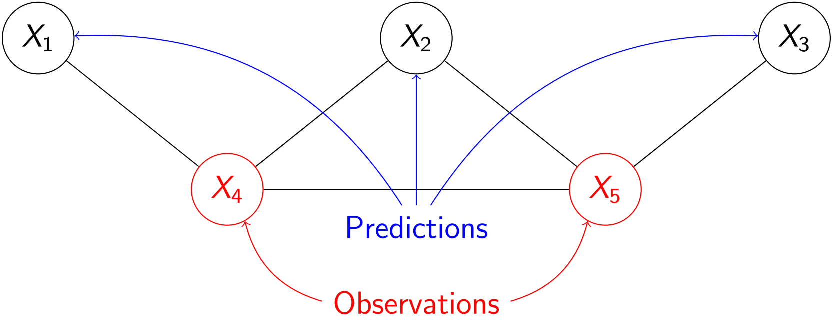


Real valued variables model

Pairwise Markov random field of real valued variables

- The joint pdf writes as a product (Hammersley-Clifford's theorem):

$$\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$$



Predicting a random variable?

Optimal prediction $\hat{\theta}_\ell(X)$

- Defined w.r.t a loss function ℓ, e.g.

$$\hat{\theta}_\ell(X) = \operatorname*{argmin}_z \mathbb{E}[\ell(X, z)]$$

- Ex:

$$\hat{\theta}_{L^2}(X) = \operatorname*{argmin}_z \mathbb{E}[(X - z)^2] = \mathbb{E}[X],$$

$$\hat{\theta}_{L^1}(X) = \operatorname*{argmin}_z \mathbb{E}[|X - z|] = q_X^{0.5},$$

$$\hat{\theta}_{\mathrm{ML}}(X) = \operatorname*{argmin}_z -\mathcal{P}_X(z).$$

Setting

A model estimation problem

- From historical data $\{\mathbf{X}^k\}_{k\in\{1..M\}}$, model estimation i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.

Setting

A model estimation problem

- From historical data $\{\mathbf{X}^k\}_{k\in\{1..M\}}$, model estimation i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.

An inference problem

- Given some (sparse) observations $\mathcal{O}$ compute the predictions $\hat{\theta}_\ell(X_i|\mathcal{O})$.

Setting

A model estimation problem

- From historical data $\{\mathbf{X}^k\}_{k\in\{1..M\}}$, model estimation i.e. $\mathbb{P}(\mathbf{X}) = \prod_{(ij)} \varphi_{ij}(X_i, X_j)$.

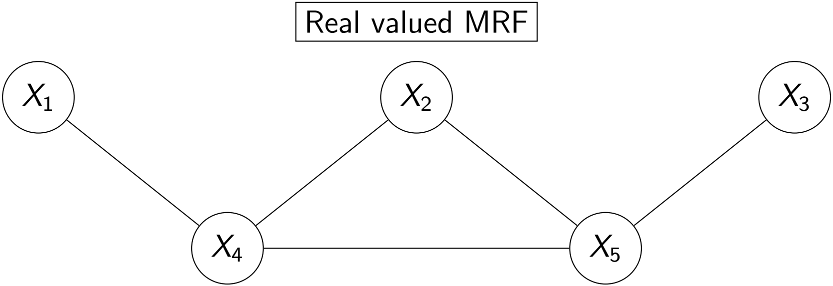
An inference problem

- Given some (sparse) observations $\mathcal{O}$ compute the predictions $\hat{\theta}_\ell(X_i|\mathcal{O})$.

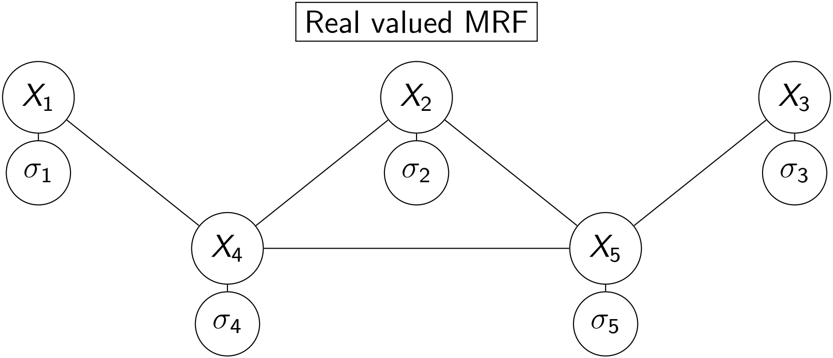
Strongly related problems

- Estimation and inference are done in approximate ways.
- Both approximations should be related...

