An Ising Model for Road Traffic Inference

Cyril Furtlehner

Chapter 1
An Ising Model for Road Traffic Inference

Cyril Furtlehner *

Abstract We review some properties of the "belief propagation" algorithm, a distributed iterative map, used to perform Bayesian inference and present some recent work where this algorithm serves as a starting point to encode observation data into a probabilistic model and to process large scale information in real time. A natural approach is based on the linear response theory and various recent instantiations are presented. We will focus on the particular situation where the data have many different statistical components, representing a variety of independent patterns. As an application, the problem of reconstructing and predicting traffic states based on floating car data is then discussed.

1.1 Introduction

The "belief propagation" algorithm BP, originated in the artificial intelligence community for inference problems on Bayesian networks [25]. It is a non-linear iterative map which propagates information on a dependency graph of variables in the form of messages between variables. It has been recognised to be a generic procedure, instantiated in various domains like error correcting codes, signal processing or constraints satisfaction problems with various names depending on the context [18]: the forward-backward algorithm for hidden Markov model selection; the Viterbi algorithm; Gallager's sum-product algorithm in information theory. It has also a nice statistical physics interpretation in the context of mean-field theories, as a minimizer of a Bethe free energy [32], a solver of the cavity equations [21] and its relation to the TAP equations in the spin-glass context [16]. As a noticeable development in the recent years, related to the connection with statistical physics, is

* INRIA Saclay – LRI, Bat. 490, Université Paris-Sud – 91405

the emergence of a new generation of algorithms for solving difficult combinatorial problems, like the survey propagation algorithm [22] for constraint satisfaction problems or the affinity propagation for clustering [6].

The subject of the present review is at first a statistical modelling problem. Assuming a set of high dimensional data, in the form of sparse observations covering a finite fraction of segments in a traffic network, we wish to encode the dependencies between the variables in a probabilistic model, which turns out to be a Markov random field (MRF). We proceed in such a way as to insure that inference on this MRF with BP is optimal in some way so that it can be fast and precise at the same time, offering the possibility to address large scale problems like inferring congestion on a macroscopic traffic network. In Section 1.2 we introduce the BP algorithm and review some of its properties. Section 1.3 is devoted to the general problem of encoding observation data, by addressing the inverse Ising problem. In Section 1.4 a traffic application is described along with the construction of an inference model. Section 1.5 is concerned with the problem of multiplicity of BP fixed points and how to turn this into an advantage when the underlying empirical distribution based on observational data is multi-modal. Finally in Section 1.6 we present some preliminary tests of the method.

1.2 The Belief Propagation Algorithm

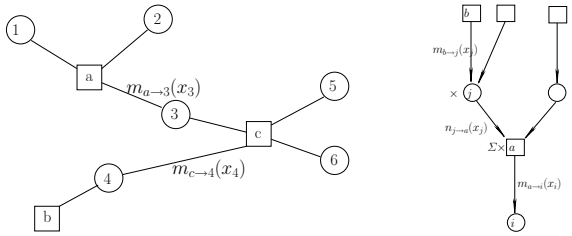
We consider a set of discrete random variables $\mathbf{x} = \{x_i, i \in \mathcal{V}\} \in \{1, \dots, q\}^{|\mathcal{V}|}$ obeying a joint probability distribution of the form

$$\mathcal{P}(\mathbf{x}) = \prod_{a \in \mathcal{F}} \psi_a(x_a) \prod_{i \in \mathcal{V}} \phi_i(x_i), \tag{1.1}$$

where ϕ_i and ψ_a are factors associated respectively to a single variable x_i and to a subset $a \in \mathcal{F}$ of variables, $\mathcal{F}$ representing a set of cliques and $x_a \stackrel{\text{def}}{=} \{x_i, i \in a\}$. The ψ_a are called the "factors" while the ϕ_i are there by convenience and could be reabsorbed in the definition of the factors. This distribution can be conveniently represented with a bi-bipartite graph, called the factor graph [18]; $\mathcal{F}$ together with $\mathcal{V}$ define the factor graph $\mathcal{G}$, which will be assumed to be connected. The set $\mathcal{E}$ of edges contains all the couples $(a, i) \in \mathcal{F} \times \mathcal{V}$ such that $i \in a$. We denote d_a (resp. d_i) the degree of the factor node a (resp. to the variable node i). The factor graph on the Figure 1.1.a corresponds for example to the following measure

$$p(x_1, \dots, x_6) = \frac{1}{Z} \psi_a(x_1, x_2, x_3) \psi_b(x_4) \psi_c(x_3, x_4, x_5, x_6)$$

with the following factor nodes $a = \{1, 2, 3\}$, $b = \{4\}$ and $c = \{3, 5, 6\}$. Assuming that the factor graph is a tree, computing the set of marginal dis-



(a) (b)

Fig. 1.1 Example of factor graph (a) and message propagation rules (b).

tributions, called the belief $b(x_i = x)$ associated to each variable i can be done efficiently. The BP algorithm does this effectively for all variables in one single procedure, by remarking that the computation of each of these marginals involves intermediates quantities called the messages $m_{a\to i}(x_i)$ [resp. $n_{i\to a}(x_i)$] "sent" by factor node a to variable node i [resp. variable node i to factor node a], and which are necessary to compute other marginals. The idea of BP is to compute at once all these messages, using the relation among them as a fixed point equation. Iterating the following message update rules sketched on Figure 1.1.b:

$$\begin{cases} m_{a\to i}(\mathbf{x}_i) \leftarrow \displaystyle\sum_{x_a \setminus x_i} \prod_{j \in a \setminus i} n_{j\to a}(x_j)\psi_a(x_a), \\ n_{i\to a}(x_i) \leftarrow \phi_i(x_i) \displaystyle\prod_{b \ni i} m_{b\to i}(x_i), \end{cases}$$

yields, when a fixed point is reached, the following result for the beliefs,

$$b(x_i) = \frac{1}{Z_i}\phi_i(x_i) \prod_{a \ni i} m_{a\to i}(x_i),$$

$$b(x_a) = \frac{1}{Z_a}\psi_a(X_a) \prod_{i \in a} n_{i\to a}(x_i).$$

This turns out to be exact if the factor graph is a tree, but only approximate on multiply connected factor graphs. As mentioned before, this set of beliefs corresponds to a stationary point of a variational problem [32]. Indeed, consider the Kullback-Leibler divergence between a test joint distribution $b(\mathbf{x})$ and the reference $p(\mathbf{x})$. The Bethe approximation leads to the following functional of the beliefs, including the joint beliefs $b_a(x_a)$ corresponding to each factor:

$$
\begin{aligned}
D_{KL}(b\|p) &= \sum_{\{x\}} b(\{x\}) \log \frac{b(\{x\})}{p(\{x\})} \\
&\approx \sum_{a,x_a} b_a(x_a) \log \frac{b_a(x_a)}{\psi(x_a)\prod_{i\in a} b_i(x_i)} + \sum_{i,x_i} \log \frac{b_i(x_i)}{\phi_i(x_i)} \\
&\stackrel{\text{def}}{=} F_{Bethe} = E - S_{Bethe}.
\end{aligned}
$$

