Pairwise MRF Calibration by Perturbation of the Bethe Reference Point

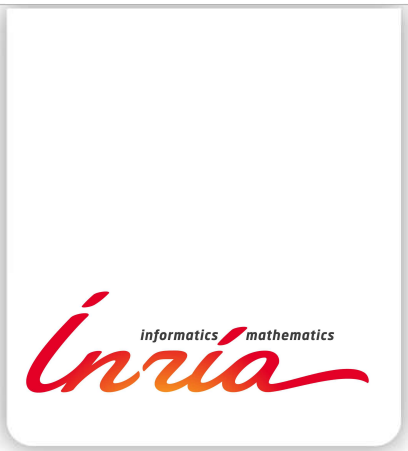
Cyril Furtlehner, Yufei Han, Jean-Marc Lasgouttes, Victorin Martin

Pairwise MRF Calibration by Perturbation of the Bethe Reference Point

Cyril Furtlehner, Yufei Han, Jean-Marc Lasgouttes, Victorin Martin

RESEARCH REPORT
N° 8059
2012
Project-Teams TAO and Imara

ISSN 0249-6399 ISRN INRIA/RR--8059--FR+ENG



Pairwise MRF Calibration by Perturbation of the Bethe Reference Point

Cyril Furtlehner*, Yufei Han†, Jean-Marc Lasgouttes†, Victorin Martin*

Project-Teams TAO and Imara

Research Report n° 8059 — 2012 — 35 pages

Abstract: We investigate different ways of generating approximate solutions to the inverse problem of pairwise Markov random field (MRF) model learning. We focus mainly on the inverse Ising problem, but discuss also the somewhat related inverse Gaussian problem. In both cases, the belief propagation algorithm can be used in closed form to perform inference tasks. Our approach consists in taking the Bethe mean-field solution as a reference point for further perturbation procedures. We remark indeed that both the natural gradient and the best link to be added to a maximum spanning tree (MST) solution can be computed analytically. These observations open the way to many possible algorithms, able to find approximate sparse solutions compatible with belief propagation inference procedures. Experimental tests on various datasets with refined L_0 or L_1 regularization procedures indicate that this approach may be a competitive and useful alternative to existing ones.

Key-words: Random Markov Fields, Ising Models, Inference, inverse problem, mean-field,belief propagation)

* Inria Saclay
† Inria Paris-Rocquencourt

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Calibration de Champ Markovien Aléatoire par Perturbation de la Solution de Bethe

Résumé : Nous étudions différentes méthodes pour trouver des solutions approchées au problème inverse de calibration de champ Markovien aléatoire à interaction de paires. Nous considérons principalement au modèle d'Ising ainsi qu'au problème lié de modèle Gaussien. En principe dans ces deux cas l'algorithme de propagation de croyance peut-être utilisé sous forme cohérente pour résoudre des problèmes d'inférence. Notre approche consiste à utiliser la solution de champ moyen de Bethe comme référence et d'effectuer différentes perturbations à partir de ce point de départ. Nous remarquons en particulier que le gradient naturel ainsi que le lien optimal à ajouter a graphe de facteurs obtenu comme arbre couvrant maximal peuvent être obtenus de façon analytique. Ces observation ouvrent un certain nombre de perspectives algorithmiques permettant de trouver des solutions sur des graphes dilués, compatibles avec la propagation de croyances. Des tests numériques portant sur différents jeux de données permettant une comparaison à des methodes de régularisation L_0 ou L_1 indiquent que cette approche peut-être une alternative compétitive aux méthodes classiques.

Mots-clés : Champs Markovien aléatoires, modèle d'Ising, Inférence, problème inverse, champ moyen, propagation de croyances

Contents

1 Introduction

The problem at stake is a model selection problem, in the MRF families, where N variables are observed pair by pair. The optimal solution is the MRF with maximal entropy obeying moment constraints. For binary variables, this happens then to be the Ising model with highest log-likelihood. It is a difficult problem, where both the graph structure and the value of the fields and coupling have to be found. In addition, we wish to ensure that the model is compatible with the fast inference algorithm "belief propagation" (BP) to be useful at large scale for real-time inference tasks. This leads us to look at least for a good trade-off between likelihood and sparsity.

Concerning the Inverse Ising Problem (IIP), the existing approaches fall mainly in the following categories:

- Purely computational efficient approaches rely on various optimization schemes of the log likelihood [21] or on pseudo-likelihood [18] along with sparsity constraints to select the only relevant features.
- Common analytical approaches are based on the Plefka expansion [33] of the Gibbs free energy by making the assumption that the coupling constants J_{ij} are small. The picture is then of a weakly correlated uni-modal probability measure. For example, the recent approach proposed in [7] is based on this assumption.
- Another possibility is to assume that relevant coupling J_{ij} have locally a tree-like structure. The Bethe approximation [37] can then be used with possibly loop corrections. Again this corresponds to having a weakly correlated uni-modal probability measure and these kind of approaches are referred as pseudo-moment matching methods in the literature for the reason

explained in the previous section. For example the approaches proposed in [24, 35, 28, 36] are based on this assumption.

- In the case where a multi-modal distribution is expected, then a model with many attraction basins is to be found and Hopfield-like models [19, 8] are likely to be more relevant. To be mentioned also is a recent mean-field methods [32] which allows one to find in some simple cases the Ising couplings of a low temperature model, i.e. displaying multiple probabilistic modes.

On the side of inverse Gaussian problem, not surprisingly similar methods have been developed by explicit performing L_0 and L_1 matrix norm penalizations on the inverse covariance matrices, so as to determine sparse non-zero couplings in estimated inverse covariance matrices for large-scale statistical inference applications [13, 20] where direct inversion is not amenable. In our context the goal is a bit different. In general cases, the underlying inverse covariance matrix is not necessarily sparse. What we aim to find is a good sparse approximation to the exact inverse covariance matrix. Furthermore, sparsity constraint is not enough for constructing graph structure that is used in conjunction with BP, known sufficient conditions referred as walk-summability [26] (WS) are likely to be imposed instead of (or in addition to) the sparsity constraint. To the best of our knowledge not much work taking this point into consideration at the noticeable exception of [2] by restricting the class of learned graph structures. To complete this overview, let us mention also that some authors proposed information based structure learning methods [30] quite in line with some approaches to be discussed in the present paper.

In some preceding work dealing with a road traffic inference application, with large scale and real time specifications [16, 15, 14], we have noticed that these methods could not be used blindly without drastic adjustment, in particular to be compatible with belief propagation. This led us to develop some heuristic models related to the Bethe approximation. The present work is an attempt to give a theoretical basis and firmer ground to these heuristics.

2 Preliminaries

2.1 Inverse Ising problem

In this section we consider binary variables ($x_i \in \{0,1\}$), which at our convenience may be also written as spin variables $s_i = 2x_i - 1 \in \{-1,1\}$. We assume that from a set of historical observations, the empirical mean $\hat{m}_i$ (resp. covariance $\hat{\chi}_{ij}$) is given for each variable s_i (resp. each pair of variable (s_i, s_j)). In this case, from Jayne's maximum entropy principle [23], imposing these moments to the joint distribution leads to a model pertaining to the exponential family, the so-called Ising model:

$$\mathcal{P}(\mathbf{s}) = \frac{1}{Z[\mathbf{J},\mathbf{h}]} \exp\Big(\sum_i h_i s_i + \sum_{i,j} J_{ij} s_i s_j\Big) \tag{2.1}$$

where the local fields $\mathbf{h} = \{h_i\}$ and the coupling constants $\mathbf{J} = \{J_{ij}\}$ are the Lagrange multipliers associated respectively to mean and covariance constraints when maximizing the entropy of $\mathcal{P}$. They are obtained as minimizers of the dual optimization problem, namely

$$(\mathbf{h}^\star, \mathbf{J}^\star) = \underset{(\mathbf{h},\mathbf{J})}{\operatorname{argmin}}\, \mathcal{L}[\mathbf{h},\mathbf{J}], \tag{2.2}$$

with

$$\mathcal{L}[\mathbf{h},\mathbf{J}] = \log Z[\mathbf{h},\mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{ij} J_{ij} \hat{m}_{ij} \tag{2.3}$$

the log likelihood. This leads to invert the linear response equations:

$$\frac{\partial \log Z}{\partial h_i}[\mathbf{h}, \mathbf{J}] = \hat{m}_i \tag{2.4}$$

$$\frac{\partial \log Z}{\partial J_{ij}}[\mathbf{h}, \mathbf{J}] = \hat{m}_{ij}, \tag{2.5}$$

$\hat{m}_{ij} = \hat{m}_i \hat{m}_j + \hat{\chi}_{ij}$ being the empirical expectation of $s_i s_j$. As noted e.g. in [7], the solution is minimizing the cross entropy, a Kullback-Leibler distance between the empirical distribution $\hat{P}$ based on historical data and the Ising model:

$$D_{KL}[\hat{\mathcal{P}} \| \mathcal{P}] = \log Z[\mathbf{h}, \mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{i<j} J_{ij} \hat{m}_{ij} - S(\hat{\mathcal{P}}). \tag{2.6}$$

The set of equations (2.4,2.5) cannot be solved exactly in general because the computational cost of Z is exponential. Approximations resorting to various mean field methods can be used to evaluate $Z[\mathbf{h}, \mathbf{J}]$.

Plefka's expansion To simplify the problem, it is customary to make use of the Gibbs free energy, i.e. the Legendre transform of the free energy, to impose the individual expectations $\mathbf{m} = \{\hat{m}_i\}$ for each variable:

$$G[\mathbf{m}, \mathbf{J}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} + F[\mathbf{h}(\mathbf{m}), \mathbf{J}]$$

(with $F[\mathbf{h}, \mathbf{J}] \stackrel{\text{def}}{=} -\log Z[\mathbf{h}, \mathbf{J}]$, $\mathbf{h}^T\mathbf{m}$ is the ordinary scalar product) where $\mathbf{h}(\mathbf{m})$ depends implicitly on $\mathbf{m}$ through the set of constraints

$$\frac{\partial F}{\partial h_i} = -m_i. \tag{2.7}$$

Note that by duality we have

$$\frac{\partial G}{\partial m_i} = h_i(\mathbf{m}), \tag{2.8}$$

and

$$\Big[\frac{\partial^2 G}{\partial m_i \partial m_j}\Big] = \Big[\frac{d\mathbf{h}}{d\mathbf{m}}\Big]_{ij} = \Big[\frac{d\mathbf{m}}{d\mathbf{h}}\Big]^{-1}_{ij} = -\Big[\frac{\partial^2 F}{\partial h_i \partial h_j}\Big]^{-1} = \big[\chi^{-1}\big]_{ij}. \tag{2.9}$$

i.e. the inverse susceptibility matrix. Finding a set of J_{ij} satisfying this last relation along with (2.8) yields a solution to the inverse Ising problem since the m's and χ's are given. Still a way to connect the couplings directly with the covariance matrix is given by the relation

$$\frac{\partial G}{\partial J_{ij}} = -m_{ij}. \tag{2.10}$$

The Plefka expansion is used to expand the Gibbs free energy in power of the coupling J_{ij} assumed to be small. Multiplying all coupling J_{ij} by α yields the following cluster expansion:

$$G[\mathbf{m}, \alpha\mathbf{J}] = \mathbf{h}^T(\mathbf{m}, \alpha)\mathbf{m} + F[\mathbf{h}(\mathbf{m}, \alpha), \alpha\mathbf{J}] \tag{2.11}$$

$$= G_0[\mathbf{m}] + \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} G_n[\mathbf{m}, \mathbf{J}] \tag{2.12}$$

where each term G_n corresponds to cluster contributions of size n in the number of links J_{ij} involved, and $\mathbf{h}(\mathbf{m},\alpha)$ depends implicitly on α in order to always fulfill (2.7). This precisely is the Plefka expansion, and each term of the expansion (2.12) can be obtained by successive derivation of (2.11). We have

$$G_0[\mathbf{m}] = \sum_i \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2}.$$

Letting

$$H_J \stackrel{\text{def}}{=} \sum_{i<j} J_{ij}s_is_j,$$

using (2.7), the two first derivatives of (2.11) w.r.t α read

$$\frac{dG[\mathbf{m},\alpha\mathbf{J}]}{d\alpha} = -\mathbb{E}_\alpha\big(H_J\big), \tag{2.13}$$

$$\frac{d^2G[\mathbf{m},\alpha\mathbf{J}]}{d\alpha^2} = -\operatorname{Var}_\alpha\big(H_J\big) - \sum_i \frac{dh_i(\mathbf{m},\alpha)}{d\alpha}\operatorname{Cov}_\alpha\big(H_J,s_i\big), \tag{2.14}$$

where subscript α indicates that expectations, variance and covariance are taken at given α. To get successive derivatives of $\mathbf{h}(\mathbf{m},\alpha)$ one can use (2.8). Another possibility is to express the fact that $\mathbf{m}$ is fixed,

$$\begin{aligned}\frac{dm_i}{d\alpha} = 0 &= -\frac{d}{d\alpha}\frac{\partial F[\mathbf{h}(\alpha),\alpha\mathbf{J}]}{\partial h_i}\\ &= \sum_j h_j'(\alpha)\operatorname{Cov}_\alpha(s_i,s_j) + \operatorname{Cov}_\alpha(H_J,s_i),\end{aligned}$$

giving

$$h_i'(\alpha) = -\sum_j [\chi_\alpha^{-1}]_{ij}\operatorname{Cov}_\alpha(H_J,s_j), \tag{2.15}$$

where χ_α is the susceptibility delivered by the model when $\alpha \neq 0$. To get the first two terms in the Plefka expansion, we need to compute these quantities at $\alpha = 0$:

$$\begin{aligned}\operatorname{Var}\big(H_J\big) &= \sum_{i<k,j} J_{ij}J_{jk}m_im_k(1-m_j^2) + \sum_{i<j} J_{ij}^2(1-m_i^2m_j^2),\\ \operatorname{Cov}\big(H_J,s_i\big) &= \sum_j J_{ij}m_j(1-m_i^2),\\ h_i'(0) &= -\sum_j J_{ij}m_j,\end{aligned}$$

(by convention $J_{ii}=0$ in these sums). The first and second orders then finally read:

$$G_1[\mathbf{m},\mathbf{J}] = -\sum_{i<j} J_{ij}m_im_j, \qquad G_2[\mathbf{m},\mathbf{J}] = -\sum_{i<j} J_{ij}^2(1-m_i^2)(1-m_j^2),$$

and correspond respectively to the mean field and to the TAP approximation. Higher order terms have been computed in [17].