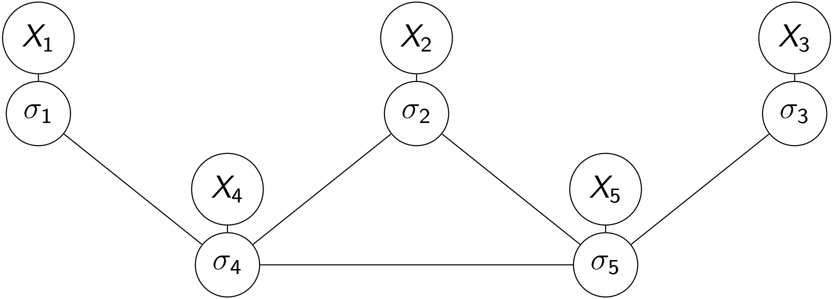
Our approximation



Our approximation

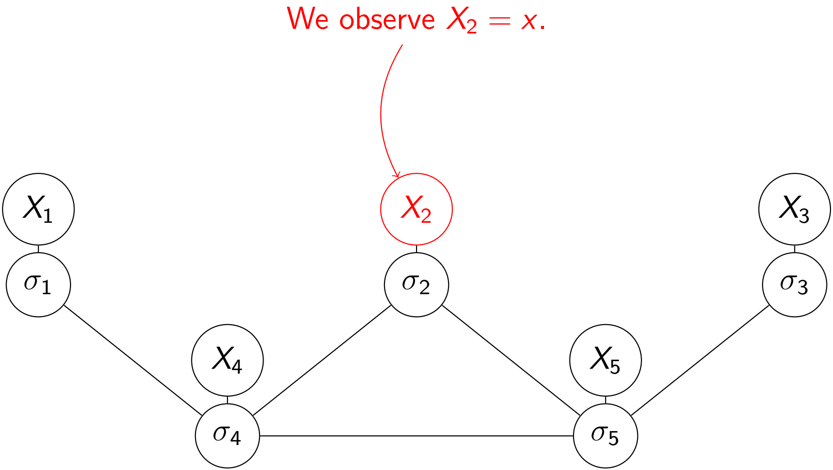


Our approximation



Binary MRF

Our approximation

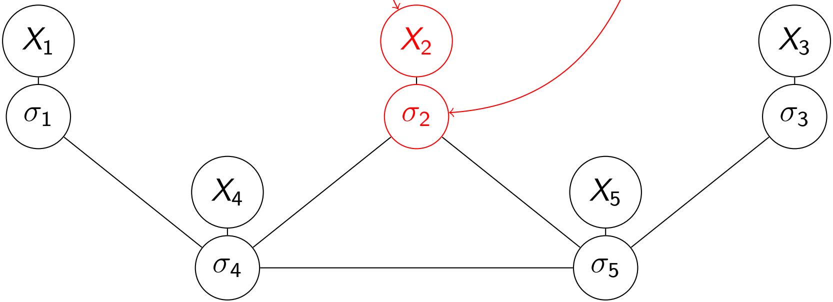


Our approximation

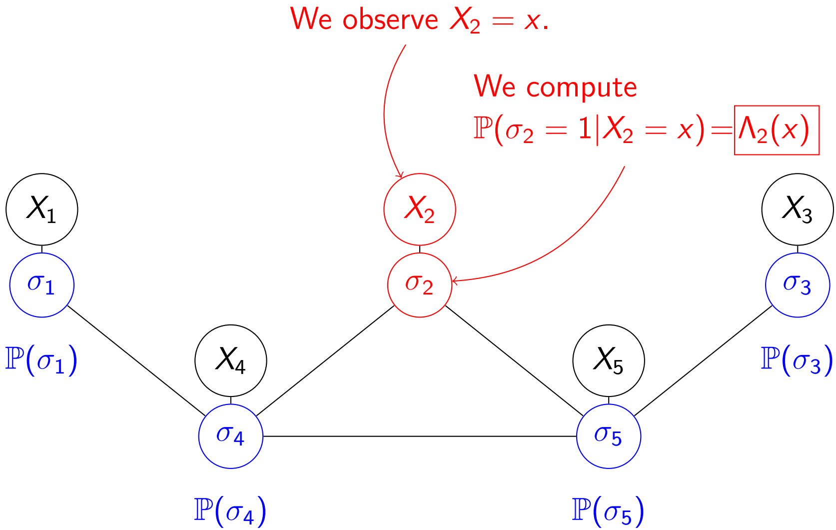
We observe $X_2 = x$.