This is equivalent to say that we look for a minimizer of $D_{KL}(b\|p)$ in the following class of joint probabilities:

$$
b(x) = \prod_a \frac{b_a(x_a)}{\prod_{i\in a} b_i(x_i)} \prod_i b_i(x_i), \tag{1.2}
$$

under the constraint that

$$
\sum_{x_a \setminus x_i} b_a(x_a) = b_i(x_i) \qquad \forall a \in \mathcal{F}, \forall i \in a,
$$

and that

$$
\sum_{x \setminus x_a} b(x) \approx b_a(x_a), \qquad \forall a \in \mathcal{F}, \tag{1.3}
$$

is valid, at least approximately. For a multi-connected factor graph, the beliefs b_i and b_a are then interpreted as pseudo-marginal distribution. It is only when $\mathcal{G}$ is simply connected that these are genuine marginal probabilities of the reference distribution p.

There are a few properties of BP that are worth mentioning at this point. Firstly, BP is a fast converging algorithm:

- Two sweeps over all edges are needed if the factor-graph is a tree.
- The complexity scales heuristically like $KN\log(N)$ on a sparse factor-graph with connectivity $K \ll N$.
- It is N^2 for a complete graph.

However, when the graph is multiply connected, there is little guarantee on the convergence [24] even so in practice it works well for sufficiently sparse graphs. Another limit in this case, is that the fixed point may not correspond to a true measure, simply because (1.2) is not normalized and (1.3) is approximate. In this sense, the obtained beliefs, albeit compatible with each other are considered only as pseudo-marginals. Finally, for such graphs, the uniqueness of fixed points is not guaranteed, but it has been shown that:

- stable BP fixed points are local minima of the Bethe free energy [13];
- the converse is not necessarily true [28].

There are two important special cases, where the BP equations simplify: (i) For binary variables: $x_i \in \{0, 1\}$. Upon normalization, the messages are

parametrised as:

$$m_{a\to i}(x_i) = m_{a\to i}x_i + (1 - m_{a\to i})(1 - x_i),$$

which is stable w.r.t. the message update rule. The propagation of information reduces then to the scalar quantity $m_{a\to i}$.
(ii) For Gaussian variables, the factors are necessarily pairwise, of the form

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

$$\phi_i(x_i) = \exp\big(-\frac{1}{2}A_{ii}x_i^2 + h_i x_i\big).$$

Since factors are pairwise, messages can be seen as sent directly from one variable node i to another j with a Gaussian form:

$$m_{i\to j}(x_j) = \exp\Big(-\frac{(x_j - \mu_{i\to j})^2}{2\sigma_{i\to j}}\Big).$$

This expression is also stable w.r.t. the message update rules. Information is then propagated via the 2-component real vector $(\mu_{i\to j}, \sigma_{i\to j})$ with the following update rules:

$$\mu_{i\to j} \longleftarrow \frac{1}{A_{ij}}\Big(h_i + \sum_{k\in\partial i\setminus j}\frac{\mu_{k\to i}}{\sigma_{k\to i}}\Big),$$

$$\sigma_{i\to j} \longleftarrow -\frac{1}{A_{ij}^2}\big[A_{ii} + \sum_{k\in\partial i\setminus j}\sigma_{k\to i}^{-1}\big].$$

At convergence the belief takes the form:

$$b_i(x) = \sqrt{\frac{\sigma_i}{2\pi}}\exp\big(-\frac{(x-\mu_i)^2}{2\sigma_i}\big)$$

with

$$\mu_i = \sigma_i\big(h_i + \sum_{j\in\partial i}\frac{\mu_{j\to i}}{\sigma_{j\to i}}\big)$$

$$\sigma_i^{-1} = A_{ii} + \sum_{j\in\partial i}\sigma_{j\to i}^{-1}$$

and the estimated covariance between x_i and x_j reads

$$\sigma_{ij} = \frac{1}{A_{ij}(1 - A_{ij}^2\sigma_{i\to j}\sigma_{j\to i})}.$$

In this case, there is only one fixed point even on a loopy graph, not necessarily stable, but if convergence occurs, the single variable beliefs provide the exact marginals [29]. In fact, for continuous variables, the Gaussian distribution is the only one compatible with the BP rules. Expectation propagation [23] is a way to address more general distributions in an approximate manner.

1.3 The inverse Ising Problem

Once the underlying joint probability measure is given, this algorithm can be very efficient for inferring hidden variables, but in real applications it is often the case that we have first to build the model from historical data. From now on we assume that we have binary variables ($x_i \in \{0,1\}$). Let $\{\hat{x}_i^j, i \in \mathcal{V}_j^\star, j = 1 \dots M\}$ be a set of observations where M represents the number of distinct, but possibly sparse, observations of the system as a whole and $\mathcal{V}_j^\star$ is the set of nodes observed for the jth observation. We can define an empirical measure based on these historical data as

$$\hat{\mathcal{P}}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{2^{M-|\mathcal{V}_j^\star|}} \prod_{i\in\mathcal{V}_j^\star} \mathbb{1}_{\{x_i=\hat{x}_i\}}.$$

As such this measure is of no use for inference and we have to make some hypothesis to find a suitable inference model. There are of course various possibilities, but a simple one is to consider that the mean and the covariance are given for respectively each variable i and each pair of variable (i,j):

$$\hat{m}_i \stackrel{\text{def}}{=} \frac{1}{\sum_j \mathbb{1}_{\{i\in\mathcal{V}_j^\star\}}} \sum_{j,\mathcal{V}_j^\star \ni i} (2 * \hat{x}_i^j - 1)$$

$$\hat{\chi}_{ij} \stackrel{\text{def}}{=} \frac{1}{\sum_k \mathbb{1}_{\{(i,j)\subset\mathcal{V}_k^\star\}}} \sum_{k,\mathcal{V}_k^\star \supset (i,j)} (2 * \hat{x}_i^k - 1)(2 * \hat{x}_j^k - 1) - \hat{m}_i\hat{m}_j.$$