At this point we are in position to find an approximate solution to the inverse Ising problem, either by inverting equation (2.9) or (2.10). To get a solution at a given order n in the coupling, solving (2.10) requires G at order $n+1$, while it is needed at order n in (2.9).

Taking the expression of G up to second order gives

$$\frac{\partial G}{\partial J_{ij}} = -m_i m_j - J_{ij}(1-m_i^2)(1-m_j^2),$$

and (2.10) leads directly to the basic mean-field solution:

$$J_{ij}^{MF} = \frac{\hat{\chi}_{ij}}{(1-\hat{m}_i^2)(1-\hat{m}_j^2)}. \tag{2.16}$$

At this level of approximation for G, using (2.8) we also have

$$h_i = \frac{1}{2}\log\frac{1+m_i}{1-m_i} - \sum_j J_{ij}m_j + \sum_j J_{ij}^2 m_i(1-m_j^2)$$

which corresponds precisely to the TAP equations. Using now (2.9) gives

$$\frac{\partial h_i}{\partial m_j} = [\chi^{-1}]_{ij} = \delta_{ij}\big(\frac{1}{1-m_i^2} + \sum_k J_{ik}^2(1-m_k^2)\big) - J_{ij} - 2J_{ij}^2 m_i m_j.$$

Ignoring the diagonal terms, the TAP solution is conveniently expressed in terms of the inverse empirical susceptibility,

$$J_{ij}^{TAP} = -\frac{2[\hat{\chi}^{-1}]_{ij}}{1+\sqrt{1-8\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}}, \tag{2.17}$$

where the branch corresponding to a vanishing coupling in the limit of small correlation i.e. small $\hat{\chi}_{ij}$ and $[\hat{\chi}^{-1}]_{ij}$ for $i \neq j$, has been chosen.

Bethe approximate solution When the graph formed by the pairs (i,j) for which the correlations $\hat{\chi}_{ij}$ are given by some observations is a tree, the following form of the joint probability corresponding to the Bethe approximation:

$$\mathcal{P}(\mathbf{x}) = \prod_{i<j}\frac{\hat{p}_{ij}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}\prod_i \hat{p}_i(x_i), \tag{2.18}$$

yields actually an exact solution to the inverse problem (2.2), where the $\hat{p}$ are the single and pair variables empirical marginal given by the observations. Using the following identity

$$\begin{aligned}\log\frac{\hat{p}_{ij}(s_i,s_j)}{\hat{p}_i(s_i)\hat{p}_j(s_j)} = {} & \frac{(1+s_i)(1+s_j)}{2}\log\frac{\hat{p}_{ij}^{11}}{\hat{p}_i^1\hat{p}_j^1} + \frac{(1+s_i)(1-s_j)}{2}\log\frac{\hat{p}_{ij}^{10}}{\hat{p}_i^1\hat{p}_j^0} \\ & + \frac{(1-s_i)(1+s_j)}{2}\log\frac{\hat{p}_{ij}^{01}}{\hat{p}_i^0\hat{p}_j^1} + \frac{(1-s_i)(1-s_j)}{2}\log\frac{\hat{p}_{ij}^{00}}{\hat{p}_i^0\hat{p}_j^0}\end{aligned}$$

where the following parametrization of the $\hat{p}$'s

$$\hat{p}_i^x \stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x\big) = \frac{1}{2}(1+\hat{m}_i(2x-1)), \tag{2.19}$$

$$\begin{aligned}\hat{p}_{ij}^{xy} &\stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x, \frac{1+s_j}{2} = y\big) \\ &= \frac{1}{4}(1+\hat{m}_i(2x-1)+\hat{m}_j(2y-1)+\hat{m}_{ij}(2x-1)(2y-1)\end{aligned} \tag{2.20}$$

relating the empirical frequency statistics to the empirical "magnetizations" $m \equiv \hat{m}$, can be used. Summing up the different terms gives us the mapping onto an Ising model (2.1) with

$$h_i = \frac{1-d_i}{2}\log\frac{\hat{p}_i^1}{\hat{p}_i^0} + \frac{1}{4}\sum_{j\in\partial i}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{10}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{00}}\Big), \tag{2.21}$$

$$J_{ij} = \frac{1}{4}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{00}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}}\Big), \qquad \forall\ (i,j)\in\mathcal{E}, \tag{2.22}$$

where d_i is the number of neighbors of i, using the notation $j \in \partial i$ for "j neighbor of i". The partition function is then explicitly given by

$$Z_{Bethe}[\hat{p}] = \exp\Big[-\frac{1}{4}\sum_{(i,j)\in\mathcal{E}}\log(\hat{p}_{ij}^{00}\ \hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}\ \hat{p}_{ij}^{11}) - \sum_i \frac{1-d_i}{2}\log(\hat{p}_i^0\ \hat{p}_i^1)\Big] \tag{2.23}$$

The corresponding Gibbs free energy can then be written explicitly using (2.21–2.23). With fixed magnetizations m_i's, and given a set of couplings $\{J_{ij}\}$, the parameters m_{ij} are implicit function

$$m_{ij} = m_{ij}(m_i, m_j, J_{ij}),$$

obtained by inverting the relations (2.22). For the linear response, we get from (2.21) a result derived first in [35]:

$$\begin{aligned}\frac{\partial h_i}{\partial m_j} = \Big[&\frac{1-d_i}{1-m_i^2}\\ &+\frac{1}{16}\sum_{k\in\partial i}\Big(\big(\frac{1}{\hat{p}_{ik}^{11}}+\frac{1}{\hat{p}_{ik}^{01}}\big)\big(1+\frac{\partial m_{ik}}{\partial m_i}\big)+\big(\frac{1}{\hat{p}_{ik}^{00}}+\frac{1}{\hat{p}_{ik}^{10}}\big)\big(1-\frac{\partial m_{ik}}{\partial m_i}\big)\Big)\Big]\delta_{ij}\\ &+\frac{1}{16}\Big(\big(\frac{1}{\hat{p}_{ij}^{11}}+\frac{1}{\hat{p}_{ij}^{10}}\big)\big(1+\frac{\partial m_{ij}}{\partial m_i}\big)+\big(\frac{1}{\hat{p}_{ij}^{00}}+\frac{1}{\hat{p}_{ij}^{01}}\big)\big(1-\frac{\partial m_{ij}}{\partial m_i}\big)\Big)\Big]\delta_{j\in\partial i}.\end{aligned}$$

Using (2.22), we can also express

$$\frac{\partial m_{ij}}{\partial m_i} = -\frac{\frac{1}{\hat{p}_{ij}^{11}}+\frac{1}{\hat{p}_{ij}^{01}}-\frac{1}{\hat{p}_{ij}^{10}}-\frac{1}{\hat{p}_{ij}^{00}}}{\frac{1}{\hat{p}_{ij}^{11}}+\frac{1}{\hat{p}_{ij}^{01}}+\frac{1}{\hat{p}_{ij}^{10}}+\frac{1}{\hat{p}_{ij}^{00}}},$$

so that with little assistance of Maple, we may finally reach the expression given in [31]

$$\begin{aligned}[\hat{\chi}^{-1}]_{ij} = \Big[&\frac{1-d_i}{1-m_i^2}+\sum_{k\in\partial i}\frac{1-m_k^2}{(1-m_i^2)(1-m_k^2)-\hat{\chi}_{ik}^2}\Big]\delta_{ij}\\ &-\frac{\hat{\chi}_{ij}}{(1-m_i^2)(1-m_j^2)-\hat{\chi}_{ij}^2}\ \delta_{j\in\partial i},\end{aligned} \tag{2.24}$$

equivalent to the original one derived in [35], albeit written in a different form, more suitable to discuss the inverse Ising problem. This expression is quite paradoxical since the inverse of the $[\chi]_{ij}$ matrix, which coefficients appear on the right hand side of this equation, should coincide with the left hand side, given as input of the inverse Ising problem. The existence of an exact solution can therefore be checked directly as a self-consistency property of the input data $\hat{\chi}_{ij}$: for a given pair (i,j) either:

- $[\hat{\chi}^{-1}]_{ij} \neq 0$, then this self-consistency relation (2.24) has to hold and J_{ij} is given by (2.22) using $\hat{m}_{ij} = \hat{m}_i \hat{m}_j + \hat{\chi}_{ij}$.
- $[\hat{\chi}^{-1}]_{ij} = 0$ then $J_{ij} = 0$ but $\hat{\chi}_{ij}$, which can be non-vanishing, is obtained by inverting $[\hat{\chi}^{-1}]$ defined by (2.24).

Finally, complete consistency of the solution is checked on the diagonal elements in (2.24). If full consistency is not verified, this equation can nevertheless be used to find approximate solutions. Remark that, if we restrict the set of equations (2.24), e.g. by some thresholding procedure, in such a way that the corresponding graph is a spanning tree, then, by construction, $\chi_{ij} \equiv \hat{\chi}_{ij}$ will be solution on this restricted set of edges, simply because the BP equations are exact on a tree. The various methods proposed for example in [28, 36] actually correspond to different heuristics for finding approximate solutions to this set of constraints. As noted in [31], a direct way to proceed is to eliminate χ_{ij} in the equations obtained from (2.22) and (2.24):

$$\chi_{ij}^2 + 2\chi_{ij}(m_i m_j - \coth(2J_{ij})) + (1-m_i^2)(1-m_j^2) = 0$$

$$\chi_{ij}^2 - \frac{\chi_{ij}}{[\chi^{-1}]_{ij}} - (1-m_i^2)(1-m_j^2) = 0.$$

This leads directly to

$$J_{ij}^{Bethe} = -\frac{1}{2}\,\mathrm{atanh}\Big(\frac{2[\hat{\chi}^{-1}]_{ij}}{\sqrt{1+4(1-\hat{m}_i^2)(1-\hat{m}_j^2)[\hat{\chi}^{-1}]_{ij}^2} - 2\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}\Big), \tag{2.25}$$

while the corresponding computed of χ_{ij}, instead of the observed one $\hat{\chi}_{ij}$, has to be inserted in (2.21) to be fully consistent. Note that J_{ij}^{Bethe} and J_{ij}^{TAP} coincide at second order in $[\hat{\chi}^{-1}]_{ij}$.

Hopfield model As mentioned in the introduction when the distribution to be modeled is multi-modal, the situation corresponds to finding an Ising model in the low temperature phase with many modes, referred to as Mattis states in the physics literature. Previous methods assume implicitly a high temperature where only one single mode, "the paramagnetic state" is selected. The Hopfield model, introduced originally to model auto-associative memories, is a special case of an Ising model, where the coupling matrix is of low rank $p \leq N$ and corresponds to the sum of outer products of p given spin vectors $\{\xi^k, k = 1 \dots p\}$, each one representing a specific attractive pattern:

$$J_{ij} = \frac{1}{p}\sum_{k=1}^{N} \xi_i^k \xi_j^k$$

In our inference context, these patterns are not given directly, the input of the model being the covariance matrix. In [8] these couplings are interpreted as the contribution stemming from the p largest principle axes of the correlation matrix. This lead in particular the authors to propose an extension of the Hopfield model by introducing repulsive patterns to take as well into account the smallest principal axes. Assuming small patterns coefficients $|\xi^k| < 1/\sqrt{N}$, they come up with the following couplings with highest likelihood:

$$J_{ij}^{Hopfield} \equiv \frac{1}{\sqrt{(1-m_i^2)(1-m_j^2)}} \sum_{k=1}^{p} \Big((1 - \frac{1}{\lambda_k}) v_i^k v_j^k - (\frac{1}{\lambda_{N-k}} - 1) v_i^{N-k} v_j^{N-k} \Big)$$

at first order of the perturbation. At this order of approximation the local fields are given by

$$h_i = \tanh(m_i) - \sum_j J_{ij}^{Hopfield} m_j.$$

In a previous study [15] we found a connection between the plain direct BP method with the Hopfield model, by considering a 1-parameter deformation of the Bethe inference model (2.18)

$$\mathcal{P}(\mathbf{x}) = \prod_{i<j} \Big(\frac{\hat{p}_{ij}(x_i, x_j)}{\hat{p}(x_i)\hat{p}(x_j)} \Big)^{\alpha} \prod_i \hat{p}_i(x_i), \tag{2.26}$$

with $\alpha \in [0, 1]$. We observed indeed that when the data corresponds to some multi-modal measure with well separated components, this measure coincides asymptotically with an Hopfield model made only of attractive pattern, representative of each component of the underlying measure. α represents basically the inverse temperature of the model and is easy to calibrate in practice.

2.2 More on the Bethe susceptibility

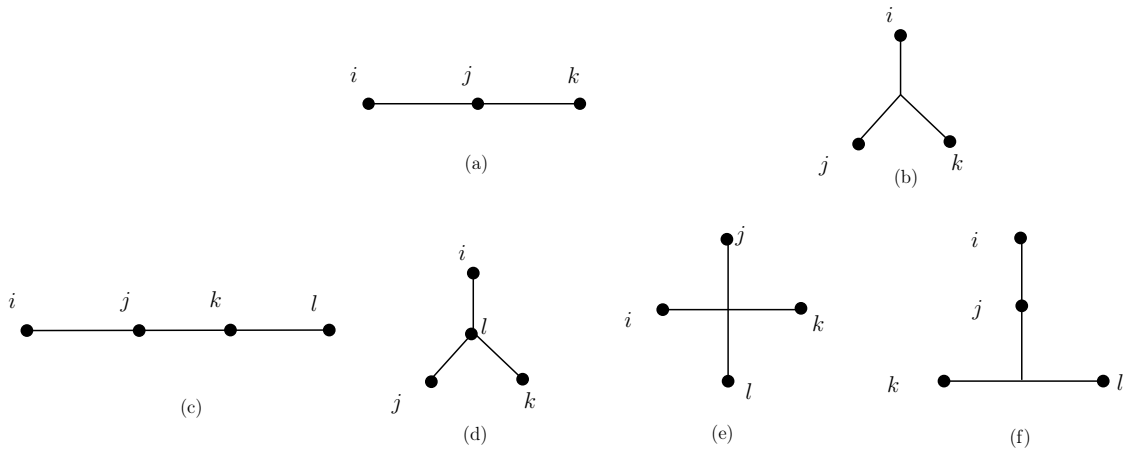


Figure 2.1: Various cumulant topologies of order three (a,b) and four (c-f).