We compute

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



Our approximation

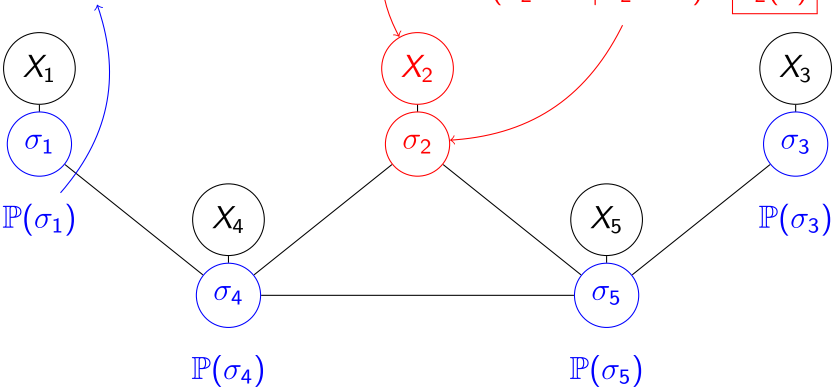


Our approximation

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

We observe $X_2 = x$.

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



Summary...

Global scheme

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

Summary...

Global scheme

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

- Let us put aside the inference (for now).
- First, how to choose Λ?

Choice of Λ

The choice of Λ is equivalent to the definition of σ

- $\Lambda(x) \stackrel{\text{def}}{=} \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x) = \mathbb{E}[\Lambda(X)]$, with $F_X(x) = \mathbb{P}(X \leq x)$

Choice of Λ

The choice of Λ is equivalent to the definition of σ

- $\Lambda(x) \stackrel{\text{def}}{=} \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x) = \mathbb{E}[\Lambda(X)]$, with $F_X(x) = \mathbb{P}(X \leq x)$

Constraints over Λ

- Increasing function (from 0 to 1), càdlàg.

Choice of Λ

The choice of Λ is equivalent to the definition of σ

- $\Lambda(x) \stackrel{\text{def}}{=} \mathbb{P}(\sigma = 1 | X = x)$
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x) = \mathbb{E}[\Lambda(X)]$, with $F_X(x) = \mathbb{P}(X \leq x)$

Constraints over Λ

- Increasing function (from 0 to 1), càdlàg.

Selection criteria

- Mutual information.
- Entropy.

Stochastic meaning of Λ

Λ is the cumulative distribution function of some random variable.

- Càdlàg, increasing from 0 to 1.
- $\Rightarrow \exists Y | \ \Lambda(x) = \mathbb{P}(Y \leq x) = F_Y(x)$.

Stochastic meaning of Λ

Λ is the cumulative distribution function of some random variable.

- Càdlàg, increasing from 0 to 1.
- $\Rightarrow \exists Y | \; \Lambda(x) = \mathbb{P}(Y \leq x) = F_Y(x)$.

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

Stochastic meaning of Λ

Λ is the cumulative distribution function of some random variable.

- Càdlàg, increasing from 0 to 1.
- $\Rightarrow \exists Y |\ \Lambda(x) = \mathbb{P}(Y \le x) = F_Y(x)$.

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

Example

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

Choice of Λ: a mutual information criterion

Maximal mutual information between σ and X, Λ_{MI}

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

Proof.

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

avec $H(x) = -x \log x - (1-x)\log(1-x)$. Right term is 0 pour $\Lambda(x) \in \{0, 1\}$, Left term maximized for $\mathbb{P}(\sigma = 1) = 0.5$. $\square$

$\sigma | X$ is deterministic & Λ is not invertible.

Choice of Λ: a Max-entropy principle

Maximal (relative) entropy of $U = \Lambda(X)$, Λ_S

- $\text{argmax}_\Lambda\, S(\Lambda) = F_X(x)$ (Cdf of X).

Proof.

The entropy is maximized for an uniform variable on $[0,1]$. The cumulative distribution function maps X to a $\mathcal{U}[0,1]$. $\square$

$\sigma|X$ is a random variable.