Let us introduce also the notation for the joint expectation of pairs of spins:

$$\hat{m}_{ij} \stackrel{\text{def}}{=} \hat{\mathbb{E}}(s_i s_j) = \hat{\chi}_{ij} + \hat{m}_i\hat{m}_j.$$

In this case from Jayne's maximum entropy principle [15], imposing these moments to the joint distribution leads to a model pertaining to the exponential family, that is an Ising model for binary variables with $s_i \stackrel{\text{def}}{=} 2x_i - 1$:

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

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. They are obtained as minimiser's of the dual optimization problem, namely

$$(\mathbf{h}^\star, \mathbf{J}^\star) = \underset{(\mathbf{h},\mathbf{J})}{\operatorname{argmin}} \log Z[\mathbf{h}, \mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{ij} J_{ij} \hat{m}_{ij} \tag{1.4}$$

which corresponds to invert the linear response equations:

$$\frac{\partial \log Z}{\partial h_i}[\mathbf{h}, \mathbf{J}] = \hat{m}_i \tag{1.5}$$

$$\frac{\partial \log Z}{\partial J_{ij}}[\mathbf{h}, \mathbf{J}] = \hat{m}_{ij}, \tag{1.6}$$

since $\hat{m}_i$ and $\hat{m}_{ij}$ are given as input to the model. As noted e.g. in [3], the solution is minimizing the cross entropy, a Kullback-Leibler distance between the empirical distribution based on observation and the Ising model:

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

The set of equations (1.5,1.6) 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 $Z[\mathbf{h}, \mathbf{J}]$.

- A common approach is based on the Plefka expansion [26], of the Gibbs free energy by making the assumption that the J_{ij} are small. The picture is then of a weakly correlated uni-modal probability measure. For example, the recent approach proposed in [3] is based on this assumption.
- A second possibility is to assume that relevant coupling J_{ij} have locally a tree-like structure. The Bethe approximation mentioned in the previous section is then used with possibly loop corrections. Again this corresponds to having a weakly correlated uni-modal probability measure and these kind of approached are referred as pseudo-moment matching methods in the literature for the reason explained in the previous section. For example the approach proposed in [17, 30, 20, 31] are based on this assumptions.
- In the case where a multi-modal distribution is expected, then a model with many attraction basin is to be found and Hopfield like model [14, 4] are likely more relevant in this case.

Gibbs free energy: 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

$$\frac{\partial F}{\partial h_i} = -m_i. \tag{1.7}$$

Note that by duality we have

$$\frac{\partial G}{\partial m_i} = h_i(\mathbf{m}), \tag{1.8}$$

and

$$\Big[\frac{\partial^2 G}{\partial m_i \partial m_j}\Big] = -\Big[\frac{\partial^2 F}{\partial h_i \partial h_j}\Big]^{-1} = \big[\chi\big]^{-1}_{ij}. \tag{1.9}$$

i.e. the inverse susceptibility matrix. Finding a set of J_{ij} satisfying this last relation along with (1.8) 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

$$\frac{\partial G}{\partial J_{ij}} = -m_{ij}. \tag{1.10}$$

Plefka's expansion: 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:

$$G[\mathbf{m},\alpha\mathbf{J}] = \mathbf{h}^T(\mathbf{m},\alpha)\mathbf{m} + F[\mathbf{h}(\mathbf{m},\alpha),\alpha\mathbf{J}] \tag{1.11}$$

$$= G_0[\mathbf{m}] + \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} G_n[\mathbf{m},\mathbf{J}] \tag{1.12}$$

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 is always fulfill (1.7). This precisely is the Plefka's expansion, and each term of the expansion (1.12) can be obtained by successive derivation of (1.11). 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

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

using (1.7), the first derivative of (1.11) w.r.t α gives

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

while the second reads:

$$\frac{d^2G[\mathbf{m},\alpha\mathbf{J}]}{d\alpha^2} = -\mathbb{E}^c_\alpha\big(H_J^2\big) - \sum_i \frac{dh_i(\mathbf{m},\alpha)}{d\alpha}\mathbb{E}^c_\alpha\big(H_J s_i\big).$$

In these expressions, it is the connected part of the expectation, noted

$$\mathbb{E}^c[XY] \stackrel{\text{def}}{=} \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

which appears when deriving on the free energy. To get successive derivative of $\mathbf{h}(\mathbf{m},\alpha)$ one can use (1.8). Another possibility is to express the fact that $\mathbf{m}$ is fixed,

$$\begin{aligned}\frac{dm_i}{d\alpha} = 0 &= -\frac{d}{d\alpha}\frac{\partial F[\mathbf{h}(\alpha),\alpha\mathbf{J}]}{\partial h_i}\\ &= \sum_{i,j} h_j'(\alpha)\mathbb{E}^c_\alpha(s_i s_j) + \mathbb{E}^c_\alpha(H_J s_i),\end{aligned}$$

giving

$$h_i'(\alpha) = -\sum_j [\chi^{-1}]_{ij}\mathbb{E}^c_\alpha(H_J s_j).$$

To get the first two terms in the Plefka's expansion we need to compute these quantities at $\alpha = 0$:

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

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

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

and correspond respectively to the mean field and to the TAP approximation. Higher order terms have been computed in [10].

Linear response approximate solution: At this point we are in position to find an approximate solution to the inverse Ising problem, either by inverting equation (1.9) or (1.10). To get a solution at a given order n in the coupling, solving (1.10) requires G at order $n+1$, while it is needed at order n in (1.9).