The explicit relation (2.24) between susceptibility and inverse susceptibility coefficients is not the only one that can be obtained. In fact, it is the specific property of a singly connected factor graph that two variables x_i and x_j, conditionally to a variable x_k are independent if k is on the path between i and j along the tree:

$$p(x_i, x_j, x_k) = p(x_i|x_k)p(x_j|x_k)p(x_k) = \frac{p(x_i, x_k)p(x_j, x_k)}{p(x_k)}$$

Using the parametrization (2.19,2.20) with $m_{ij} = m_i m_j + \chi_{ij}$ yields immediately

$$\chi_{ij} = \frac{\chi_{ik}\chi_{jk}}{1 - m_k^2}, \qquad \forall\ k \in (i, j) \text{ along } \mathcal{T}. \tag{2.27}$$

By recurrence we get, as noticed in e.g. [29], given the path $i_0 = i, i_1, \ldots, i_{n+1} = j$ between i and j along the tree $\mathcal{T}$

$$\chi_{ij} = \frac{\prod_{a=0}^{n} \chi_{i_a i_{a+1}}}{\prod_{a=1}^{n}(1 - m_{i_a}^2)},$$

reflecting the factorization of the joint measure. This expression actually coincides with (2.24) only on a tree. On a loopy graph, this last expression should be possibly replaced by a sum over paths.

Higher order susceptibility coefficients are built as well in terms of elementary building blocks given by the pairwise susceptibility coefficients χ_{ij}. The notations generalize into the following straightforward manner:

$$\begin{aligned} m_{ijk} &\stackrel{\text{def}}{=} \mathbb{E}(s_i s_j s_k) \stackrel{\text{def}}{=} m_i m_j m_k + m_i \chi_{jk} + m_j \chi_{ik} + m_k \chi_{ij} + \chi_{ijk} \\ m_{ijkl} &\stackrel{\text{def}}{=} \mathbb{E}(s_i s_j s_k s_l) \stackrel{\text{def}}{=} m_i m_j m_k m_l \\ &\quad + m_i m_j \chi_{kl} + m_i m_k \chi_{jl} + m_i m_l \chi_{jk} + m_j m_k \chi_{il} + m_j m_l \chi_{ik} + m_k m_l \chi_{ij} \\ &\quad + m_i \chi_{jkl} + m_j \chi_{ikl} + m_k \chi_{ijl} + m_l \chi_{ijk} + \chi_{ijkl}, \end{aligned}$$

where χ_{ijk} and χ_{ijkl} are respectively three and four point susceptibilities. The quantities being related to the corresponding marginals similarly to (2.19,2.20):

$$\begin{aligned} p(s_i, s_j, s_k) &= \frac{1}{8}\big(1 + m_i s_i + m_j s_j + m_k s_k \\ &\quad + m_{ij} s_i s_j + m_{ik} s_i s_k + m_{jk} s_j s_k + m_{ijk} s_i s_j s_k\big) \\ p(s_i, s_j, s_k, s_l) &= \frac{1}{16}\big(1 + m_i s_i + m_j s_j + m_k s_k + m_l s_l \\ &\quad + m_{ij} s_i s_j + m_{ik} s_i s_k + m_{il} s_i s_l + m_{jk} s_j s_k + m_{jl} s_j s_l + m_{kl} s_k s_l \\ &\quad + m_{ijk} s_i s_j s_k + m_{ijl} s_i s_j s_l + m_{ikl} s_i s_k s_l + m_{jkl} s_j s_k s_l + m_{ijkl} s_i s_j s_k s_l\big) \end{aligned}$$

Using the basic fact that, on the tree

$$p(s_i, s_j, s_k) = \frac{p(s_i, s_j) p(s_j, s_k)}{p(s_i)}$$

when j is on the path $\widehat{ik}$ given by $\mathcal{T}$, and

$$p(s_i, s_j, s_k) = \sum_{s_l} \frac{p(s_i, s_l) p(s_j, s_l) p(s_k, s_l)}{p(s_l)^2}$$

when path $\widehat{ij}$, $\widehat{ik}$ and $\widehat{jk}$ along $\mathcal{T}$ intersect on vertex l, we obtain

$$\chi_{ijk} = \begin{cases} -2\dfrac{m_l}{(1-m_l^2)^2} \chi_{il}\chi_{jl}\chi_{kl} & \text{with } \{l\} = (i,j) \cap (i,k) \cap (j,k) \text{ along } \mathcal{T}, \\ -2 m_j \chi_{ik} & \text{if } j \in (i,k) \text{ along } \mathcal{T}. \end{cases}$$

For the fourth order, more cases have to be distinguished. When i, j, k and l are aligned as on Figure 2.1.c, in this order on the path $\widehat{il}$ along $\mathcal{T}$ we have

$$p(s_i, s_j, s_k, s_l) = \frac{p(s_i, s_j) p(s_j, s_k, s_l)}{p(s_j)^2}$$

which leads to

$$\chi_{ijkl} = 4 m_k m_j \chi_{il} - \chi_{ik}\chi_{jl} - \chi_{il}\chi_{jk}.$$

When path $\widehat{ij}$, $\widehat{ik}$ and $\widehat{jk}$ along $\mathcal{T}$ intersect on vertex l as on Figure 2.1.d, we obtain instead [1]

$$\chi_{ijkl} = 2\frac{3m_l^2 - 1}{1 - m_l^2} \chi_{ij}\chi_{kl}.$$

[1] This apparently non-symmetric expression can be symmetrized with help of (2.27).

For the situation corresponding to Figure 2.1.e, we have

$$p_(s_i, s_j, s_k, s_l) = \sum_{s_q} \frac{p(s_i, s_j, s_q)p(s_q, s_k, s_l)}{p(s_q)^2}$$

which leads to

$$\chi_{ijkl} = 2\frac{3m_q^2 - 1}{1 - m_q^2}\chi_{ij}\chi_{kl}.$$

Finally, for the situation corresponding to Figure 2.1.f, we have

$$p_(s_i, s_j, s_k, s_l) = \sum_{s_q} \frac{p(s_i, s_j)p(s_j, s_k, s_l)}{p(s_j)^2}$$

leading to

$$\chi_{ijkl} = -\chi_{ik}\chi_{jl} - \chi_{jk}\chi_{il} + 4\frac{m_k m_q}{1 - m_q^2}\chi_{ij}\chi_{lq}.$$

2.3 Sparse inverse estimation of covariance matrix

Let us leave the Ising modeling issue aside for a while and introduce another related graph selection problem, named sparse inverse covariance estimation, which is defined on real continuous random variables. This method aims at constructing a sparse factor graph structure by identifying conditionally independent pairs of nodes in the graph, given empirical covariances of random variables. Assuming that all nodes in the graph follow a joint multi-variate Gaussian distribution with mean μ and covariance matrix C, the existing correlation between the nodes i and j given all the other nodes in the graph are indicated by the non-zero ijth entry of the precision matrix C^{-1}, while zero entries correspond to independent pairs of variables. Therefore, under the joint normal distribution assumption, selection of factor graph structures amounts to finding the sparse precision matrix that best describes the underlying data distribution, given the fixed empirical covariance matrix. When the derived inverse estimation is sparse, it becomes easier to compute marginal distribution of each random variable and conduct statistical inference. To achieve that goal, optimizations methods have been developed based on L_0 or L_1 norm penalty for the estimation of C^{-1}, to enhance its sparse structure constraint on the estimated inverse of covariance matrix and discover underlying conditionally independent parts.

Let $\hat{C} \in \mathbb{R}^{n\times n}$ be the empirical covariance matrix of n random variables (represented as the nodes in the graph model). The sparsity penalized maximum likelihood estimation A of the precision matrix C^{-1} can be derived by solving the following positive definite cone program:

$$A = \underset{X \succ 0}{\operatorname{argmin}} -\mathcal{L}(X) + \lambda P(X) \tag{2.28}$$

where

$$\mathcal{L}(A) \stackrel{\text{def}}{=} \log\det(A) - \operatorname{Tr}(A\hat{C}), \tag{2.29}$$

is the log likelihood of the distribution defined by A, $\log\det$ being the logarithm of the determinant, and $P(A)$ is a sparsity inducing regularization term [13]. λ is the regularization coefficient balancing the data-fitting oriented likelihood and sparsity penalty. Since the precision matrix of joint normal distribution should be positive definite, any feasible solution to this optimization problem is thus required to locate within a positive definite cone. The penalty term $P(A)$ is typically constructed using sparsity inducing matrix norm, also known as sparse learning in the domain of statistical learning. There are two typical configurations of $P(A)$:

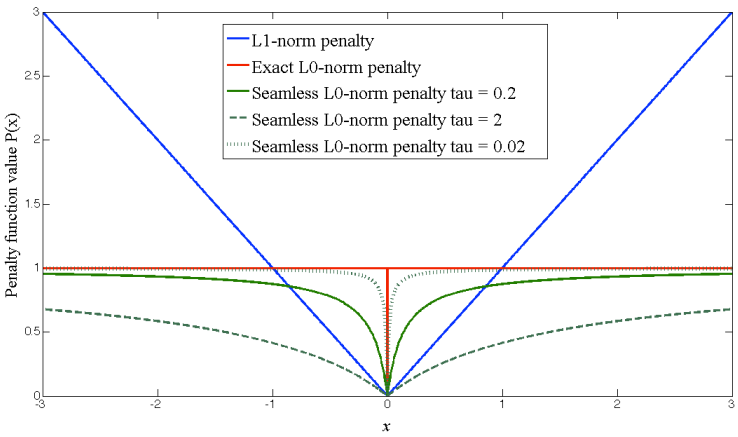


Figure 2.2: Demonstration of L_1 norm, exact L_0 norm and Seamless L_0 norm based penalty function.

- The L_0 norm $\|A\|_{L_0}$ of the matrix X, which counts the number of non-zero elements in the matrix. It is also known as the cardinality or the non-zero support of the matrix. Given its definition, it is easy to find that L_0 norm based penalty is a strong and intuitive appeal for sparsity structure of the estimated precision matrix. However, it is computationally infeasible to solve exact L_0-norm minimization directly, due to the fact that exact L_0 norm penalty is discontinuous and non-differentiable. In practice, one either uses a continuous approximation to the form of the L_0-penalty, or solve it using a greedy method. Due to the non-convexity of the exact L_0 norm penalty, only a local optimum of the feasible solution can be guaranteed. Nevertheless, L_0 norm penalty usually leads to much sparser structure in the estimation, while local optimum is good enough for most practical cases.
- The L_1 matrix norm $\|A\|_{L_1} = \sum_{i,j}^{n} |A_{ij}|$. L_1 norm penalty was firstly introduced into the standard least square estimation problem by Tibshiharni [34], under the name "Lasso regression". Minimizing the L_1 norm based penalty encourages sparse non-zero entries in the estimated precision matrix A, which achieves a selection of informative variables for regression and reduces complexity of regression model efficiently. Further extension of the basic L_1-norm penalty function allows one assigning different non-negative weight values λ_{ij} to different entries A_{ij}, as $\sum_{i,j}^{n} \lambda_{ij}|A_{ij}|$. This weighted combination can constrain the sparse penalties only on the off-diagonal entries, so as to avoid unnecessary sparsity on the diagonal elements. Furthermore, this extension allows us to introduce prior knowledge about the conditional independence structure of the graph into the joint combination problem.

For further understanding of the relation between the exact L_0 and L_1 norm penalty, we illustrate them with respect to one scalar variable in Figure 2.2. As we can see, within $[-1, 1]$, L_1-norm penalty plays as a convex envelope of the exact L_0-norm penalty. Due to the convexity property of L_1 norm penalty, the global optimum of the convex programming problem can be achieved with even linear computational complexity [34, 13]. However, although L_1 norm based penalty leads to computationally sound solutions to the original issue, it also introduces modeling bias into the penalized maximum likelihood estimation. As illustrated in the figure, when the underlying true values of the matrix entries are sufficiently large, the corresponding L_1 norm

based regularization performs linearly increased penalty to those entries, which thus results in a severe bias w.r.t. the maximum likelihood estimation [12]. In contrast, L_0 norm based penalty avoids such issue by constraining the penalties of non-zero elements to be constant. It has been proved in [34] that the L_1 norm penalty discovers the underlined sparse structure when some suitable assumptions are satisfied. However, in general cases, the quality of the solutions is not clear.

3 Link selection at the Bethe point

3.1 Bethe approximation and Graph selection

As observed in the previous section, when using the Bethe approximation to find an approximate solution to the IIP, the consistency check should then be that either the factor graph be sparse, nearly a tree, either the coupling are small. There are then two distinct ways of using the Bethe approximation:

- the direct way, where the form of the joint distribution (2.18) is assumed with a complete graph. There is then a belief propagation fixed point for which the beliefs satisfy all the constraints. This solution to be meaningful requires small correlations, so that the belief propagation fixed point be stable and unique, allowing the corresponding log likelihood to be well approximated. Otherwise, this solution is not satisfactory, but a pruning procedure, which amounts to select a sub-graph based on mutual information, can be used. The first step is to find the MST with these weights. Adding new links to this baseline solution in a consistent way is the subject of the next section.
- the indirect way consists in first inverting the potentially non-sparse correlation matrix. If the underlying interaction matrix is actually a tree, this will be visible in the inverse correlation matrix, indicated directly by the non-zero entries. This seems to work better than the previous one when no sparsity but weak coupling is assumed. It corresponds in fact to the equations solved iteratively by the susceptibility propagation algorithm [28].

Let us first justify the intuitive assertion concerning the optimal model with tree like factor graphs, which was actually first proposed in [6] and which is valid for any type of MRF.