Decoding Function

Global scheme

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

Choosing Γ

If Λ is invertible,

- We can pick $\Gamma = \Lambda^{-1}$.
- $\Lambda^{-1}(b)$ is the only X-value such as $\mathbb{P}(\sigma = 1|X = x) = b$.

Choosing Γ

If Λ is invertible,

- We can pick $\Gamma = \Lambda^{-1}$.
- $\Lambda^{-1}(b)$ is the only X-value such as $\mathbb{P}(\sigma = 1|X = x) = b$.

General case, $\Gamma^{\mathcal{P}}$

- Deconditioning w.r.t σ yields a distribution $\hat{F}$:

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

  with $F^s(x) = \mathbb{P}(X \leq x|\sigma = s)$.
- We can compute a given statistic of $\hat{F}$ (mean, median, ...).
- It doesn't matter if Λ is invertible or not.

Prediction without observation

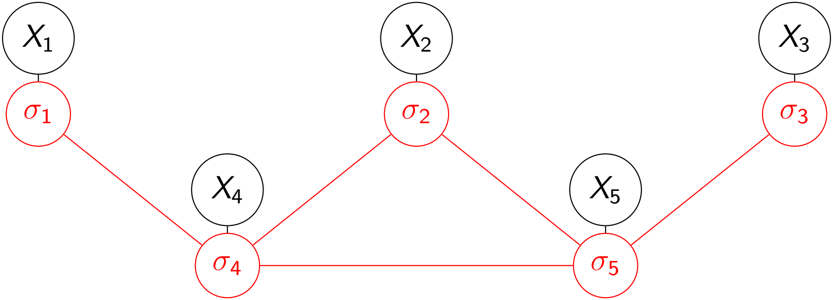
General case

- In all cases $\hat{F} = F_X$.
- $\Gamma^{\mathcal{P}}(\mathbb{P}(\sigma = 1))$ is always the optimal predictor $\hat{\theta}_\ell(X)$.

Invertible Cdf case

- $\mathbb{P}(\sigma = 1) = \frac{1}{2}$.
- $F_X^{-1}(\mathbb{P}(\sigma = 1)) = q_X^{0.5} \Rightarrow$ optimal only for the L^1 loss function.

Ising Model Estimation



Ising Model

Estimation of $\mathbb{P}(\sigma_i, \sigma_j)$

Binary pairwise distribution

- Marginal distributions fixed by the choice of Λ

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

- The correlation parameter remains to be fixed: $\mathbb{P}(\sigma_i \sigma_j = 1)$.

Estimation of $\mathbb{P}(\sigma_i, \sigma_j)$

Binary pairwise distribution

- Marginal distributions fixed by the choice of Λ

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

- The correlation parameter remains to be fixed: $\mathbb{P}(\sigma_i \sigma_j = 1)$.

Two methods

- Moment matching: $\mathbb{E}[\Lambda_1(X_1)\Lambda_2(X_2)] = \langle \Lambda_1(X_1)\Lambda_2(X_2) \rangle$:

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

- Maximum Likelihood (using EM algorithm).

Inference

Global scheme

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

Setting

We want to approximate the marginals

- From a product form.

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

Setting

We want to approximate the marginals

- From a product form.

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

(Loopy) Belief Propagation (BP)

- Message-Passing algorithm.
- Yields the exact marginals when the graph is a tree.
- Minimization of a "distance" to the true marginals.
- No general result about convergence...

Algorithm definition

Update rules

- Message sent from a node i to a node j

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

- After convergence is reached, we compute

$$b_i(\sigma_i) \propto \gamma_i(\sigma_i)\prod_{j\in\partial i} m_{j\to i}(\sigma_i)$$

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

which are compatible $\sum_{\sigma_j} b_{ij}(\sigma_i,\sigma_j) = b_i(\sigma_i)$

How to take our observations into account?

Our observations of X_i gives us the distribution of σ_i.

- Fixing σ_i value is natural with BP but not fixing its distribution.

How to take our observations into account?

Our observations of X_i gives us the distribution of σ_i.

- Fixing σ_i value is natural with BP but not fixing its distribution.

Variational point of view of BP

{Stable BP fixed points} $\subset$ {Local minima of $\mathrm{KL}_{\mathrm{Bethe}}(b||\mathbb{P})$}

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

subject to

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

and assuming $\sum_{\boldsymbol{\sigma} \setminus \sigma_i, \sigma_j} b(\boldsymbol{\sigma}) = b_{ij}(\sigma_i, \sigma_j)$ (Bethe approximation).