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 (1.10) leads directly for the basic mean-field solution to:

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

At this level of approximation for G, using (1.8) 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 correspond precisely to the TAP equations. Using now (1.9) 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,

$$J_{ij}^{TAP} = -\frac{2[\hat{\chi}^{-1}]_{ij}}{1+\sqrt{1-8\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}}, \tag{1.13}$$

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 approximation: In this case we remark first that when the graph formed by the observed correlations $\hat{\chi}_{ij}$ is a tree then the form (1.2) of the joint probability corresponding to the Bethe approximation yields actually an exact solution to the inverse problem (1.4)

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

where the $\hat{p}$ are the single and pair variables empirical marginal given by the observations. Rewriting this expression as an Ising model results in the following expressions for the parameters

$$h_i = \frac{1-d_i}{2}\log\frac{\hat{p}_i^1}{\hat{p}_i^0} + \frac{1}{4}\sum_{j\in i}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{10}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{00}}\Big), \tag{1.14}$$

$$J_{ij} = \frac{1}{4}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{00}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}}\Big), \tag{1.15}$$

while the partition function simply reads

$$Z_{Bethe}[\hat{p}] = \exp\big(-\frac{1}{4}\sum_{ij}\log(\hat{p}_{ij}^{00}\ \hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}\ \hat{p}_{ij}^{11}) - \sum_i \frac{1-d_i}{2}\log(\hat{p}_i^0\ \hat{p}_i^1)\big), \tag{1.16}$$

and where the $\hat{p}$'s are parametrized as

$$\hat{p}_i^{x_i} \stackrel{\text{def}}{=} \hat{p}(x_i = \frac{1+s}{2}) = \frac{1}{2}(1+m_i s), \tag{1.17}$$

$$\hat{p}_{ij}^{x_i x_j} \stackrel{\text{def}}{=} \hat{p}(x_i = \frac{1+s}{2}, x_j = \frac{1+s'}{2}) = \frac{1}{4}(1+m_i s + m_j s' + m_{ij} s s') \tag{1.18}$$

are the empirical frequency statistics given by the observations for $m \equiv \hat{m}$. The corresponding Gibbs free energy can then be written explicitly using (1.14,1.15,1.16). With fixed magnetization's 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 (1.15). For the linear response, we get from (1.14):

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

Using (1.15), 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 [2]

$$[\hat{\chi}^{-1}]_{ij} = \Big[\frac{1-d_i}{1-m_i^2} + \sum_{k\in\partial i} \frac{1-m_k^2}{(1-m_i^2)(1-m_k^2)-\chi_{ik}^2}\Big]\delta_{ij}$$
$$- \frac{\chi_{ij}}{(1-m_i^2)(1-m_j^2)-\chi_{ij}^2}\ \delta_{j\in\partial i}. \tag{1.19}$$

equivalent to the original one derived in [30] 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 has to hold and J_{ij} is given by (1.15) using $\chi_{ij} = \hat{\chi}_{ij}$.
- $[\hat{\chi}^{-1}]_{ij} = 0$ then $J_{ij} = 0$ while $\hat{\chi}_{ij}$ can be non-zero, because (1.15) does not hold in that case.

Finally complete consistency of the solution is checked on the diagonal elements in (1.19). If full consistency is not verified, these equation can nevertheless be used to find approximate solutions. Remark that if we restrict the set of equations (1.19), 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 [20, 31] actually correspond to different heuristics for finding approximate solutions to this set of constraints. As noted in [2] a direct way to proceed is to eliminate χ_{ij} in the equations obtained from (1.15) and (1.19):

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

This leads directly to

$$J_{ij}^{Bethe} = -\frac{1}{2}\,\mathrm{atanh}\Big(\frac{2[\hat{\chi}^{-1}]_{ij}}{\sqrt{1+4(1-\hat{m}_i^2)(1-\hat{m}_j^2)[\hat{\chi}^{-1}]_{ij}^2} - 2\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}\Big), \tag{1.20}$$

while the corresponding computed of ξ_{ij}, instead of the observed one $\hat{\xi}_{ij}$, has to be inserted in (1.14) to be fully consistent. Note that J_{ij}^{Bethe} and J_{ij}^{TAP} coincide at second order in $[\hat{\chi}^{-1}]_{ij}$.

1.4 Application context

1.4.1 Road Traffic Inference

Once the underlying joint probability measure is given, the BP algorithm can be very efficient for inferring hidden variables, but in real applications it is often the case that we have first to build the model. This is precisely the case for the application that we are considering concerning the reconstruction and prediction of road traffic conditions, typically on the secondary network from sparse observations. Existing solutions for traffic information are classically based on data coming from static sensors (magnetic loops) on main arterial roads. These devices are far too expensive to be installed everywhere on the traffic network and other sources of data have to be found. One recent solution comes from the increasing number of vehicles equipped with GPS and able to exchange data through cellular phone connections for example, in the form of so-called Floating Car Data (FCD). Our objective in this context is to build an inference schema adapted to these FCD, able to run in real time and adapted to large scale road networks, of size ranging from 10^3 to 10^5 road segments. In this respect, the BP algorithm seems very well suited, but the difficulty is to construct a model based on these FCD. To set an inference schema, we assume that a large amount of FCD sent by probe vehicles concerning some area of interest are continuously collected over a reasonable period of time (one year or so) such as to allow a finite fraction (a few percents) of road segments to be covered in real time. Schematically the inference method works as follows:

- Historical FCD are used to compute empirical dependencies between contiguous segments of the road network.
- These dependencies are encoded into a graphical model, which vertices are (segment,timestamps) pairs attached with a congestion state, i.e. typically CONGESTED/NOT-CONGESTED.
- Congestion probabilities of segments that are unvisited or sit in the short-term future are computed with BP, conditionally to real-time data.

On the factor-graph, the information is propagated both temporally and spatially. In this perspective, reconstruction and prediction are on the same footing, even though prediction is expected to be less precise than reconstruction.

1.4.2 An Ising model for traffic.

Binary latent state and traffic index

When looking at standard traffic information systems, the representation of the congestion network suggests two main traffic states: uncongested (green) or congested (red) as shown on Figure 1.1. If we take seriously this seemingly

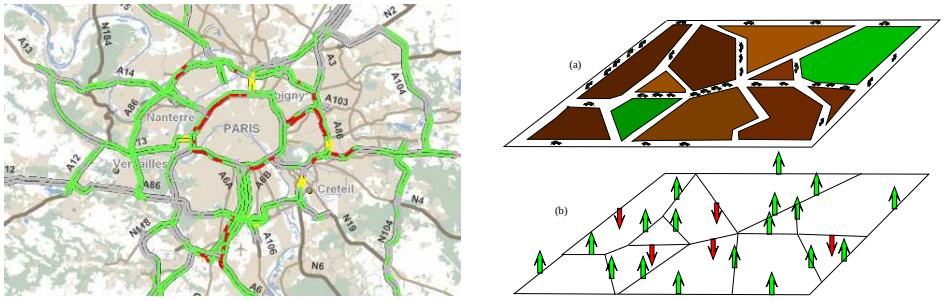


Fig. 1.1 Underlying Ising modelling of traffic configurations.

empirical representation, we are asking the question: Is it possible to encode traffic data on the basis of a binary latent state $s_{i,t} \in \{-1, 1\}$ (Ising) corresponding to congested/non-congested state. As a corollary, what is the proper criteria to define the congested/uncongested state and for which purpose? In some recent work we have proposed an answer to this question [9, 19]. As said before, static sensors and probe vehicles delivers real-valued information, i.e. respectively speed and density, and speed and travel time. For each segments, we may potentially collect a distribution $\hat{f}$ of travel time and it is not clear how to decide from this distribution, whether a link is congested or not given an newly observed travel time. A straightforward possibility is to consider the mean travel time, or even more robust, the median travel time as a separator of the two state. The way we actually see this encompass this possibility, but is not limited to it. The idea is to define the latent binary state τ $(= \frac{1+s}{2})$ associated to some travel time x in an abstract way through the mapping:

$$\Lambda(x) \stackrel{\text{def}}{=} P(\tau = 1|x).$$

This means that an observation x is translated into a conditional probability for the considered segment to be congested. This number $\Lambda(x) \in [0, 1]$, represents our practical definition for the *traffic index*.