Proposition 3.1. *The optimal model with tree like factor graphs to the inverse pairwise MRF is obtained by finding the MST on the graph of weighted links with weights given by mutual information between variables.*

Proof. On a tree, the Bethe distribution relating marginal p_i of single and pair p_{ij} variables distributions

$$\mathcal{P}(\mathbf{x}) = \prod_{(i,j)\in\mathcal{T}} \frac{p_{ij}(x_i,x_j)}{p_i(x_i)p_j(x_j)} \prod_i p_i(x_i)$$

is exact. Assuming first that the associated factor graph is given by a tree $\mathcal{T}$, the expression (2.6) of the log likelihood leads to the optimal solution given that tree:

$$p_i = \hat{p}_i \qquad \text{and} \qquad p_{ij} = \hat{p}_{ij}, \qquad \forall i \in V,\ (i,j) \in \mathcal{T}.$$

In turn we have the following expression for the log likelihood:

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{T}} I_{ij} - \sum_i H_i,$$

with

$$H_i \stackrel{\text{def}}{=} -\sum_{x_i} \hat{p}_i(x_i) \log \hat{p}_i(x_i)$$

$$I_{ij} \stackrel{\text{def}}{=} \sum_{x_i,x_j} \hat{p}_{ij}(x_i,x_j) \log \frac{\hat{p}_{ij}(x_i,x_j)}{\hat{p}_i(x_i)\hat{p}_j(x_j)}.$$

respectively the single variable entropy and the mutual information between two variables. Since the single variable contributions to the entropy is independent of the chosen graph $\mathcal{T}$, the optimal choice for $\mathcal{T}$ correspond to the statement of the proposition. ■

3.2 Optimal 1-link correction to the Bethe point

Adding one link Suppose now that we want to add one link to the max spanning tree. The question is which link will produce the maximum improvement to the log likelihood. This question is a special case of how to correct a given factor for a general pairwise model. Let $\mathcal{P}_0$ be the reference distribution to which we want to add (or modify) one factor ψ_{ij} to produce the distribution

$$\mathcal{P}_1(\mathbf{x}) = \mathcal{P}_0(\mathbf{x}) \times \frac{\psi_{ij}(x_i,x_j)}{Z_\psi} \tag{3.1}$$

with

$$Z_\psi = \int dx_i dx_j p^0_{ij}(x_i,x_j)\psi_{ij}(x_i,x_j).$$

The log likelihood corresponding to this new distribution now reads

$$\mathcal{L}_1 = \mathcal{L}_0 + \int d\mathbf{x}\hat{\mathcal{P}}(\mathbf{x}) \log \psi_{ij}(x_i,x_j) - \log Z_\psi.$$

The functional derivative w.r.t. ψ reads

$$\frac{\partial \mathcal{L}_1}{\partial \psi_{ij}(x_i,x_j)} = \frac{\hat{p}(x_i,x_j)}{\psi_{ij}(x_i,x_j)} - \frac{p^0_{ij}(x_i,x_j)}{Z_\psi}, \qquad \forall (x_i,x_j) \in \Omega^2.$$

which yield immediately the solution

$$\psi_{ij}(x_i,x_j) = \frac{\hat{p}_{ij}(x_i,x_j)}{p^0_{ij}(x_i,x_j)} \qquad \text{with} \qquad Z_\psi = 1, \tag{3.2}$$

where $p_0(x_i,x_j)$ is the reference marginal distribution obtained from $\mathcal{P}_0$. The correction to the Log likelihood simply reads

$$\Delta\mathcal{L} = D_{KL}(\hat{p}_{ij} \| p^0_{ij}). \tag{3.3}$$

Sorting all the links w.r.t. this quantity yields the (exact) optimal 1-link correction to be made.

As a check, consider the special case of the Ising model. Adding one link amounts to set one J_{ij} to some finite value, but since this will perturb at least the local magnetization of i and j, we have also to modify the local fields h_i and h_j. The modified measure therefore reads

$$\mathcal{P}_{1link} = \mathcal{P}_{Bethe} \frac{\exp\big(J_{ij}s_is_j + h_is_i + h_js_j\big)}{\Delta Z(J,h)},$$

where $\Delta Z(J,h)$ is multiplicative correction factor to the partition function. We have

$$\Delta Z(h_i,h_j,J_{ij}) = z_0 + z_1\hat{m}_i + z_2\hat{m}_j + z_3 m_{ij}^{Bethe},$$

after introducing the following quantities

$$z_0 \stackrel{\text{def}}{=} e^{J_{ij}+h_i+h_j} + e^{-J_{ij}-h_i+h_j} + e^{-J_{ij}+h_i-h_j} + e^{J_{ij}-h_i-h_j}$$

$$z_1 \stackrel{\text{def}}{=} e^{J_{ij}+h_i+h_j} - e^{-J_{ij}-h_i+h_j} + e^{-J_{ij}+h_i-h_j} - e^{J_{ij}-h_i-h_j}$$

$$z_2 \stackrel{\text{def}}{=} e^{J_{ij}+h_i+h_j} + e^{-J_{ij}-h_i+h_j} - e^{-J_{ij}+h_i-h_j} - e^{J_{ij}-h_i-h_j}$$

$$z_3 \stackrel{\text{def}}{=} e^{J_{ij}+h_i+h_j} - e^{-J_{ij}-h_i+h_j} - e^{-J_{ij}+h_i-h_j} + e^{J_{ij}-h_i-h_j}.$$

The correction to the log likelihood is then given by

$$\Delta\mathcal{L}(h_i,h_j,J_{ij}) = \log\Delta Z(h_i,h_j,J_{ij}) - h_i\hat{m}_i - h_j\hat{m}_j - J_{ij}\hat{m}_{ij} \tag{3.4}$$

This is a concave function of h_i, h_j and J_{ij}, and the (unique) maximum is obtained when the following constraints are satisfied:

$$\frac{\partial\log(\Delta Z)}{\partial h_i} = \frac{z_1 + z_0\hat{m}_i + z_3\hat{m}_j + z_2 m_{ij}^{Bethe}}{\Delta Z(h_i,h_j,J_{ij})} = \hat{m}_i,$$

$$\frac{\partial\log(\Delta Z)}{\partial h_j} = \frac{z_2 + z_3\hat{m}_i + z_0\hat{m}_j + z_1 m_{ij}^{Bethe}}{\Delta Z(h_i,h_j,J_{ij})} = \hat{m}_j,$$

$$\frac{\partial\log(\Delta Z)}{\partial J_{ij}} = \frac{z_3 + z_2\hat{m}_i + z_1\hat{m}_j + z_0 m_{ij}^{Bethe}}{\Delta Z(h_i,h_j,J_{ij})} = \hat{m}_{ij}.$$

This constraints can be solved as follows. First let

$$U \stackrel{\text{def}}{=} (1 + \hat{m}_i + \hat{m}_j + m_{ij}^{Bethe})\ e^{2(h_i+h_j)},$$

$$V \stackrel{\text{def}}{=} (1 - \hat{m}_i + \hat{m}_j - m_{ij}^{Bethe})\ e^{2(h_j-J_{ij})},$$

$$W \stackrel{\text{def}}{=} (1 + \hat{m}_i - \hat{m}_j - m_{ij}^{Bethe})\ e^{2(h_i-J_{ij})}.$$

to obtain the linear system

$$U(1-\hat{m}_i) - V(1+\hat{m}_i) + W(1-\hat{m}_i) = (1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})(1+\hat{m}_i),$$

$$U(1-\hat{m}_j) + V(1-\hat{m}_j) - W(1+\hat{m}_j) = (1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})(1+\hat{m}_j),$$

$$U(1-\hat{m}_{ij}) - V(1+\hat{m}_{ij}) - W(1+\hat{m}_{ij}) = -(1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})(1-\hat{m}_{ij}).$$

Inverting this system yields the following solution

$$U = \frac{(1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})}{(1 - \hat{m}_i - \hat{m}_j + \hat{m}_{ij})}(1 + \hat{m}_i + \hat{m}_j + \hat{m}_{ij})$$

$$V = \frac{(1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})}{(1 - \hat{m}_i - \hat{m}_j + \hat{m}_{ij})}(1 - \hat{m}_i + \hat{m}_j - \hat{m}_{ij})$$

$$W = \frac{(1 - \hat{m}_i - \hat{m}_j + m_{ij}^{Bethe})}{(1 - \hat{m}_i - \hat{m}_j + \hat{m}_{ij})}(1 + \hat{m}_i - \hat{m}_j - \hat{m}_{ij})$$

Using the parametrization (2.20), we finally arrive at

$$e^{h_i s_i + h_j s_j + J_{ij} s_i s_j} = \frac{\hat{p}_{ij}(s_i, s_j)}{b_{ij}(s_i, s_j)} \left(\frac{b_{ij}(1,1) b_{ij}(-1,1) b_{ij}(1,-1) b_{ij}(-1,-1)}{\hat{p}_{ij}(1,1) \hat{p}_{ij}(-1,1) \hat{p}_{ij}(1,-1) \hat{p}_{ij}(-1,-1)} \right)^{1/4}.$$

where $b_{ij}(s_i, s_j)$ is the joint marginal of s_i and s_j, obtained from the Bethe reference point. From these expressions, we can assess for any new potential link the increase (3.4) in the log likelihood. After rearranging all terms, it takes indeed as announced the following simple form

$$\Delta\mathcal{L}(h_i, h_j, J_{ij}) = \sum_{s_i, s_j} \hat{p}_{ij} \log \frac{\hat{p}_{ij}(s_i, s_j)}{b_{ij}(s_i, s_j)} = D_{KL}(\hat{p} \| b). \tag{3.5}$$

The interpretation is therefore immediate: the best candidate is the one for which the Bethe solution yields the most distant joint marginal b_{ij} to the targeted one $\hat{p}_{ij}$ given by the data. Note that the knowledge of the $\{b_{ij}, (ij) \notin \mathcal{T}\}$ requires a sparse matrix inversion through equation (2.24), which renders the method a bit expensive in the Ising case. For Gaussian MRF, the situation is different, because in that case the correction to the log likelihood can be evaluated directly by another means. Indeed, the correction factor (3.2) reads in that case

$$\psi_{ij}(x_i, x_j) = \exp\Big(-\frac{1}{2} (x_i, x_j) \big([\hat{C}^{ij}]^{-1} - [C^{ij}]^{-1} \big) (x_i, x_j)^T \Big),$$

where $[\hat{C}^{ij}]$ and $[C^{ij}]$ represent the restricted 2×2 covariance matrix corresponding to the pair (x_i, x_j) of respectively the reference model and the current model specified by precision matrix $A = C^{-1}$. With a small abuse of notation the new model obtained after adding or changing link (ij) reads

$$A' = A + [\hat{C}^{ij}]^{-1} - [C^{ij}]^{-1} \stackrel{\text{def}}{=} A + V.$$

with a log likelihoods variation given by:

$$\Delta\mathcal{L} = \frac{C_{ii}\hat{C}_{jj} + C_{jj}\hat{C}_{ii} - 2C_{ij}\hat{C}_{ij}}{C_{ii}C_{jj} - C_{ij}^2} - 2 - \log \frac{\hat{C}_{ii}\hat{C}_{jj} - \hat{C}_{ij}^2}{C_{ii}C_{jj} - C_{ij}^2}.$$

Let us notice the following useful formula (see e.g. [4]):

$$\begin{aligned} (A + [V^{ij}])^{-1} &= A^{-1} - A^{-1}[V^{ij}](1 + A^{-1}[V^{ij}])^{-1} A^{-1} \\ &= A^{-1} - A^{-1}[V^{ij}](1 + [C^{ij}][V^{ij}])^{-1} A^{-1}, \end{aligned} \tag{3.6}$$

valid for a 2×2 perturbation matrix $[V^{ij}]$. Using this formula, the new covariance matrix reads

$$C' = A'^{-1} = A^{-1} - A^{-1}[C^{ij}]^{-1} \big(1 - [\hat{C}^{ij}][C^{ij}]^{-1} \big) A^{-1}. \tag{3.7}$$

Therefore the number of operations needed to maintain the covariance matrix after each add-on is $\mathcal{O}(N^2)$.

Let us now examine under which condition adding/modifying links in this way let the covariance matrix remain positive semi-definite. By adding a 2×2 matrix, we expect a quadratic

correction to the determinant:

$$\begin{aligned}\det(A') &= \det(A)\det(1+A^{-1}V)\\ &= \det(A)\Big[1+V_{ii}C_{ii}+V_{ij}C_{ji}+V_{ji}C_{ij}+V_{jj}C_{jj}\\ &+\big(V_{ii}V_{jj}-V_{ij}V_{ji}\big)\big(C_{ii}C_{jj}-C_{ij}C_{ji}\big)\Big],\\ &= \det A \times \frac{\det([C^{ij}])}{\det([\hat{C}^{ij}])}\end{aligned}$$

which is obtained directly because $A^{-1}V$ has non zero entries only on column i and j. Multiplying V by some parameter $\alpha \geq 0$, define

$$P(\alpha) \stackrel{\text{def}}{=} \det(1+\alpha A^{-1}V) = \alpha^2 \det\big([C^{ij}][\hat{C}^{ij}]^{-1} - \frac{\alpha-1}{\alpha}\big).$$

When increasing α from 0 to 1, $P(\alpha)$ will vary from 1 to $\det([C^{ij}])/\det([\hat{C}^{ij}])$ without canceling at some point iff $[C^{ij}][\hat{C}^{ij}]^{-1}$ is definite positive. $P(\alpha)$ is proportional to the characteristic polynomial of a 2×2 matrix $[C^{ij}]/[\hat{C}^{ij}]^{-1}$ of argument $(\alpha-1)/\alpha$, so A' remains positive definite if $[C^{ij}][\hat{C}^{ij}]^{-1}$ is definite positive. Since the product of eigenvalues given by $\det([C^{ij}])/\det([\hat{C}^{ij}])$ is positive, one has to check for the sum, given by the trace of $[C^{ij}]/[\hat{C}^{ij}]^{-1}$:

$$C_{ii}\hat{C}_{jj} + C_{jj}\hat{C}_{ii} - 2C_{ij}\hat{C}_{ij} > 0. \tag{3.8}$$

Since $[C^{ij}]$ and $[\hat{C}^{ij}]$ are individually positive definite, we have

$$C_{ii}C_{jj} - {C_{ij}}^2 > 0 \qquad \text{and} \qquad \hat{C}_{ii}\hat{C}_{jj} - \hat{C}_{ij}^2 > 0$$

from which we deduce that

$$\Big(\frac{C_{ii}\hat{C}_{jj}+C_{jj}\hat{C}_{ii}}{2}\Big)^2 > C_{ii}\hat{C}_{jj}C_{jj}\hat{C}_{ii} > {C_{ij}}^2\hat{C}_{ij}^2,$$

giving finally that (3.8) is always fulfilled when both $[C^{ij}]$ and $[\hat{C}^{ij}]$ are non-degenerate.

Each time a link is added to the graph, its number of loops increases by one unit, so in a sense (3.3) represent a 1-loop correction to the bare Bethe tree solution.

Removing one link To use this in an algorithm, it would also be desirable to be able to remove links as well, such that with help of a penalty coefficient per link, the model could be optimized with a desired connectivity level.