BP update rules are obtained from the stationary points of the corresponding Lagrangian.

Variational definition of our BP variant

New "soft" constraints: Mirror BP

For each observation, $X_i = x_i$ we add a constraint

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

Variational definition of our BP variant

New "soft" constraints: Mirror BP

For each observation, $X_i = x_i$ we add a constraint

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

Modified version of Belief Propagation:

- $m_{i\to j}(\sigma_j)$ is the same as usual if σ_i is not subject to soft constraints.
- else:

$$m_{i\to j}(\sigma_j) \propto \sum_{\sigma_i} \varphi_{ij}(\sigma_i, \sigma_j) \frac{b_i^*(\sigma_i)}{m_{j\to i}(\sigma_i)} = \sum_{\sigma_i} \frac{b_i^*(\sigma_i)}{b_i^{\text{BP}}(\sigma_i)} \varphi_{ij}(\sigma_i, \sigma_j)\gamma_i(\sigma_i) \prod_{k\in\partial j\setminus i} m_{k\to i}(\sigma_i)$$

Variational definition of our BP variant

New "soft" constraints: Mirror BP

For each observation, $X_i = x_i$ we add a constraint

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

Modified version of Belief Propagation:

- $m_{i\to j}(\sigma_j)$ is the same as usual if σ_i is not subject to soft constraints.
- else:

$$m_{i\to j}(\sigma_j) \propto \sum_{\sigma_i} \varphi_{ij}(\sigma_i,\sigma_j)\frac{b_i^*(\sigma_i)}{m_{j\to i}(\sigma_i)} = \sum_{\sigma_i} \frac{b_i^*(\sigma_i)}{b_i^{\text{BP}}(\sigma_i)} \varphi_{ij}(\sigma_i,\sigma_j)\gamma_i(\sigma_i)\prod_{k\in\partial j\setminus i} m_{k\to i}(\sigma_i)$$

The information doesn't cross node i anymore, as if σ_i is fixed.

Variational definition of our BP variant

New "soft" constraints: Mirror BP

For each observation, $X_i = x_i$ we add a constraint

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

Modified version of Belief Propagation:

- $m_{i\to j}(\sigma_j)$ is the same as usual if σ_i is not subject to soft constraints.
- else:

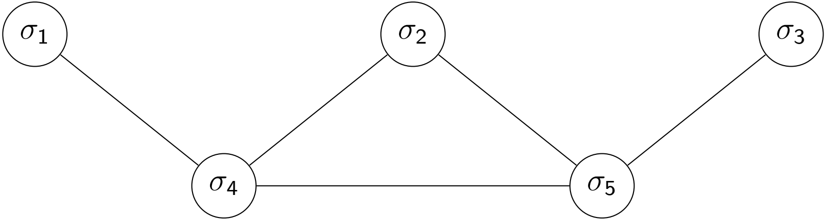
$$m_{i\to j}(\sigma_j) \propto \sum_{\sigma_i} \varphi_{ij}(\sigma_i, \sigma_j) \frac{b_i^*(\sigma_i)}{m_{j\to i}(\sigma_i)} = \sum_{\sigma_i} \frac{b_i^*(\sigma_i)}{b_i^{\text{BP}}(\sigma_i)} \varphi_{ij}(\sigma_i, \sigma_j)\gamma_i(\sigma_i) \prod_{k\in\partial j\setminus i} m_{k\to i}(\sigma_i)$$

The information doesn't cross node i anymore, as if σ_i is fixed.

$$b_{ij}^{\text{Mirror}}(\sigma_i, \sigma_j) = \frac{b_i^*(\sigma_i)}{b_i^{\text{BP}}(\sigma_i)} b_{ij}^{\text{BP}}(\sigma_i, \sigma_j) \quad \text{similar to Jeffrey's rule.}$$

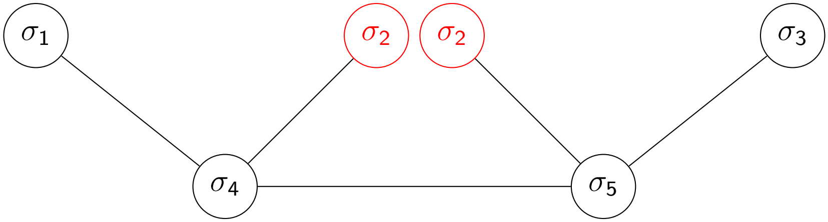
Stability of Mirror BP

Fixing the belief of a node has the effect of graph cutting at this node.