Using Bayes rules and the Boolean notation $\bar{\tau} \stackrel{\text{def}}{=} 1 - \tau$, we obtain

$$P(x|\tau) = \Big(\frac{\Lambda(x)}{p_\Lambda}\tau + \frac{1 - \Lambda(x)}{1 - p_\Lambda}\bar{\tau}\Big)\hat{f}(x). \tag{1.21}$$

where $p_\Lambda \stackrel{\text{def}}{=} P(\tau = 1)$. The normalization constraint imposes

$$p_\Lambda = \int \Lambda(x)\hat{f}(x)dx. \tag{1.22}$$

A certain amount of information can be stored in this mapping. A special case mentioned before corresponds to having for Λ a step function, i.e.

$$\Lambda(x) = \mathbb{1}_{\{x>x^*\}}, \tag{1.23}$$

with an adjustable parameter corresponding to the threshold x^*. Another parameter free possibility is to use the empirical cumulative distribution:

$$\Lambda(x) = \hat{F}(x) \stackrel{\text{def}}{=} P(\hat{x} < x). \tag{1.24}$$

Now, Given a map Λ, an obvious way to convert back a probability $u = P(\tau = 1)$ into a travel time consist then simply in using, when it exits, the inverse map:

$$\hat{x} = \Lambda^{-1}(u). \tag{1.25}$$

Actually another legitimate way to proceed is based on the conditional probability (1.21) to yields the following estimator:

$$\hat{x} = \underset{y}{\text{argmin}}\ \mathbb{E}\big(\|x - y\|_r\big),$$

where the expectation is taken from the probability distribution

$$P(x) = P(x|\tau = 1)u + P(x|\tau = 0)(1 - u),$$

and where $\|x-y\|_r$ represents the loss function, measuring the error between the prediction x and the actual value y. In this last case, the natural requirement that we seek for Λ is that the mutual information between x and τ be maximal. This reads

$$\begin{aligned} I(x,\tau) &\stackrel{\text{def}}{=} \sum_{\tau\in\{0,1\}} \int dx P(x,\tau) \log \frac{P(x,\tau)}{P(x)P(\tau)}, \\ &= \int dx \big(\Lambda(x)\log\Lambda(x) + (1-\Lambda(x))\log(1-\Lambda(x))\big)\hat{f}(x) - h(p_\Lambda), \\ &= \int du\ h\big[\Lambda\big(\hat{F}^{-1}(u)\big)\big] - h(p_\Lambda), \end{aligned}$$

after introducing the binary information function $h(x) \stackrel{\text{def}}{=} x\log x + (1 - x)\log(1 - x)$. In this form, h being convex, reaching its maximum at $x = 0$ and $x = 1$, its minimum at $x = 1/2$, it is then straightforward to obtain that the step function (1.23) with $x^* = \hat{F}^{-1}(1/2)$ corresponding to the median

observation is the limit function which maximizes $I(x,\tau)$. If instead we use the inverse map Λ^{-1}, the mutual information between x and τ is not relevant. Without any specific hypothesis on the distribution of beliefs that BP should generate, a simple requirement is then to impose a minimum information i.e. a maximum entropy contained in the variable $u = \Lambda(x)$, which probability density is given by

$$dF(u) \stackrel{\text{def}}{=} \int \delta\big(u - \Lambda(x)\big) d\hat{F}(x)$$

$$= \frac{d\hat{F}}{d\Lambda}\big(\Lambda^{-1}(u)\big).$$

Using this and the change of variable $x = \Lambda^{-1}(u)$ yields the entropy

$$S[u] = -\int d\hat{F}(x) \frac{d\hat{F}}{d\Lambda}(x) = -D_{KL}(\hat{F}\|\Lambda),$$

expressed as the opposite of the relative entropy between F and Λ. Without any further constraint, this leads to the fact that $\Lambda = F$ is the optimal mapping. In both cases, additional constraints comes from the fact that we want a predictor $\hat{x}$ minimizing a loss function $\|\hat{x} - x\|_r$ which depends on the choice of the Euclidean norm $\mathbb{L}_r$ (see [19] for details).

Global inference model:

In fact the mapping between real-valued observations and the binary latent states is only one element of the model. The general schema of our Ising based inference model is sketched on Figure 1.2. It can be decomposed into 4 distinct pieces:

- A single variable statistical model translating real-valued observations into binary latent states.
- A pairwise statistical model of the dependency between latent states.
- A MRF model to encode the network of dependencies.
- The Belief propagation algorithm to decode a partially observed network.

It is based on a statistical description of traffic data which is obtained by spatial and temporal discretization, in terms of road segments i and discrete time slots t corresponding to time windows of typically a few minutes, leading to consider a set of vertices $\mathcal{V} = \{\alpha = (i,t)\}$. To each vertex is attached a microscopic degree of freedom $x_\alpha \in E$, as a descriptor of the corresponding segment state (e.g. $E = \{0,1\}$, 0 for congested and 1 for fluid). The model itself is based on historical data in form of empirical marginal distributions $\hat{p}(x_\alpha)$, $\hat{p}(x_\alpha, x_\beta)$, giving reference states and statistical interactions between degrees of freedom. Finally, reconstruction and prediction are produced in

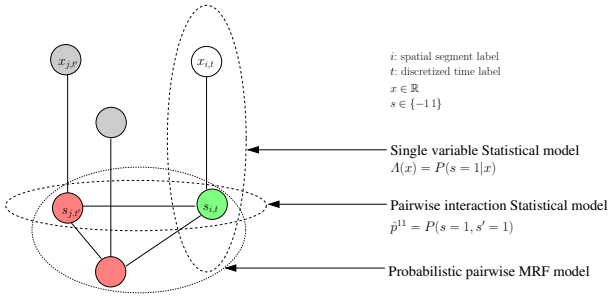


Fig. 1.2 Sketch of the Ising based inference schema.

the form of conditional marginal probability distribution $p(x_\alpha|\mathcal{V}^*)$ of hidden variables in $\mathcal{V}\backslash\mathcal{V}^*$, conditionally to the actual state of the observed variables in the set $\mathcal{V}^*$.