For the Gaussian model, if A is the coupling matrix, removing the link (i,j) amounts to chose a factor ψ_{ij} in (3.1) of the form:

$$\psi_{ij}(x_i,x_j) = \exp\big(A_{ij}x_ix_j\big)$$

(x_i and x_j are assumed centered as in the preceding section). Again, let V denote the perturbation in the precision matrix such that $A' = A+V$ is the new one. The corresponding change in the log likelihood then reads

$$\Delta\mathcal{L} = \log\det(1+A^{-1}V) - \mathrm{Tr}(V\hat{C}).$$

Arranging this expression leads to

$$\Delta\mathcal{L} = \log\big(1 - 2A_{ij}C_{ij} - A_{ij}^2 \det([C^{ij}])\big) + 2A_{ij}\hat{C}_{ij}.$$

Using again (3.6) we get for the new covariance matrix

$$C' = C - \frac{A_{ij}}{1 - 2A_{ij}C_{ij} - A_{ij}^2 \det([C^{ij}])}\; C \begin{bmatrix} A_{ij}C_{jj} & 1 - A_{ij}C_{ij} \\ 1 - A_{ij}C_{ij} & A_{ij}C_{ii} \end{bmatrix} C, \tag{3.9}$$

with again a slight abuse of notation, the 2×2 matrix being to be understood as a $N \times N$ matrix with non-zero entries corresponding to (i,i), (i,j), (j,i) and (j,j). To check for the positive-definiteness property of A', let us observe first that

$$\det(A') = \det(A) \times P(A_{ij}),$$

with

$$P(x) = (1 - x(C_{ij} - \sqrt{C_{ii}C_{jj}}))(1 - x(C_{ij} + \sqrt{C_{ii}C_{jj}})).$$

When x varies from 0 to A_{ij}, $P(x)$ should remains strictly positive to insure that A' is definite positive. This results in the following condition:

$$\frac{1}{C_{ij} - \sqrt{C_{ii}C_{jj}}} < A_{ij} < \frac{1}{\sqrt{C_{ii}C_{jj}} + C_{ij}}.$$

We are now equipped to define algorithms based on addition/deletion of links.

3.3 Imposing walk-summability

Having a sparse GMRF gives no guaranty to its compatibility with belief propagation. In order to be able to use the Gaussian Belief Propagation (GaBP) algorithm for performing inference tasks, stricter constraints have to be imposed. The more precise condition known for convergence and validity of the GaBP algorithm is called walk-summability (WS) and is extensively described in [26]. The two necessary and sufficient conditions for WS that we consider here are:

(i) $\mathrm{Diag}(A) - |R(A)|$ is definite positive, with $R(A) \stackrel{\text{def}}{=} A - \mathrm{Diag}(A)$ the off-diagonal terms of A.

(ii) The spectral radius $\rho(|R'(A)|) < 1$, with $R'(A)_{ij} \stackrel{\text{def}}{=} \frac{R(A)_{ij}}{\sqrt{A_{ii}A_{jj}}}$.

Adding one link Using the criterion developed in the previous section suppose that in order to increase the likelihood we wish to add a link (i,j) to the graph. The model is modified according to

$$A' = A + [\hat{C}^{ij}]^{-1} - [C^{ij}]^{-1} \stackrel{\text{def}}{=} A + V.$$

We assume that A is WS and we want to express conditions under which A' is still WS.

Using the definition (i) of walk summability and mimicking the reasoning leading to (3.8) we can derive a sufficient condition for WS by replacing A with $W(A) \stackrel{\text{def}}{=} \mathrm{Diag}(A) - |R(A)|$. It yields

$$\begin{aligned} \det\left(W(A + \alpha V)\right) &= \det\left(W(A)\right) \det\left(1 + \alpha W(A)^{-1} W(V)\right) \\ &= \det\left(W\right) \Big[1 + \alpha \left\{W_{ii}^{-1}V_{ii} + W_{jj}^{-1}V_{jj} - 2W_{ij}^{-1}|V_{ij}|\right\} \\ &+ \alpha^2 \left(W_{ii}^{-1}W_{jj}^{-1} - (W_{ji}^{-1})^2\right)\left(V_{ii}V_{jj} - |V_{ij}|^2\right)\Big] \\ &= \det(W) Q(\alpha), \end{aligned}$$

since $R(A)_{ij} = 0$ and shortening $W(A)$ in W. A sufficient condition for WS of A' is that Q does not have any root in $[0, 1]$. Note that checking this sufficient condition imposes to keep track of the matrix $(\mathrm{Diag}(A) - |R|)^{-1} = W^{-1} \stackrel{\text{def}}{=} IW$ and requires $\mathcal{O}(N^2)$ operations at each step, using (3.6). A more compact expression of Q is

$$Q(\alpha) = 1 + \alpha \,\mathrm{Tr}\left(IW^{ij}W(V)\right) + \alpha^2 \det\left(W(V)\right) \det\left(IW^{ij}\right),$$

First let's tackle the special case where $\det\left(W(V)IW^{ij}\right) = 0$, the condition for WS of A' is then

$$\mathrm{Tr}(IW^{ij}W(V)) > -1.$$

Of course if the roots are not real, i.e. $\mathrm{Tr}(IW^{ij}W(V))^2 < 4 \det\left(W(V)IW^{ij}\right)$, A' is WS. If none of these conditions is verified we have to check that both roots

$$\frac{-\mathrm{Tr}(IW^{ij}W(V) \pm \sqrt{\mathrm{Tr}(IW^{ij}W(V))^2 - 4\det\left(W(V)IW^{ij}\right)}}{2\det\left(W(V)IW^{ij}\right)},$$

are not in $[0, 1]$.

Modifying one link This is equivalent to adding one link in the sense that (3.1) and (3.2) are still valid. If we want to make use of (i) the only difference is that $R(A)_{ij}$ is not zero before adding V so $W(A + \alpha V) = W(A) + \alpha W(V)$ does not hold in general. Instead we have

$$W(A + \alpha V) = W(A) + \phi(\alpha V, A),$$

with

$$\phi(V, A) \stackrel{\text{def}}{=} \begin{bmatrix} V_{ii} & -|V_{ij} + A_{ij}| + |A_{ij}| \\ -|V_{ji} - A_{ji}| + |A_{ij}| & V_{jj} \end{bmatrix}$$

So we can derive a condition for A' to be WS using, as for the link addition,

$$\begin{aligned} \det(W(A + \alpha V)) &= \det(W) \det\left(1 + W^{-1}\phi(\alpha V, A)\right) \\ &= \det(W)\,\Theta(\alpha) \end{aligned}$$

But now $\Theta(\alpha)$, the equivalent of $Q(\alpha)$, is a degree 2 polynomial only by parts. More precisely, if $\alpha V_{ij} - A_{ij} > 0$ we have $\Theta(\alpha) \stackrel{\text{def}}{=} Q_p(\alpha)$ and else $\Theta(\alpha) \stackrel{\text{def}}{=} Q_m(\alpha)$ with both Q_p and Q_m degree 2 polynomial. So by checking the sign of $\alpha V_{ij} - A_{ij}$ and the roots of Q_p and Q_m we have sufficient conditions for WS after modifying one link.

Another possible way for both adding or modifying one link is to estimate the spectral radius of $|R'(A')|$ through simple power iterations and concludes using (ii). Indeed if a matrix M as a unique eigenvalue of modulus equal to its spectral radius then the sequence

$$\mu_k \stackrel{\text{def}}{=} \frac{\langle b_k, Mb_k\rangle}{\langle b_k, b_k\rangle}, \text{ with } b_{k+1} \stackrel{\text{def}}{=} \frac{Mb_k}{\|\,Mb_k\|}$$

converges to this eigenvalue. While the model remains sparse, with connectivity K, a power iteration of $R'(A')$ requires $\mathcal{O}(KN)$ operations, and it is then possible to conclude about the WS of A' in $\mathcal{O}(KN)$. Keeping track of $W(A)^{-1}$, which requires $\mathcal{O}(N^2)$ operations, at each step is not needed anymore but we have to test WS for each possible candidate link. Note that computing the spectral radius gives us a more precise information about the WS status of the model.

Removing one link Removing one link of the graph will change the matrix A in A' such as $|R'(A')| \leq |R'(A)|$ where the comparison ($\leq$) between two matrices must be understood as the element-wise comparison. Then dealing with positive matrices elementary results gives us $\rho(|R'(A')|) \leq \rho(|R'(A)|)$ and thus removing one link of a WS model provide a new WS model.

3.4 Greedy Graph construction Algorithms

Algorithm 1: Incremental graph construction by link addition

S1 INPUT: the MST graph, and corresponding covariance matrix C.

S2 : select the link with highest $\Delta\mathcal{L}$ compatible with the WS preserving condition of A' in the Gaussian case. Update C according to (3.7) for the Gaussian model. For the Ising model, C is updated by first running BP to generate the set of beliefs and co-beliefs supported by the current factor graph, which in turn allows one to use (2.24) to get all the missing entries in C by inverting χ^{-1}.

S3 : repeat S2 until convergence (i) or until a target connectivity is reached (ii)

S4 : if (ii) repeat S2 until convergence by restricting the link selection in the set of existing ones.

The complexity is respectively $\mathcal{O}(N^3)$ both for the Gaussian and the Ising model in the sparse domain where $\mathcal{O}(N)$ links are added and respectively $\mathcal{O}(N^4)$ and $\mathcal{O}(N^5)$ in the dense domain. Indeed, the inversion of χ^{-1} in the Ising case costs $\mathcal{O}(N^2)$ operations as long as the factor graph remains sparse, but $\mathcal{O}(N^3)$ for dense matrices.

Algorithm 2: Graph surgery by link addition/deletion

S1 INPUT: the MST graph, and corresponding covariance matrix C, a link penalty coefficient ν.

S2 : select the modification with highest $\Delta\mathcal{L} - s\nu$, with $s = +1$ for an addition and $s = -1$ for a deletion, compatible with the WS preserving condition of A' in the Gaussian case. Update C according to (3.7) and (3.9) respectively for an addition or a change and a deletion for the Gaussian model. For the Ising model, C is updated by first running BP to generate the set of beliefs and co-beliefs supported by the current factor graph, which in turn allows one to use (2.24) to get all the missing entries in C by inverting χ^{-1}.

S3 : repeat S2 until convergence.

In absence of penalty ($\nu = 0$) the algorithm will simply generate a model for all different mean connectivities, hence delivering an almost continuous Pareto set of solutions, with all possible trade-off between sparsity and likelihood as long as walk summability is satisfied.

Instead, with a fixed penalty the algorithm is converging toward a solution with a connectivity depending implicitly on ν; it corresponding roughly to the point $K^\star$ where the slope is

$$\frac{\Delta\mathcal{L}}{N\Delta K}(K^\star) = \nu.$$

If we want to use the backtracking mechanism allowed by the penalty term without converging to a specific connectivity, we may also let ν be adapted dynamically. A simple way is to adapt ν with the rate of information gain by letting

$$\nu = \eta\Delta\mathcal{L}_{add}, \qquad \text{with} \qquad \eta \in [0, 1[,$$

where $\Delta\mathcal{L}_{add}$ corresponds to the gain of the last link addition. With such a setting ν is always maintained just below the information gain per link, allowing thus the algorithm to carry on toward higher connectivity. Of course this heuristic assumes a concave Pareto front.

4 Perturbation theory near the Bethe point

4.1 Linear response of the Bethe reference point

The approximate Boltzmann machines described in the introduction are obtained either by perturbation around the trivial point corresponding to a model of independent variables, the first order yielding the Mean-field solution and the second order the TAP one, either by using the linear response delivered in the Bethe approximation. We propose to combine in a way the two procedures, by computing the perturbation around the Bethe model associated to the MST with weights given by mutual information. We denote by $\mathcal{T} \subset \mathcal{E}$, the subset of links corresponding to the MST, considered as given as well as the susceptibility matrix $[\chi_{Bethe}]$ given explicitly by its inverse through (2.24), in term of the empirically observed ones $\hat{\chi}$. Following the same lines as the one given in Section 2, we consider again the Gibbs free energy to impose the individual expectations $\mathbf{m} = \{\hat{m}_i\}$ given for each variable. Let $\mathbf{J}^{Bethe} = \{K_{ij}, (i,j) \in \mathcal{T}\}$ the set of Bethe-Ising couplings, i.e. the set of coupling attached to the MST s.t. corresponding susceptibilities are fulfilled and $\mathbf{J} = \{J_{ij}, (i,j) \in \mathcal{E}\}$ a set of Ising coupling corrections. The Gibbs free energy reads now

$$G[\mathbf{m}, \mathbf{J}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} + F\big[\mathbf{h}(\mathbf{m}), \mathbf{J}^{Bethe} + \mathbf{J}\big]$$

where $\mathbf{h}(\mathbf{m})$ depends implicitly on $\mathbf{m}$ through the same set of constraints (2.7) as before. The only difference resides in the choice of the reference point. We start from the Bethe solution given by the set of coupling $\mathbf{J}^{Bethe}$ instead of starting with an independent model.

The Plefka expansion is used again to expand the Gibbs free energy in power of the coupling J_{ij} assumed to be small. Following the same lines as in Section 2.1, but with G_0 now replaced by

$$G_{Bethe}[\mathbf{m}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} - \log Z_{Bethe}\big[\mathbf{h}(\mathbf{m}), \mathbf{J}^{Bethe}\big],$$

and h_i, J^{Bethe} and Z_{Bethe} given respectively by (2.21,2.22,2.23) where $\mathcal{E}$ is now replaced by $\mathcal{T}$, letting again

$$H^1 \stackrel{\text{def}}{=} \sum_{i<j} J_{ij} s_i s_j,$$

and following the same steps (2.13,2.14,2.15) leads to the following modification of the local fields

$$h_i = h_i^{Bethe} - \sum_j [\chi^{-1}_{Bethe}]_{ij} \operatorname{Cov}_{Bethe}(H^1, s_i) \qquad \forall i \in \mathcal{V}$$

to get the following Gibbs free energy at second order in α (after replacing H^1 by αH^1):

$$G[\mathbf{m}, \alpha J] = G_{Bethe}(\mathbf{m}) - \alpha \mathbb{E}_{Bethe}(H^1)$$
$$-\frac{\alpha^2}{2}\Big(\operatorname{Var}_{Bethe}(H^1) - \sum_{ij} [\chi^{-1}_{Bethe}]_{ij} \operatorname{Cov}_{Bethe}(H^1, s_i) \operatorname{Cov}_{Bethe}(H^1, s_j)\Big) + o(\alpha^2).$$

This is the general expression for the linear response near the Bethe reference point that we now use.

$$G_{BLR}[\mathbf{J}] \stackrel{\text{def}}{=} -\mathbb{E}_{Bethe}(H^1) \tag{4.1}$$

$$-\frac{1}{2}\Big(\mathrm{Var}_{Bethe}(H^1) - \sum_{i,j}[\chi^{-1}_{Bethe}]_{ij}\,\mathrm{Cov}_{Bethe}(H^1, s_i)\,\mathrm{Cov}_{Bethe}(H^1, s_j)\Big). \tag{4.2}$$

represents the Gibbs free energy at this order of approximation. It it is given explicitly through

$$\mathbb{E}_{Bethe}(H^1) = \sum_{i<j} J_{ij} m_{ij}$$

$$\mathrm{Var}_{Bethe}(H^1) = \sum_{i<j,k<l} J_{ij} J_{kl}\big(m_{ijkl} - m_{ij}m_{kl}\big)$$

$$\mathrm{Cov}_{Bethe}(H^1, s_k) = \sum_{i<j} J_{ij}(m_{ijk} - m_{ij}m_k)$$

where

$$m_i \stackrel{\text{def}}{=} \mathbb{E}_{Bethe}(s_i), \qquad m_{ij} \stackrel{\text{def}}{=} \mathbb{E}_{Bethe}(s_i s_j)$$

$$m_{ijk} \stackrel{\text{def}}{=} \mathbb{E}_{Bethe}(s_i s_j s_k), \qquad m_{ijkl} \stackrel{\text{def}}{=} \mathbb{E}_{Bethe}(s_i s_j s_k s_l)$$

are the moments delivered by the Bethe approximation. With the material given in Section 2.2 these are given in closed form in terms of the Bethe susceptibility coefficients χ_{Bethe}. Concerning the log-likelihood, it is given now by:

$$\mathcal{L}[\mathbf{J}] = -G_{Bethe}(\mathbf{m}) - G_{BLR}[\mathbf{J}] - \sum_{ij}(J^{Bethe}_{ij} + J_{ij})\hat{m}_{ij} + o(J^2). \tag{4.3}$$

G_{BLR} is at most quadratic in the J's and contains the local projected Hessian of the log likelihood onto the magnetization constraints (2.7) with respect to this set of parameters. This is nothing else than the Fisher information matrix associated to these parameter J which is known to be positive-semidefinite, which means that the log-likelihood associated to this parameter space is convex. Therefore it makes sense to use the quadratic approximation (4.3) to find the optimal point.