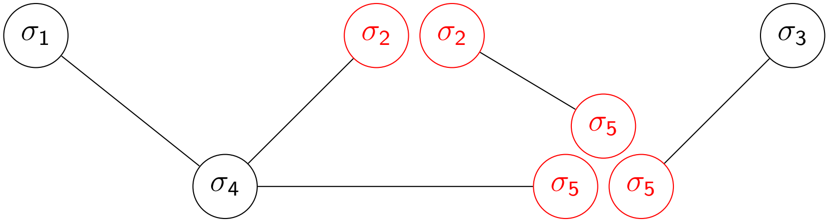
Stability of Mirror BP

Fixing the belief of a node has the effect of graph cutting at this node.



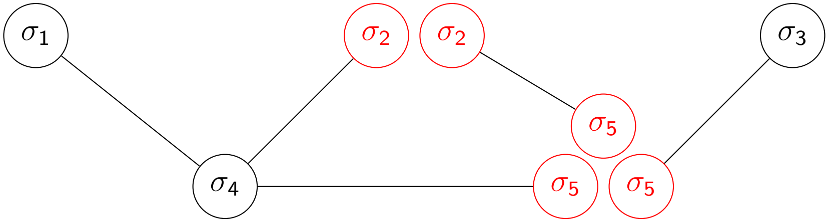
Stability of Mirror BP

Fixing the belief of a node has the effect of graph cutting at this node.



Stability of Mirror BP

Fixing the belief of a node has the effect of graph cutting at this node.

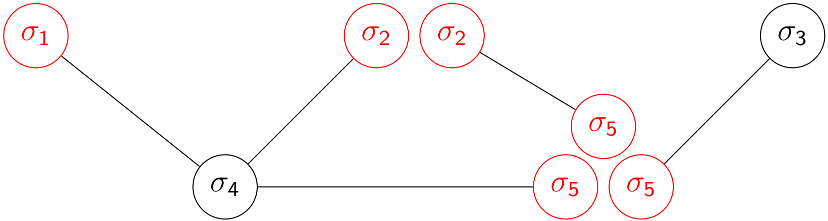


Weak (theoretical) result

If the resulting graph is formed by disconnected trees containing no more than two observed leaves, Mirror-BP converges to a unique fixed point.

Stability of Mirror BP

Fixing the belief of a node has the effect of graph cutting at this node.



Weak (theoretical) result

If the resulting graph is formed by disconnected trees containing no more than two observed leaves, Mirror-BP converges to a unique fixed point.

A decimation experiment

What do we do?

- Reveal the variables in a random order.
- Predict the non observed variables.
- Compute the mean L^1 prediction error.

A decimation experiment

What do we do?

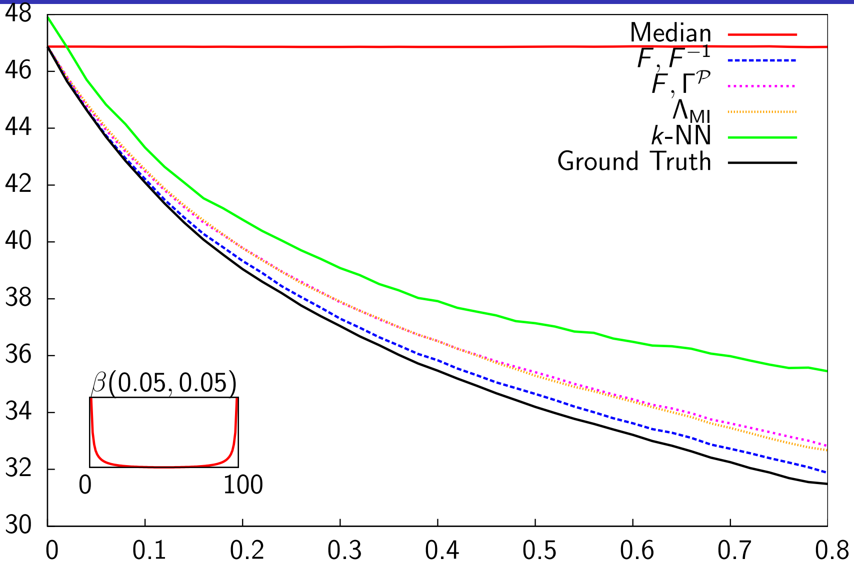
- Reveal the variables in a random order.
- Predict the non observed variables.
- Compute the mean L^1 prediction error.