In addition to this microscopic view, it is highly desirable to enrich the description with macroscopic variables, able in particular to capture and encode the temporal dynamics of the global system. These can be obtained by some linear analysis, e.g. PCA or with non-linear methods of clustering providing possibly hierarchical structures. Once some relevant variables are identified, we can expect to have a macroscopic description of the system, which can potentially be easily coupled to the microscopic one, by adding some nodes into the factor-graph. These additional degrees of freedom would be possibly interpreted in terms of global traffic indexes, associated to regions or components.

The binary latent states are used to model the interactions in a simplified way enabling for large scale applications. Trying to model exactly the pairwise dependencies at the observation level is potentially too expensive from the statistical as well as the computational viewpoint. So the pairwise model sketched on Figure 1.2 corresponds to

$$P(x_i, x_j) = \sum_{\tau,\tau'} \hat{p}_{ij}(\tau, \tau') P(x_i|\tau) P(x_j|\tau'),$$

with $P(x|\tau)$ given in (1.21) and $\hat{p}_{ij}$ to be determined from empirical frequency statistics. Since a probability law of two binary variables requires three independent parameters; two of them are already being given by individuals marginals probabilities $\hat{p}_i^1 \stackrel{\text{def}}{=} P(\tau_i = 1)$ according to (1.22). For each pair of variables, one parameter remains therefore to be fixed. By convenience we consider the coefficient

$$p_{ij}^{11} \stackrel{\text{def}}{=} P(\tau_i = 1, \tau_j = 1),$$

and write a moment matching constraint in the traffic index space[2]. We obtain

$$\hat{p}_{ij}^{11} = \hat{p}_i^1 \hat{p}_j^1 + \frac{\widehat{\mathrm{cov}}\big[\Lambda_i(x_i), \Lambda_j(x_j)\big]}{\big(2\hat{\mathbb{E}}[\Lambda_i(x)] - 1\big)\big(2\hat{\mathbb{E}}[\Lambda_j(x)] - 1\big)},$$

involving the empirical expectation of indexes, $\hat{\mathbb{E}}[\Lambda_i(x)]$ and empirical covariance between indexes $\widehat{\mathrm{cov}}\big[\Lambda_i(x_i), \Lambda_j(x_j)\big]$ obtained from observation data.

1.4.3 MRF model and pseudo moment matching calibration

At the microscopic level, the next step is to define the MRF i.e. the Ising model, on which to run BP with good inference properties. Recall that we try to answer two related questions:

- Given the set of coefficients $\hat{p}(\tau_{i,t})$ and $\hat{p}(\tau_{i,t}, \tau_{j,t})$, considered now as model input, what is the joint law $P(\{\tau_{i,t}, (i,t) \in \mathcal{V}\})$?
- Given actual observations $\{x^*_{i,t}, (i,t) \in \mathcal{V}^*\}$, how to infer $\{x_{i,t}, (i,t) \in \mathcal{V}\backslash\mathcal{V}^*\}$?

The solution that we have been exploring [9] is based on the the Bethe approximation described in Section 1.4.2. It consists to use the Bethe approximation (1.2) for the encoding and the belief-propagation for the decoding, such that the calibration of the model is coherent with the inference algorithm. In particular, when there is no real time observation, the reference point is given by the set of historical beliefs, so we expect that running BP on our MRF delivers precisely these beliefs. Stated differently, we look for the ϕ and ψ defining the MRF in (1.1) such that the beliefs match the historical marginals:

$$b_i(\tau_i) = \hat{p}_i(\tau_i), \qquad \text{and} \qquad b_{ij}(\tau_i, \tau_j) = \hat{p}_{ij}(\tau_i, \tau_j).$$

As explained in Section 1.3 there is an explicit solution to this problem, because BP is coherent with the Bethe approximation [32], and thus any BP fixed point $\mathbf{b}$ has to verify

$$\mathcal{P}(\tau) = \prod_{i\in\mathcal{V}} \phi_i(\tau_i) \prod_{(i,j)\in\mathcal{F}} \psi_{ij}(\tau_i, \tau_j) \propto \prod_{i\in\mathcal{V}} b_i(\tau_i) \prod_{(i,j)\in\mathcal{F}} \frac{b_{ij}(\tau_i, \tau_j)}{b_i(\tau_i) b_i(\tau_j)}. \quad (1.26)$$

As a result, a *canonical choice* for the functions ϕ and ψ is simply

$$\phi_i(\tau_i) = \hat{p}_i(\tau_i), \qquad \psi_{ij}(\tau_i, \tau_j) = \frac{\hat{p}_{ij}(\tau_i, \tau_j)}{\hat{p}_i(\tau_i)\hat{p}_j(\tau_j)}, \quad (1.27)$$

[2] Potentially any arbitrary mapping $\phi(x)$ could be considered to perform the moment matching

along with $m_{i\to j}(x_j) \equiv 1$ as a particular BP fixed point. In addition, from the re-parametrization property of BP [27], any other choices verifying (1.26) produces the same set of fixed points with the same convergence properties. Note that more advanced methods than the strict use of the Bethe approximation, presented in the preceding section could be used as well, but as we shall see in the next sections, the hypothesis that traffic data could be well represented by one single BP fixed point might not be fulfilled. In that case the linear response, which takes a BP fixed point as a reference starting point might be of limited efficiency. Instead of trying to use more accurate version of the linear response, we have followed a different route, by enriching the Bethe approximation with an adjustable parameter, interpreted as an inverse temperature, in order to better calibrate the model in a multiple BP fixed point context. This will be explained in the next section.