4.2 Line search along the natural gradient in a reduced space

Finding the corresponding couplings still amounts to solve a linear problem of size N^2 in the number of variables which will hardly scale up for large system sizes. We have to resort to some simplifications which amounts to reduce the size of the problem, i.e. the number of independent couplings. To reduce the problem size we can take a reduced number of link into consideration, i.e. the one associated with a large mutual information or to partition them in a way which remains to decide, into a small number q of group $\mathcal{G}_\nu, \nu = 1, \dots q$. Then, to each group ν is associated a parameter α_ν with a global perturbation of the form

$$H^1 = \sum_{\nu=1}^{q} \alpha_\nu H_\nu$$

where each H_ν involves the link only present in $\mathcal{G}_\nu$.

$$H_\nu \stackrel{\text{def}}{=} \sum_{(i,j)\in\mathcal{G}_\nu} J_{ij} s_i s_j,$$

and the J_{ij} are fixed in some way to be discussed soon. The corresponding constraints, which ultimately insures a max log-likelihood in this reduced parameter space are then

$$\frac{\partial G_{BLR}}{\partial \alpha_\nu} = -\hat{\mathbb{E}}(H_\nu).$$

This leads to the solution:

$$\alpha_\mu = \sum_{\nu=1}^{q} \mathcal{I}^{-1}_{\mu\nu}\big(\hat{\mathbb{E}}(H_\nu) - \mathbb{E}_{Bethe}(H_\nu)\big)$$

where the Fisher information matrix $\mathcal{I}$ has been introduced and which reads in the present case

$$\mathcal{I}_{\mu\nu} = \big[\mathrm{Cov}_{Bethe}(H_\mu, H_\nu) - \sum_{\substack{i\neq j\\(i,j)\in\mathcal{T}}} [\chi^{-1}_{Bethe}]_{ij}\,\mathrm{Cov}_{Bethe}(H_\mu, s_i)\,\mathrm{Cov}_{Bethe}(H_\nu, s_j)\big] \tag{4.4}$$

The interpretation of this solution is to look in the direction of the natural gradient [1, 3] of the log likelihood. The exact computation of the entries of the Fisher matrix involves up to 4th order moments and can be computed using results of Section 2.2. At this point, the way of choosing the groups of edges and the perturbation couplings J_{ij} of the corresponding links, leads to various possible algorithms. For example, to connect this approach to the one proposed in Section 3.2, the first group of links can be given by the MST, with parameter α_0 and their actual couplings $J_{ij} = J^{Bethe}_{ij}$ at the Bethe approximation; making a short list of the $q-1$ best links candidates to be added to the graph, according to the information criteria 3.3, defines the other groups as singletons. It is then reasonable to attach them the value

$$J_{ij} = \frac{1}{4}\log\frac{\hat{p}^{11}_{ij}\,\hat{p}^{00}_{ij}\,p^{01}_{ij}\,p^{10}_{ij}}{p^{11}_{ij}\,p^{00}_{ij}\,\hat{p}^{01}_{ij}\,\hat{p}^{10}_{ij}},$$

of the coupling according to (3.2), while the modification of the local fields as a consequence of (3.2) can be dropped since the Gibbs free energy take it already into account implicitly, in order to maintain single variable magnetization $m_i = \hat{m}_i$ correctly imposed.

4.3 Reference point at low temperature

Up to now we have considered the case where the reference model is supposed to be a tree and is represented by a single BP fixed point. From the point of view of the Ising model this corresponds to perturb a high temperature model in the paramagnetic phase. In practice the data encountered in applications are more likely to be generated by a multi-modal distribution and a low temperature model with many fixed points should be more relevant. In such a case we assume that most of the correlations are already captured by the definition of single beliefs fixed points and the residual correlations is contained in the co-beliefs of each fixed point. For a multi-modal distribution with q modes with weight $w_k, k = 1\dots q$ and a pair of variables (s_i, s_j) we indeed have

$$\begin{aligned}\chi_{ij} &= \sum_{k=1}^{q} w_k\,\mathrm{Cov}(s_i, s_j|k) + \sum_{k=1}^{q} w_k(\mathbb{E}(s_i|k) - \mathbb{E}(s_i))(\mathbb{E}(s_j|k) - \mathbb{E}(s_j))\\ &\stackrel{\text{def}}{=} \chi^{intra}_{ij} + \chi^{inter}_{ij},\end{aligned}$$

where the first term is the average intra cluster susceptibility while the second is the inter cluster susceptibility. All the preceding approach can then be followed by replacing the single Bethe susceptibility and higher order moments in equations (4.1,4.4) in the proper way by their multiple BP fixed point counterparts. For the susceptibility coefficients, the inter cluster susceptibility coefficients χ^{inter} are given directly from the single variable belief fixed points. The intra cluster susceptibilities χ^k are treated the same way as the former Bethe susceptibility. This means that the co-beliefs of fixed points $k \in \{1, \dots q\}$ are entered in formula (2.24) which by inversion yields the χ^k's, these in turn leading to χ^{intra} by superposition. Higher order moments are obtain by simple superposition. Improved models could be then searched along the direction indicated by this natural gradient.

5 L_0 norm penalized sparse inverse estimation algorithm

We propose here to use the Doubly Augmented Lagrange (DAL) method [22, 11, 10] to solve the penalized log-determinant programming in (2.28). For a general problem defined as follows:

$$\min_x F(x) = f(x) + g(x) \tag{5.1}$$

where $f(x)$ and $g(x)$ are both convex. DAL splits the combination of $f(x)$ and $g(x)$ by introducing a new auxiliary variable y. Thus, the original convex programming problem can be formulated as :

$$\begin{aligned} \min_{x,y} F(x) &= f(x) + g(y) \\ \text{s.t.} \quad x - y &= 0 \end{aligned} \tag{5.2}$$

Then it advocates an augmented Lagrangian method to the extended cost function in (5.2). Given penalty parameters μ and γ, it minimizes the augmented Lagrangian function

$$L(x, y, \nu, \tilde{x}, \tilde{y}) = f(x) + g(y) + \langle \nu, x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2 + \frac{\gamma}{2}\|x - \tilde{x}\|_2^2 + \frac{\gamma}{2}\|y - \tilde{y}\|_2^2 \tag{5.3}$$

where $\tilde{x}$ and $\tilde{y}$ are the prior guesses of x and y that can obtained either from a proper initialization or the estimated result in the last round of iteration in an iterative update procedure. Since optimizing jointly with respect to x and y is usually difficult, DAL optimizes x and y alternatively. That gives the following iterative alternative update algorithm with some simple manipulations:

$$\begin{aligned} x^{k+1} &= \min_x f(x) + \frac{\mu}{2}\|x - y^k + \tilde{\nu}^k\|_2^2 + \frac{\gamma}{2}\|x - x^k\|_2^2 \\ y^{k+1} &= \min_y g(y) + \frac{\mu}{2}\|x^{k+1} - y + \tilde{\nu}^k\|_2^2 + \frac{\gamma}{2}\|y - y^k\|_2^2 \\ \tilde{\nu}^{k+1} &= \tilde{\nu}^k + x^{k+1} - y^{k+1} \end{aligned} \tag{5.4}$$

where $\tilde{\nu} = \frac{1}{\mu}\nu$. As denoted in [10] and [22], DAL improves basic augmented Lagrangian optimization by performing additional smooth regularization on estimations of x and y in successive iteration steps. As a result, it guarantees not only the convergence of the scaled dual variable $\tilde{\nu}$, but also that of the proximal variables x^k and y^k, which could be divergent in basic augmented Lagrangian method.

We return now to the penalized log-determinant programming in sparse inverse estimation problem, as seen in (2.28). The challenge of optimizing the cost function is twofold. Firstly, the exact L_0-norm penalty is non-differentiable, making it difficult to find an analytic form of gradient for optimization. Furthermore, due to the log-determinant term in the cost function, it implicitly requires that any feasible solution to the sparse approximation A of the precision

matrix should be strictly positive definite. The gradient of the log-determinant term is given by $\hat{C} - A^{-1}$, which is not continuous in the positive definite domain and makes it impossible to obtain any second-order derivative information to speed up the gradient descent procedure. We hereafter use S_{++} as the symmetric positive definite symmetric matrices that form the feasible solution set for this problem. By applying DAL to the cost function (2.28), we can derive the following formulation:

$$\begin{aligned}\hat{J}(A, Z, \tilde{A}, \tilde{Z}, \nu) = &- \log\det(A) + \mathrm{Tr}(\hat{C}A) + \lambda P(Z) + \langle \nu, A - Z\rangle \\ &+ \frac{\mu}{2}\|A - Z\|_2^2 + \frac{\gamma}{2}\|A - \tilde{A}\|_2^2 + \frac{\gamma}{2}\|Z - \tilde{Z}\|_2^2 \\ \text{s.t. } &A, Z \in S_{++}\end{aligned} \tag{5.5}$$

where Z is the auxiliary variable that has the same dimension as the sparse inverse estimation A. $\tilde{A}$ and $\tilde{Z}$ are the estimated values of A and Z derived in the last iteration step. The penalty parameter γ controls the regularity of A and Z. By optimizing A and Z alternatively, the DAL procedure can be easily formulated as an iterative process as follows, for some $\delta > 0$:

$$\begin{aligned}A^{k+1} &= \underset{A}{\mathrm{argmin}} - \log\det(A) + \mathrm{Tr}(\hat{C}A) + \lambda P(Z^k) + \langle \nu^k, A - Z^k\rangle \\ &\quad + \frac{\mu}{2}\|A - Z^k\|_2^2 + \frac{\gamma}{2}\|A - A^k\|_2^2 \\ Z^{k+1} &= \underset{Z}{\mathrm{argmin}}\, \lambda P(Z) + \langle \nu^k, A^{k+1} - Z\rangle + \frac{\mu}{2}\|A^{k+1} - Z\|_2^2 \\ &\quad + \frac{\gamma}{2}\|Z - Z^k\|_2^2 \\ \nu^{k+1} &= \nu^k + \delta(A^{k+1} - Z^{k+1}) \\ \text{s.t. } &A^{k+1}, Z^{k+1} \in S_{++}\end{aligned} \tag{5.6}$$

By introducing the auxiliary variable Z, the original penalized maximum likelihood problem is decomposed into two parts. The first one is composed mainly by the convex log-determinant programming term. Non-convex penalty is absorbed into the left part. Separating the likelihood function and the penalty leads to the simpler sub-problems of solving log-determinant programming using eigenvalue decomposition and L_0 norm penalized sparse learning alternatively. Each sub-problem contains only one single variable, making it applicable to call gradient descent operation to search local optimum. Taking $\tilde{\nu} = \frac{1}{\mu}\nu$, we can derive the following scaled version of DAL for the penalized log-determinant programming:

$$\begin{aligned}A^{k+1} &= \underset{A}{\mathrm{argmin}} - \log\det(A) + \mathrm{Tr}(\hat{C}A) + \frac{\mu}{2}\left\|A - Z^k + \tilde{\nu}^k\right\|_2^2 + \frac{\gamma}{2}\left\|A - A^k\right\|_2^2 \\ Z^{k+1} &= \underset{Z}{\mathrm{argmin}}\, \frac{\mu}{2}\left\|A^{k+1} - Z + \tilde{\nu}^k\right\|_2^2 + \frac{\gamma}{2}\left\|Z - Z^k\right\|_2^2 + \lambda P(Z) \\ \tilde{\nu}^{k+1} &= \tilde{\nu}^k + A^{k+1} - Z^{k+1} \\ \text{s.t. } &A^{k+1}, Z^{k+1} \in S_{++}\end{aligned} \tag{5.7}$$

To attack the challenge caused by non-differentiability of the exact L_0 norm penalty, we make use of a differentiable approximation to L_0-norm penalty in the cost function $\hat{J}$, named as "seamless L_0 penalty" (SELO) in [25]. The basic definition of this penalty term is given as:

$$P_{\mathrm{SELO}}(Z) = \sum_{i,j} \frac{1}{\log(2)} \log\left(1 + \frac{|Z_{i,j}|}{|Z_{i,j}| + \tau}\right) \tag{5.8}$$

where $Z_{i,j}$ denotes individual entry in the matrix Z and $\tau > 0$ is a tuning parameter. As seen in Figure 2.2, as τ gets smaller, $P(Z_{i,j})$ approximates better the L_0 norm $I(Z_{i,j} \neq 0)$. SELO penalty is differentiable, thus we can calculate the gradient of $P(Z)$ explicitly with respect to each $Z_{i,j}$ and make use of first-order optimality condition to search local optimum solution. Due to its continuous property, it is more stable than the exact L_0 norm penalty in optimization. As proved in [25], the SELO penalty has the oracle property with proper setting of τ. That's to say, the SELO penalty is asymptotically normal with the same asymptotic variance as the unbiased OLS estimator in terms of Least Square Estimation problem.