Simulated data

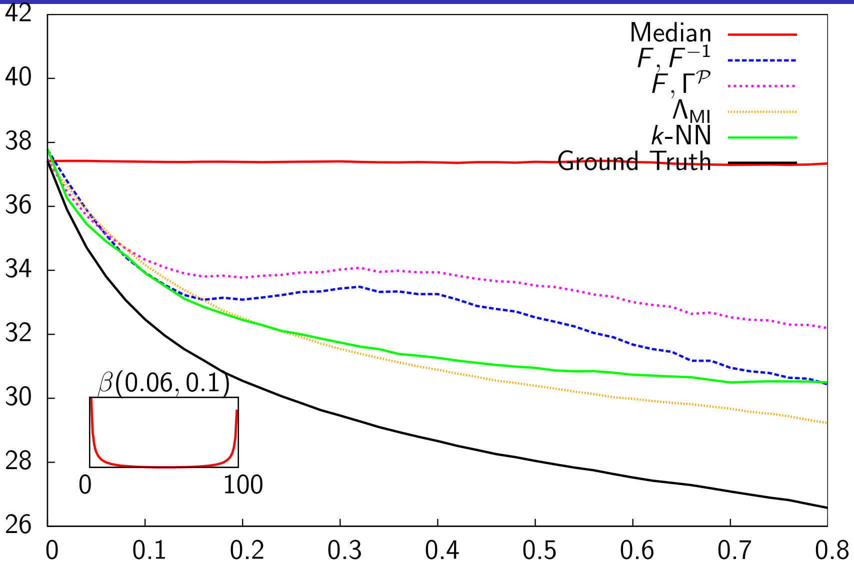
Variables $X_i \sim \beta(a, b)$ over a tree.

$$\mathrm{pdf}(\beta(a, b)) \propto x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x).$$

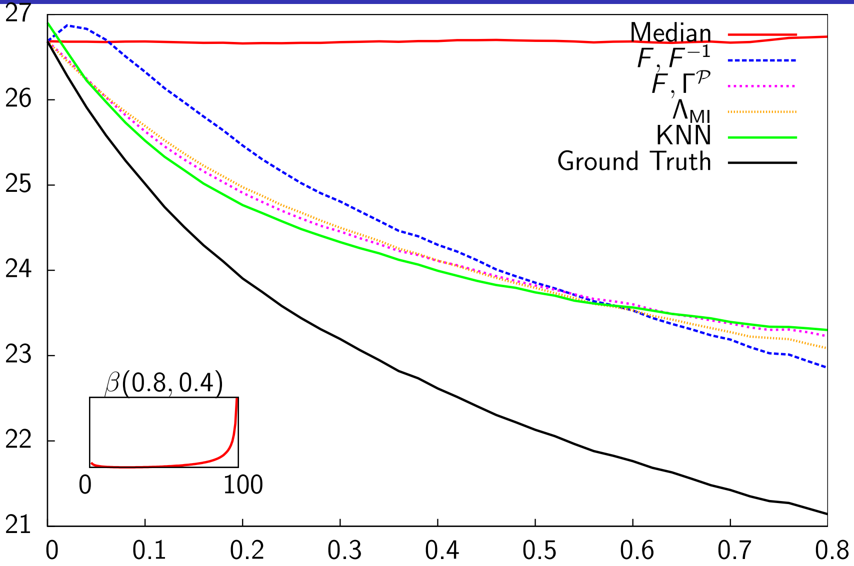
Binary limit on a binary tree.