Next, for the decoding part, inserting information in real time in the model is done as follows. In practice, observations are in the form of real numbers like speed or travel time. One possibility is to project such an observation onto the binary state $\tau_i = 0$ or $\tau_i = 1$, but this proves to be too crude. As explained in Section 1.4.2, since the output of BP is anyway in the form of beliefs, i.e. real numbers in $[0, 1]$, the idea is to exploit the full information by defining a correspondence between observations x_i and probabilities $p^*(\tau_i = 1)$. The optimal way of inserting this quantity into the BP equations is obtained variationally by imposing the additional constraint $b_i(\tau_i) = p^*(\tau_i)$, which results in modified messages sent from $i \in \mathcal{V}^*$, now reading [7]

$$n_{i\to j}(x_i) = \frac{p_i^*(\tau_i)}{m_{j\to i}(\tau_i)}.$$

This leads to a new version of BP which convergence properties have been analyzed in [19]. This works well in practice, in particular when compared to some heuristic method consisting in giving a bias to the local field of the observed variables as shown on the Figure 1.2 discussed in the last Section 1.6.

1.5 Multiple BP fixed points for multiple traffic patterns

Some experiments with a preliminary version of this procedure [9] indicate that many BP fixed point can exist in absence of information, each one corresponding to some congestion pattern like e.g. congestion/free flow. We have analyzed in [8] the presence of multiple fixed points by looking at a study case, and we outline some of the results in this section. In this study, we considered a generative hidden model of traffic in the form of a probabilistic mixture, with each component having a simple product form:

$$P_{\text{hidden}}(\tau) \stackrel{\text{def}}{=} \frac{1}{C}\sum_{c=1}^{C}\prod_{i\in\mathcal{V}} p_i^c(\tau_i). \tag{1.28}$$

C represents the number of mixture components. Although (1.28) is quite general, the tests are conducted with $C \ll N$, with well separated components of the mixture. The single sites probabilities $p_i^c \stackrel{\text{def}}{=} p_i^c(1)$, corresponding to each component c, are generated randomly as i.i.d. variables,

$$p_i^c = \frac{1}{2}(1+\tanh h_i^c)$$

with h_i^c uniformly distributed in some fixed interval $[-h_{max}, +h_{max}]$. The mean of p_i^c is therefore $1/2$ and its variance reads

$$v \stackrel{\text{def}}{=} \frac{1}{4}\mathbb{E}_h\big(\tanh^2(h)\big) \in [0, 1/4].$$

This parameter v implicitly fixed by h_{max} represents the average level of "polarizability" of the variables in each cluster: $v = 0$ corresponds to $p_i^c = 1/2$ while $v = 1/4$ corresponds to $p_i^c \in \{0, 1\}$ with even probabilities. The interpretation of this model is that traffic congestion is organized in various patterns, which can show up at different times. We then studied the behavior of our inference model on the data generated by this hidden probability by adding a single parameter α into its definition (1.27):

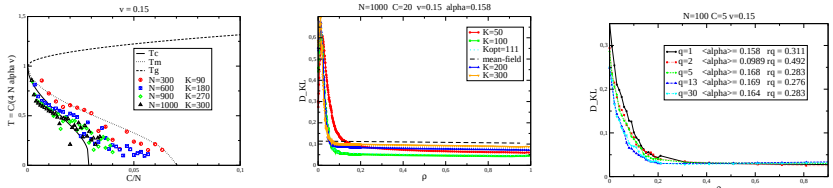
$$\phi_i(\tau_i) = \hat{p}_i(\tau_i), \quad \psi_{ij}(\tau_i,\tau_j) = \Big(\frac{\hat{p}_{ij}(\tau_i,\tau_j)}{\hat{p}_i(\tau_i)\hat{p}_j(\tau_j)}\Big)^{\alpha} \tag{1.29}$$

where $\hat{p}_i$ and $\hat{p}_{ij}$ are again the $1-$ and $2-$ variable frequency statistics that constitute the input of the model, while (1.28) is assumed to be unknown. This parameter α, which can be interpreted as an inverse temperature in the Ising model, is there to compensate for saturation effects ($\alpha < 1$), when the coupling between variables are too large. This is due to some over-counting of the dependencies between variables which may occur in a multiply connected graph. Still, in complement, some sparsity can be imposed to the factor graph with help of some link selection procedure that reduce the mean connectivity to K.

The typical numerical experiment we perform, given a configuration randomly sampled from (1.28), is to reveal gradually the variables $\tau_{\mathcal{V}^*}$ in a random order and compute conditional predictions for the remaining unknown variables. We then compare the beliefs obtained with the true conditional marginal probabilities $P(\tau_i = \tau|\tau_{\mathcal{V}^*})$ computed with (1.28), using an error measure based on the Kullback-Leibler distance:

$$D_{\text{KL}} \stackrel{\text{def}}{=} \Big\langle \sum_{\tau\in\{0,1\}} b_i(\tau)\log\frac{b_i(\tau)}{P(\tau_i=\tau|\tau_{\mathcal{V}^*})}\Big\rangle_{\mathcal{V}^*},$$

where $<>_{\mathcal{V}^*}$ means an average taken on the set of hidden variables. A sample



(a) (b) (c)

Fig. 1.1 (a) Phase diagram of the Hopfield model and optimal points found experimentally.(b) D_{KL} error as a function of observed variables ρ for the single parameter model with $N = 1000$ and $C = 20$ and various pruning levels and for the muliparameter model $N = 100$ $C = 5$ with various number of calibrated parameters ranging from 1 to 30 (c).

test shown on Figure 1.1.b indicates for example that, on a system with 10^3 variables, it is possible with our model to infer with good precision a mixture of 20 components by observing 5% of the variables. To interpret these results, letting $s_i = 2\tau_i - 1$, we first identify the Ising model corresponding to the MRF given by (1.29):

$$\mathcal{P}(\mathbf{s}) = \frac{1}{Z} e^{-\beta H[\mathbf{s}]},$$

with an inverse temperature β and the Hamiltonian

$$H[\mathbf{s}] \stackrel{\text{def}}{=} -\frac{1}{2} \sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i.$$

The identification reads:

$$\beta J_{ij} = \frac{\alpha}{4} \log \frac{\hat{p}_{ij}(1,1)\hat{p}_{ij}(0,0)}{\hat{p}_{ij}(0,1)\hat{p}_{ij}(1,0)},$$

$$\beta h_i = \frac{1 - \alpha K_i}{2} \log \frac{\hat{p}_i(1)}{\hat{p}_i(0)} + \frac{\alpha}{4} \sum_{j \in i} \log \frac{\hat{p}_{ij}(1,1)\hat{p}_{ij}(1,0)}{\hat{p}_{ij}(0,1)\hat{p}_{ij}(0,0)},$$