The first two steps in (5.7) are performed with the positive definite constrains imposed on A and Z. The minimizing with respect to A is accomplished easily by performing Singular Vector Decomposition (SVD). By calculating the gradient of $\hat{J}$ with respect to A in (5.7), based on the first-order optimality, we derive:

$$\hat{C} - A^{-1} + \mu(A - Z^k + \tilde{\nu}^k) + \gamma(A - A^k) = 0 \tag{5.9}$$

Based on generalized eigenvalue decomposition, it is easy to verify that $A^{k+1} = V\,\mathrm{Diag}(\beta)V^T$, where V and $\{d_i\}$ are the eigenvectors and eigenvalues of $\mu(Z^k - \tilde{\nu}^k) - \hat{C} + \gamma A^k$. β_i is defined as:

$$\beta_i = \frac{d_i + \sqrt{{d_i}^2 + 4(\tilde{\nu} + \gamma)}}{2(\tilde{\nu} + \gamma)} \tag{5.10}$$

Imposing $Z \in S_{++}$ directly in minimizing the cost function with respect to Y make the optimization difficult to solve. Thus, instead, we can derive a feasible solution to Z by a continuous search on μ. Based on spectral decomposition, it is clear that X^{k+1} is guaranteed to be positive definite, while it is not necessarily sparse. In contrast, Z is regularized to be sparse while not guaranteed to be positive definite. μ is the regularization parameter controlling the margin between the estimated X^{k+1} and the sparse Z^{k+1}. Increasingly larger μ during iterations makes the sequences $\{X^k\}$ and $\{Z^k\}$ converge to the same point gradually by reducing margin between them. Thus, with enough iteration steps, the derived Z^k follows the positive definite constraint and sparsity constraint at the same time. We choose here to increase μ geometrically with a positive factor $\eta > 1$ after every N_μ iterations until its value achieves a predefined upper bound $\mu_{\max}$. With this idea, the iterative DAL solution to the L_0 norm penalty is given as:

$$\begin{aligned}
A^{k+1} &= \underset{A}{\operatorname{argmin}} - \log\det(A) + \mathrm{Tr}(\hat{C}A) + \frac{\mu}{2}\big\|A - Z^k + \tilde{\nu}^k\big\|_2^2 + \frac{\gamma}{2}\big\|A - A^k\big\|_2^2, \\
Z^{k+1} &= \underset{Z}{\operatorname{argmin}} \frac{\mu}{2}\big\|A^{k+1} - Z + \tilde{\nu}^k\big\|_2^2 + \frac{\gamma}{2}\big\|Z - Z^k\big\|_2^2 + \lambda P(Z) \\
\tilde{\nu}^{k+1} &= \tilde{\nu}^k + A^{k+1} - Z^{k+1} \\
\mu^{k+1} &= \min\big(\mu\,\eta^{\lfloor k/N_\mu \rfloor}, \mu_{\max}\big).
\end{aligned} \tag{5.11}$$

In the second step of (5.11), we calculate the gradient of the cost function with respect to Z and achieve the local minimum by performing the first-order optimum condition on it. Therefore, the updated value of each entry of Z is given by a root of a cubic equation, as defined below:

$$\begin{aligned}
&\text{if } Z_{i,j} > 0,\ Z_{i,j} \text{ is the positive root of} \\
&\quad 2{Z_{i,j}}^3 + (3\tau - 2\theta_{i,j}){Z_{i,j}}^2 + (\tau^2 - 3\tau\theta_{i,j})Z_{i,j} - \tau^2\theta_{i,j} + \frac{\lambda\tau}{\mu + \gamma} = 0 \\
&\text{if } Z_{i,j} < 0,\ Z_{i,j} \text{ is the negative root of} \\
&\quad 2{Z_{i,j}}^3 - (3\tau + 2\theta_{i,j}){Z_{i,j}}^2 + (\tau^2 + 3\tau\theta_{i,j})Z_{i,j} - \tau^2\theta_{i,j} - \frac{\lambda\tau}{\mu + \gamma} = 0 \\
&\text{else } Z_{i,j} = 0
\end{aligned} \tag{5.12}$$

where $Z_{i,j}$ is one single entry of Z and

$$\theta_{i,j} = \frac{\gamma Z^k_{i,j} + \mu(A^{k+1}_{i,j} + \tilde{\nu}^k)}{\mu + \gamma}.$$

Solving the cubic equations can be done rapidly using Cardano's formula within a time cost $O(n^2)$. Besides, the spectral decomposition procedure has the general time cost $O(n^3)$. Given the total number of iterations K, theoretical computation complexity of DAL is $O(Kn^3)$. For our experiments, we initialize μ to 0.06, the multiplier factor η to 1.3 and the regularization penalty parameter γ to 10^{-4}. To approximate the L_0 norm penalty, τ is set to be $5 \cdot 10^{-4}$. In our experiment, to derive the Pareto curve of the optimization result, we traverse different values of λ. Most learning procedures converge with no more than $K = 500$ iteration steps.

To validate performance of sparse inverse estimation based on the L_0 norm penalty, we involve an alternative sparse inverse matrix learning method using L_1 norm penalization for comparison. Taking $P(A)$ in (2.28) to be the L_1 matrix norm of A, we strengthen conditional dependence structure between random variables by jointly minimizing the negative log likelihood function and the L_1 norm penalty of the inverse matrix. Since L_1 norm penalty is strictly convex, we can use a quadratic approximation to the cost function to search for the global optimum, which avoids singular vector decomposition with complexity of $O(p^3)$ and improves the computational efficiency of this solution to $O(p)$, where p is the number of random variables in the GMRF model. This quadratic approximation based sparse inverse matrix learning is given in [5], named as QUIC. We perform it directly on the empirical covariance matrix with different settings of the regularization coefficient λ. According to works in compressed sensing, the equality between L_1 norm penalty and L_0 norm penalty holds if and only if the design matrix satisfies restricted isometry property. However, restricted isometry property is sometimes too strong in practical case. Furthermore, to our best knowledge, there is no similar necessary condition guaranteeing equivalence between L_1 and L_0 norm penalty in sparse inverse estimation problem. Therefore, in our case, L_1 norm penalized log-determinant programming is highly likely to be biased from the underlying sparse correlation structure in the graph, which leads to much denser inverse matrices.

6 Experiments

In this section, various solutions based on the different methods exposed before are compared. We look first at the intrinsic quality, given either by the exact log likelihood for the Gaussian case, or by the empirical one for the Ising model, and then at its compatibility with belief propagation for inference tasks.

Inverse Ising problem Let us start with the inverse Ising problem. The first set of experiments illustrates how the linear-response approach exposed in Section 2 works when the underlying model to be found is itself an Ising model. The quality of the solution can then be assessed directly by comparing the couplings J_{ij} found with the actual ones. Figure 6.1 are obtained by generating at random 10^3 Ising models of small size $N = 10$ either with no local fields ($h_i = 0, \forall i = 1 \dots N$) or with random centered ones $h_i = U[0,1] - 1/2$ and with couplings $J_{ij} = \frac{J}{\sqrt{N/3}}(2 * U[0,1] - 1)$, centered with variance J^2/N, J being the common rescaling factor corresponding to the inverse temperature. A glassy transition is expected at $J = 1$. The couplings are then determined using (2.16), (2.17), (2.22) and (2.25) respectively for the mean-field, TAP, BP and Bethe (equivalent to susceptibility propagation) solutions. Figure 6.1.a shows that

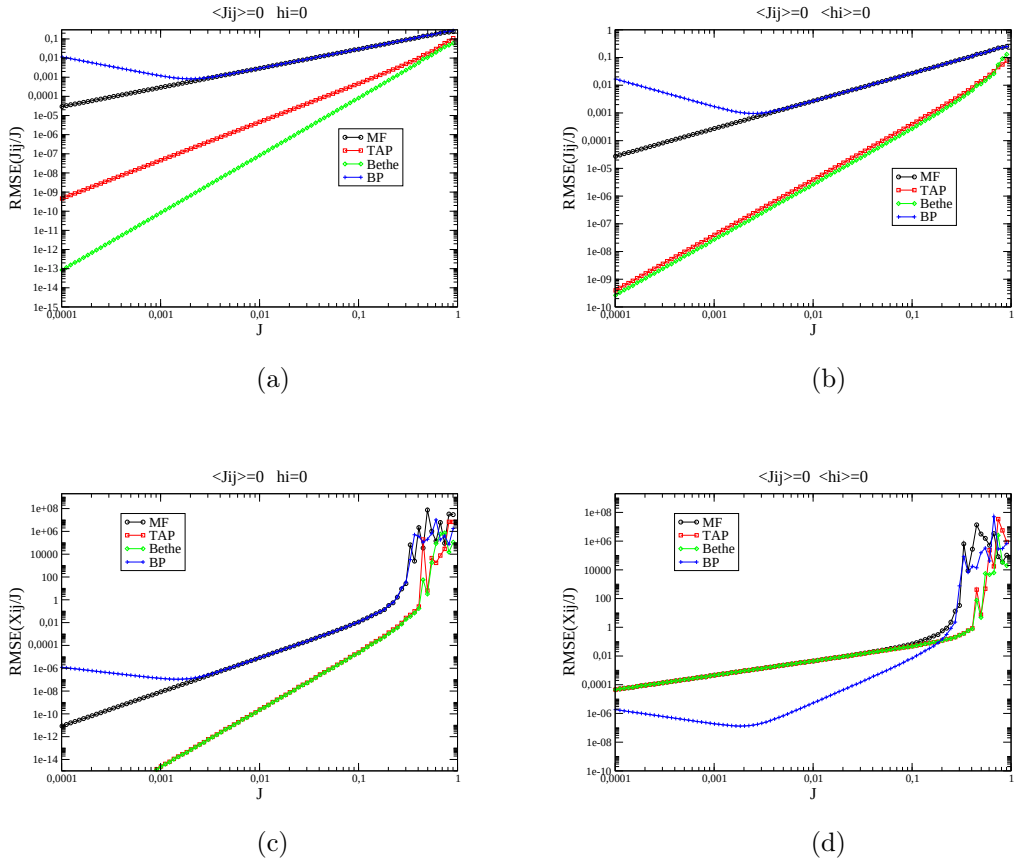
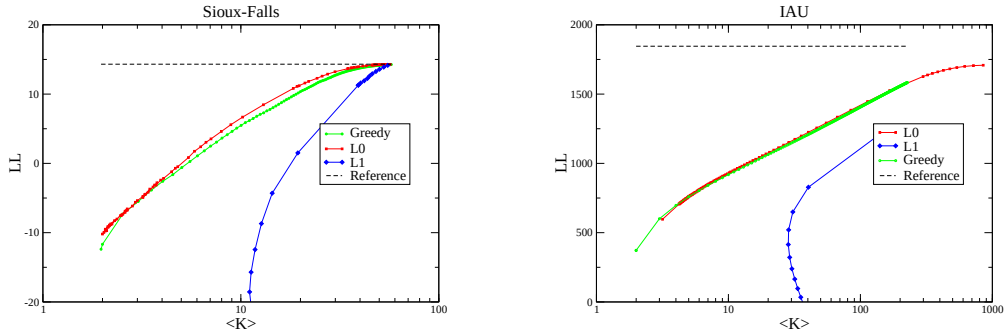


Figure 6.1: Comparison between various approximate solutions to the inverse Ising problem. RMSE errors as a function of the temperature are plotted in (a) and (b) for the couplings J_{ij}, in (c) and (d) for the susceptibility matrix χ_{ij} obtained from the corresponding BP fixed point. Local fields h_i are zero in (a) and (c) and finite but zero in average in (b) and (d).

the Bethe approximation yields the most precise results in absence of local fields while it is equivalent to TAP when a local field is present as shown on Figure 6.1.b. Since we want to use these methods in conjunction with BP we have also compared the BP-susceptibilities they deliver. To do that, we simply run BP to get a set of belief and co-beliefs in conjunction with equation (2.24) which after inversion yields a susceptibility matrix to be compared with the exact ones. The comparison shown on Figure 6.1.c indicates that Bethe and TAP yield the best results in absence of local field which are less robust when compared to the more naive BP method when local fields are present as seen on Figure 6.1.d. This is due to the fact that BP delivers exact beliefs when model (2.22,2.21) is used, which is not necessarily the case for other methods when the local fields are non-vanishing. It is actually not a problem of accuracy but of BP compatibility which is raised by this plot.