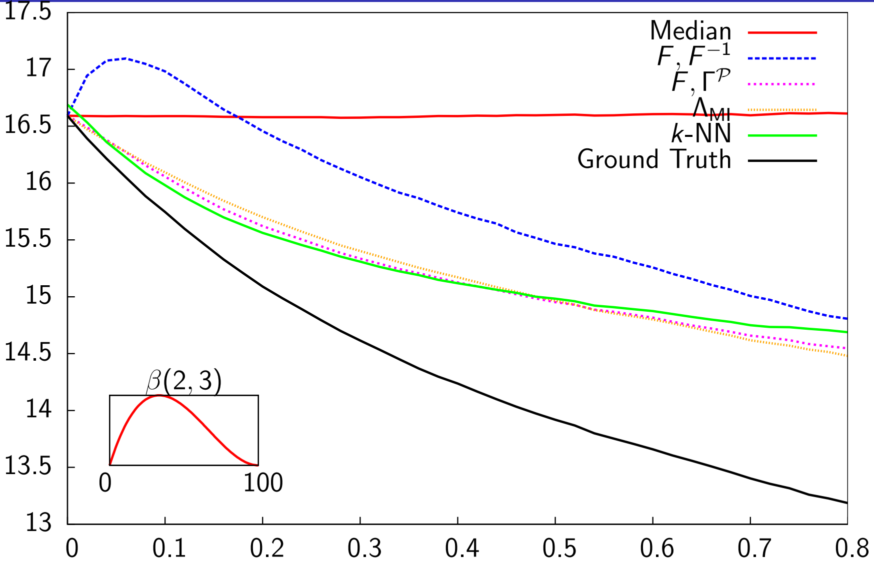
Non-symmetric variables on 4-ary tree.



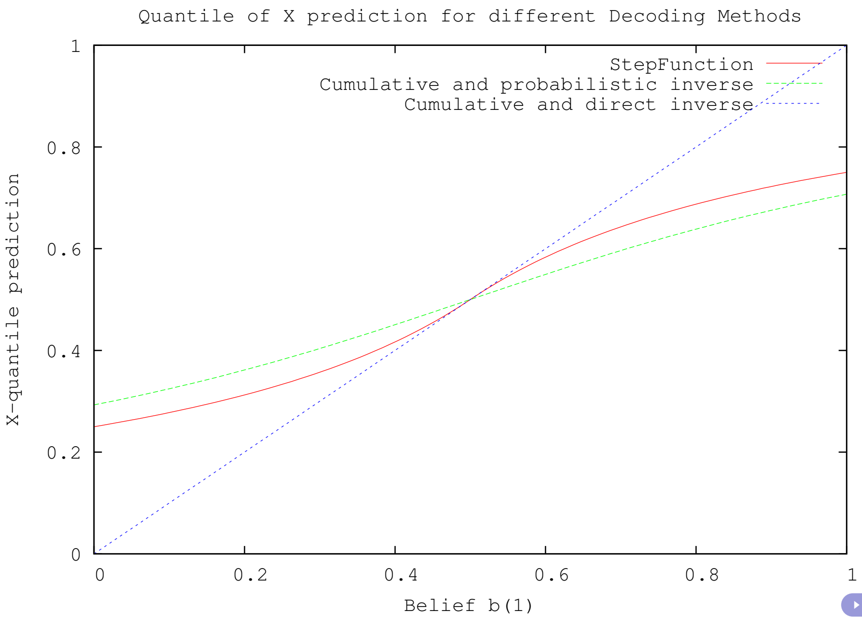
Away from binary variables on a binary tree.



Unimodal variables on a binary tree.



Thank you for your attention!


Quantile of X prediction for different Decoding Methods
StepFunction
Cumulative and probabilistic inverse
Cumulative and direct inverse
X-quantile prediction
Belief b(1)
0
0.2
0.4
0.6
0.8
1

Max-entropy principle (again)

Optimal prediction without observation

- New constraint in the entropy maximization.
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x) = \Lambda(\hat{\theta}(X))$.

Max-entropy principle (again)

Optimal prediction without observation

- New constraint in the entropy maximization.
- $\mathbb{P}(\sigma = 1) = \int_x \Lambda(x) dF_X(x) = \Lambda(\hat{\theta}(X))$.

Distortion of the cdf.

$$\begin{cases} \Lambda_S^{\hat{\theta}(X)}(x) = \dfrac{1}{\alpha} \log\left(\alpha F(x) + 1\right), \forall x \leq \hat{\theta}(X) \\ \Lambda_S^{\hat{\theta}(X)}(x) = 1 + \dfrac{1}{\alpha} \log\left(\alpha (F(x) - 1) + 1\right), \forall x > \hat{\theta}(X), \end{cases}$$

with $F(\hat{\theta}(X)) = \frac{1+e^{\alpha}(\alpha-1)}{\alpha(e^{\alpha}-1)}$. When $\hat{\theta}(X) \to q_X^{0.5}$, $\alpha \to 0$ and then $\Lambda_S^{\hat{\theta}(X)}(x) \to F_X(x)$.