Sparse inverse models Let us now test the approach proposed in Section 3.2 to build a model link by link for comparison with more conventional optimization schema based on L_0 and L_1 penalizations. We show the results of tests only for the Gaussian case where the link surgery



(a) (b)

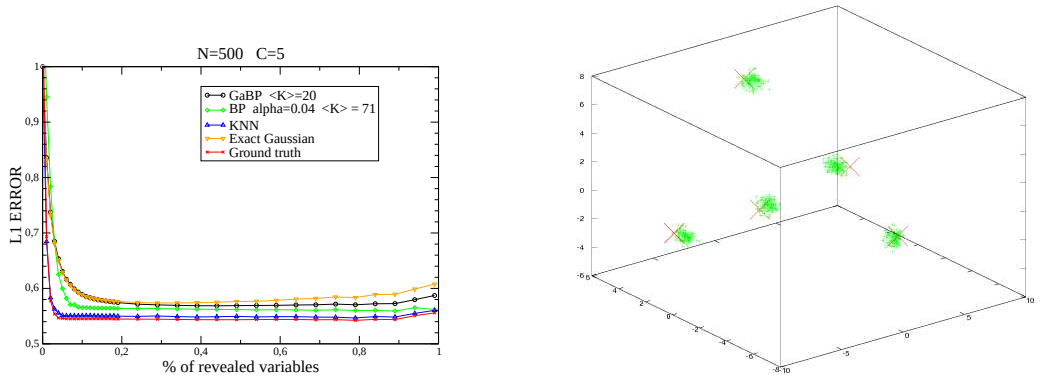
Figure 6.2: Comparison of the greedy information based MRF inference method with L_0 and L_1 norm penalized optimizations. (a) corresponds to the Sioux-Fall data of size $N = 60$. (b) corresponds to the IAU data of size $N = 1000$. $< K >$ is the ratio of links to nodes N. The log likelihood on the y axes is unormalized and corresponds to (2.29). The reference is the log likelihood given by the full inverse covariance matrix.

can be treated exactly. In the Ising case the method can be used only marginally to propose little correction on the maximum spanning tree or any other sparse model. In general we cannot expect the method to be able to compete with the inverse susceptibility propagation schema i.e. what we call here the Bethe inverse model (2.25). The reason is that the LL gain given by one link is more costly to assess than in the Gaussian case and it is also only approximate. So the stability of the schema is more difficult to control when many links have to be added because the condition of the validity of the Bethe approximation are not controlled without paying an additional computational price. For the Gaussian case instead the situation is much more favorable because the gain can be computed exactly with low computational cost even when the graph is dense. The test we show on Figure 6.2 are done on simulated data produced by the traffic simulator METROPOLIS [9], our original motivation for this work being related to traffic inference [16, 14]. The first set corresponds to a small traffic network called Sioux-Fall consisting of 72 links, from which we extract the $N = 60$ most varying ones (the other one being mostly idle). The second set (IAU) is obtained for a large scale albeit simplified network of the Paris agglomeration of size 13626 links, out of which we extracted a selection of the $N = 1000$ most varying ones. Each sample data is a N-dimensional vector of observed travel times $\{\hat{t}_i, i = 1 \dots N\}$, giving a snapshot of the network at a given time in the day. The total number of samples is $S = 3600$ for Sioux-Falls and $S = 7152$ for IAU, obtained by generating many days of various traffic scenarios. Then for each link the travel time distribution is far from being Gaussian, having heavy tails in particular. So to deal with normal variables (if taken individually) we make the following standard transformation:

$$y_i = F^{-1}_{Gauss} \hat{F}_i(t_i), \qquad \forall i = 1 \dots N \tag{6.1}$$

which map the travel time t_i to a genuine Gaussian variable y_i, where $\hat{F}_i$ and F_{Gauss} are respectively the empirical cdf of t_i and of a centered normal variable. The input of the different algorithms under study is then the covariance matrix $\mathrm{Cov}(y_i, y_j)$. This mapping will actually be important in the next section when using the devised MRF for inference tasks. Figure 6.2 displays the comparison between various methods. Performances of the greedy method are com-

parable to the L_0 penalized optimization. To generate one solution both methods are comparable also in term of computational cost, but the greedy is faster in very sparse regime, and since it is incremental, it generate a full Pareto subset for the cost of one solution. On this figure we see also that the L_1 method is simply not adapted to this problem. From the figures, we can see that the estimated inverse matrix derived based on L_1 norm penalty needs distinctively more non-zero entries to achieve similar log-likelihood level as the L_0 penalty, indicating its failure of discovering the underlying sparse structure, the thresholding of small non-zero entries being harmful w.r.t. positive definitness. The reason might be that is adapted to situations where a genuine sparse structure exists, which is not the case in the present data.



(a) (b)

Figure 6.3: Comparison of decimation curves in the case of a multi-modal distribution with five cluster for $N = 500$ variables (a). Projection of the dataset in the 3-d dominant PCA space along with the corresponding BP fixed points projections obtained with the Ising model.

Inverse models for inference We turn now to experiments related to the original motivation of this work, which is to use calibrated model for some inference tasks. The experiments goes as follows: we have an historical data set consisting of a certain number of samples, each one being a N-dimensional variable vector, say travel times, which serves to build the models [2]. Given a sample test data we want to infer the $(1-\rho)N$ hidden variables when a certain fraction ρ of the variables are revealed. In practice we proceed gradually on each test sample by revealing one by one the variables in a random order and plot as a function of ρ the L_1 error made by the inference model on the hidden variables. Both for Ising and Gaussian MRF, the inference is not performed in the original variable space, but in an associated one obtained through a mapping (a traffic index) using the empirical cumulative distribution of each variable. For the Gaussian model the inference is performed in the index space defined previously by (6.1). For the Ising models we have studied a variety of possible mapping [27] in order to associate a binary variable to a real one such that a belief associated to a binary state can be converted back into a travel time prediction. Without entering into the details (see [27] for details), to define in practice this binary state σ_i, either we make use of the median value $x_i^{1/2} = F_i^{-1}(1/2)$ in the distribution of x_i for all $i = 1 \dots N$:

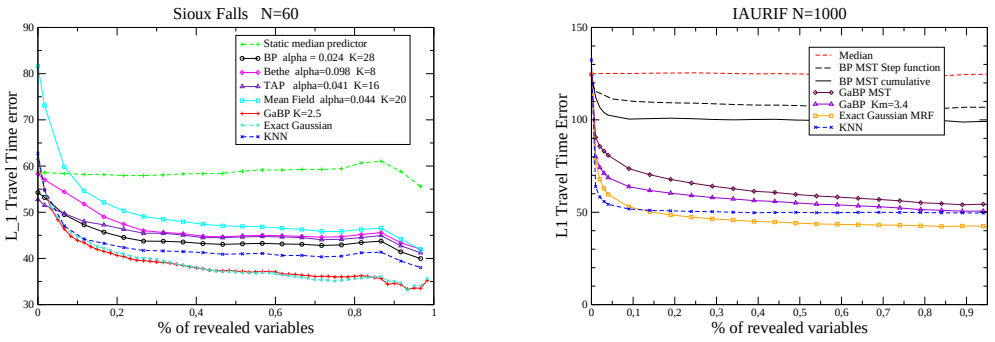
$$\sigma_i = \mathbb{1}_{\{x_i > x_i^{1/2}\}} \qquad (i).$$

[2] In fact the pairwise MRF models exploit only pairwise observations but for sake of comparison with a KNN predictor we generate complete historical samples data.

Either we perform a soft mapping using the cdf:

$$P(\sigma_i = 1) = \hat{F}_i(x_i) \qquad (ii),$$

the last one having the advantage of being functionally invertible if $\hat{F}_i^{-1}$ is defined, while the former one being inverted using Bayes rule. The data we are considering are "low temperature" data in the sense that correlations are too strong for an Ising model with one single fixed point. This is reflected in the fact that none of the basic methods given in the Section 2 is working. To overcome this we use a simple heuristic which consists in to add a parameter $\alpha \in [0, 1]$ in the BP model like e.g. in (2.26) or to multiply the J_{ij} by α for the MF, TAP and Bethe models, the local field being consequently modified owing to their dependency on the J_{ij}. Concerning the factor-graph we have considered various graph selection procedures. All are based on the mutual information given empirically between variables. A global/local threshold can be used to construct the graph, the parameter being the mean/local connectivity K; the MST can be used conveniently as a backbone and additional links are obtained through the thresholding selection procedures. These two parameter α and K are calibrated such as to optimize the performance for each type of model so that fair comparisons can be made afterward. One important difference



(a) (b)

Figure 6.4: Comparison of decimation curves between various MRF for Sioux-Falls data (a) and IAU data (b)

between the Ising model and the Gaussian one is that multiple fixed points may show up in the Ising case while only a single one, stable or not stable, is present in the Gaussian case. This can be an advantage in favor of the Ising model when the data have well separated clusters. Figure 6.3 illustrates this point. The data are sampled from a distribution containing 5 modes, each one being a product form over N random bimodal distributions attached to each link. On Figure 6.3.a which displays the error as a function of the fraction of revealed variables we see that the Ising model obtained with (2.26), encoded with the median value (i), gives better prediction than the exact Gaussian model or the approximated GaBP compatible one. Indeed Figure 6.3.b shows a projection of the data in the most relevant 3-d PCA space along with the projected position of BP fixed points (given by their sets of beliefs) delivered by the Ising model. As we see, the model is able to attach one BP fixed point to each component of the distribution. Ideally we would like a perfect calibration of the Ising model in order that these fixed points be located at the center of each cluster. The method proposed in Section 4 could help to do this, but has not been implemented yet. On Figure 6.3.a we see also that the KNN predictor performs optimally in this case, since the error curve coincides exactly with the one given by the hidden generative

model of the data (ground truth). Figure 6.4 shows how the different models compare on the data generated by the traffic simulator. On the Sioux-Falls network, the Gaussian model gives the best results, and a sparse version obtained with the greedy algorithm of section 3.4 reach the same level of performance and outperforms KNN. The best Ising model is obtained with the (2.26) with type (ii) encoding. For IAU the full Gaussian model is also competitive w.r.t KNN, but the best sparse GaBP model is not quite able to follow. In fact the correlations are quite high in this data, which explain why the best Ising model shows very poor performance. The best Ising model in that case corresponds to the plain BP model with type (ii) encoding and MST graph.

7 Conclusion

This paper is based on the observation that the Bethe approximation can be in many case a good starting point for building inverse models from data observations. We have developed different ways of perturbing such a mean-field solution valid both for binary and Gaussian variables, and leading to an efficient algorithm in the Gaussian case to generated sparse approximation models compatible with BP. The additional requirement that the model be compatible with BP for large scale application discards dense models and simplifies in a way the search space on model selection. More experimental tests on various data should help to refine and settle the general methods proposed here.

Acknowledgments This work was supported by the French National Research Agency (ANR) grant N° ANR-08-SYSC-017.

References

[1] Amari, S. Natural gradient works efficiently in learning. *Neural Computation 10*, 2 (1998), 251–276.

[2] Anandkumar, A., Tan, V., Huang, F., and Willsky, A. High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *JMLR* (2012), 2293–2337.

[3] Arnold, L., Auger, A., Hansen, N., and Ollivier, Y. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *ArXiv e-prints* (2011).

[4] Chang, F. C. Inversion of a perturbed matrix. *Applied Mathematics Letters 19*, 2 (2006), 169 – 173.

[5] Cho-Jui Hsieh, Matyas A.Sustik, I. S., and Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advancees in Neural Information Processing Systems* (2011), vol. 24.

[6] Chow, C., and Liu, C. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on 14*, 3 (1968), 462 – 467.

[7] Cocco, S., and Monasson, R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. arXiv:1110.5416, 2011.

[8] Cocco, S., Monasson, R., and Sessak, V. High-dimensional inference with the generalized Hopfield model: Principal component analysis and corrections. *Phys. Rev. E 83* (2011), 051123.

[9] de Palma, A., and Marchal, F. Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems. *Networks and Spatial Economics 2*, 4 (2002), 347–369.

[10] Dong, B., and Zhang, Y. An efficient algorithm for l_0 minimization in wavelet frame based image restoration. *Journal of Scientific Computing In press* (2012).

[11] Eckstein, J. Nonlinear proximal point algorithm using Bregman functions, with applications to convex programming. *Mathematics of Operations Research 18*, 11-48 (1993), 11–50.

[12] Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association 96*, 456 (2001), (1348–1360).

[13] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 3 (2008), 432–441.

[14] Furtlehner, C., Han, Y., Lasgouttes, J.-M., Martin, V., Marchal, F., and Moutarde, F. Spatial and temporal analysis of traffic states on large scale networks. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on* (2010), pp. 1215 –1220.

[15] Furtlehner, C., Lasgouttes, J.-M., and Auger, A. Learning multiple belief propagation fixed points for real time inference. *Physica A: Statistical Mechanics and its Applications 389*, 1 (2010), 149–163.

[16] Furtlehner, C., Lasgouttes, J.-M., and de La Fortelle, A. A belief propagation approach to traffic prediction using probe vehicles. In *Proc. IEEE 10th Int. Conf. Intel. Trans. Sys.* (2007), pp. 1022–1027.

[17] Georges, A., and Yedidia, J. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General 24*, 9 (1991), 2173.

[18] Höfling, H., and Tibshirani, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *JMLR 10* (2009), 883–906.

[19] Hopfield, J. J. Neural network and physical systems with emergent collective computational abilities. *Proc. of Natl. Acad. Sci. USA 79* (1982), 2554–2558.

[20] Hsieh, C., Sustik, M., Dhillon, I., and Ravikumar, K. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems 24*. 2011, pp. 2330–2338.

[21] in Lee, S., Ganapathi, V., and Koller, D. Efficient structure learning of Markov networks using L_1-regularization. In *NIPS* (2006).

[22] Iusem, A. Augmented Lagrangian methods and proximal point methods for convex optimization. *Investigacion Operativa* (1999), 11–50.

[23] Jaynes, E. T. *Probability Theory: The Logic of Science (Vol 1)*. Cambridge University Press, 2003.

[24] Kappen, H., and Rodríguez, F. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation 10*, 5 (1998), 1137–1156.

[25] Lee Dicker, B. H., and Lin, X. Variable selection and estimation with the seamless-l_0 penalty. *Statistica Sinica In press* (2012).

[26] Malioutov, D., Johnson, J., and Willsky, A. Walk-sums and belief propagation in Gaussian graphical models. *The Journal of Machine Learning Research 7* (2006), 2031–2064.

[27] Martin, V., Lasgouttes, J., and Furtlehner, C. Encoding dependencies between real-valued observables with a binary latent MRF. to be submitted, 2012.

[28] Mézard, M., and Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris 103*, 1-2 (2009), 107 – 113.

[29] Mora, T. *Géométrie et inférence dans l'optimisation et en théorie de l'information*. Thèse de doctorat, Université Paris Sud - Paris XI, 2007.

[30] Netrapalli, P., Banerjee, S., Sanghavi, S., and Shakkottai, S. Greedy learning of Markov network structure. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on* (2010), pp. 1295 –1302.

[31] Nguyen, H., and Berg, J. Bethe-Peierls approximation and the inverse Ising model. *J. Stat. Mech.*, 1112.3501 (2012), P03004.

[32] Nguyen, H., and Berg, J. Mean-field theory for the inverse Ising problem at low temperatures. *Phys. Rev. Lett. 109* (2012), 050602.

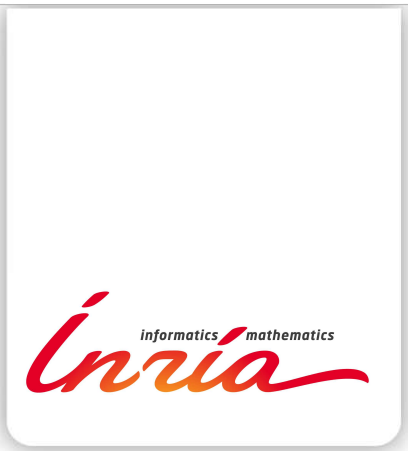
[33] Plefka, T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A: Mathematical and General 15*, 6 (1982), 1971.

[34] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society,Series B 58* (1996), 267–288.

[35] Welling, M., and Teh, Y. Approximate inference in Boltzmann machines. *Artif. Intell. 143*, 1 (2003), 19–50.

[36] Yasuda, M., and Tanaka, K. Approximate learning algorithm in Boltzmann machines. *Neural Comput. 21* (2009), 3130–3178.

[37] Yedidia, J. S., Freeman, W. T., and Weiss, Y. Generalized belief propagation. *Advances in Neural Information Processing Systems* (2001), 689–695.



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399