Then, in the limit $C \gg 1$, $N \gg C$ and fixed average connectivity K, we get asymptotically a mapping to the Hopfield model [14]. The relevant parameters in this limit are $\eta = C/N$ and the variance $v \in [0, 1/4]$ of the variable bias in the components. In this limit, the Hamiltonian is indeed similar to the one governing the dynamics of the Hopfield neural network model:

$$H[\mathbf{s}] = -\frac{1}{2N} \sum_{i,j,c} \xi_i^c \xi_j^c s_i s_j - \sum_{i,c} h_i^c \xi_i^c s_i,$$

$$\text{with} \quad \xi_i^c \stackrel{\text{def}}{=} \frac{p_i^c(1) - \frac{1}{2}}{\sqrt{v}} \quad \text{and} \quad h_i^c = \frac{C}{2\alpha K\sqrt{v}} - \frac{2C\sqrt{v}}{K} \sum_{j \in i} \text{Cov}(\xi_i^c, \xi_j^c),$$

the inverse temperature given by the mapping reads

$$\beta = \frac{4\alpha v K}{C}.$$

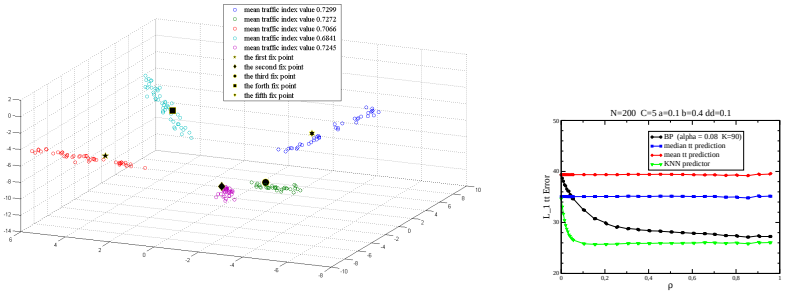
Using mean-field methods, the phase diagram of this model has been established [1]. There are 3 phases, separated by the transition lines T_g, between the paramagnetic phase and the spin glass phase, and T_c, between the spin glass phase and the ferromagnetic phase (see Figure 1.1.a). The latter corresponds to the so-called *Mattis states*, i.e. to spin configurations correlated with one of the mixture components, of direct relevance w.r.t. inference. Locating the various models obtained in this diagram as on Figure 1.1.a, helps to understand whether inference is possible or not with our MRF model.

We have also tested a multi-parameter version of the model in [8], where the links sorted according to the mutual information they contribute for, and grouped them into a certain number of quartiles: to each quartile q we associated a parameter $\alpha_q < 1$. Using a calibration procedure based on a stochastic optimization algorithm CMAES [12], we can see on some examples a significant improvement of the model, as seen on the example presented on Figure 1.1.c.

To summarize, the main lessons of this theoretical study are

- the various components of a probabilistic mixture with weak internal correlations maybe correctly accounted for by our inference model. It is able to to associate in an unsupervised way one BP fixed point to each component of the mixture.
- the mechanism for that can be understood by some asymptotic analysis which reveals a connection with a Hopfield model, where the main patterns corresponds to the components of the mixture. The phase diagram of the Hopfield model gives then relevant indications on whether inference will be easy, difficult or impossible depending on the ratio $N_{states}/N_{variables}$ and on the mean internal variance of the variables v within each state.
- the model can be easily generalized to a multi-parameter version to improve its accuracy with help of a calibration based on a robust optimization strategy like the CMAES algorithm for example.

An example of BP fixed points associated to the mixture's components is given on Figure 1.2. Note that in this figure, the 3-d projection space corresponds to the first principal components of the travel time vectors. The set of beliefs corresponding to each fixed point is converted into travel time through the inverse mapping given in 1.25 and projected on this 3-d space. The reconstruction experiments shown on Figure 1.2 but explained in the next section, shows that the model, albeit very economical as compared to the K-nearest



(a) (b)

Fig. 1.2 Segmentation and BP fixed point identification (a) for synthetic travel time data corresponding to a mixture with five components with internal correlations. Corresponding reconstruction experiment where the L_1 travel time error is plotted against the fraction ρ of observed variables and compared with a K-NN predictor considered here as ground truth.

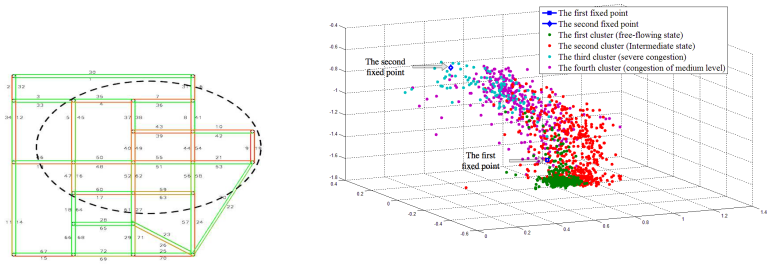
neighbor (K-NN) predictor, is able to predict correctly real-valued hidden variables.

1.6 Experiments with Synthetic and Real Data

Using both synthetic and real data we perform two kind of numerical tests:

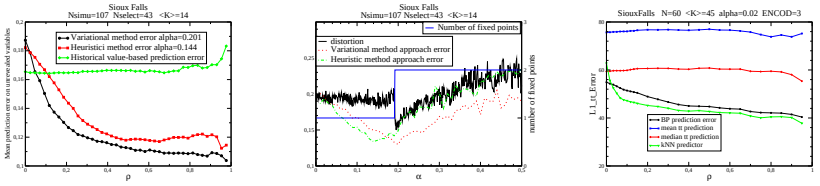
(i) reconstruction/prediction experiments
(ii) automatic segmentation and BP fixed points identification

In experiments of the type (i), the data set is divided into two parts, one corresponding to the learning set, used to build the model and the other part corresponding to the test set, used to perform the tests. In the reconstruction tests traffic configuration corresponding to a single time layer are extracted from the test set; a fraction ρ of variables are chosen at random to be revealed, and while this fraction is progressively increased, travel time are inferred for the hidden one. In prediction experiments, traffic configuration corresponding to successive time layers are selected, with present time t_0 separating the time layers into 2 equal parts, one corresponding to past and the other to future. Observed variable are necessarily selected in the past window time. Variables with time stamp $t = t0$ (present) or $t > t0$ (future) are inferred, i.e. reconstructed or predicted respectively. In the type (ii) experiments, on one hand an automatic clustering of the data on reduced dimensional space is performed with machine learning techniques[11]. On the other hand the BP fixed points obtained at $\rho = 0$ are listed[8, 7] and compared to segmentation in the reduced dimensional space.



(a) (b)

Fig. 1.1 (a) Siouxfalls network. (b) Automatic segmentation of simulated Siouxfalls data and BP fixed point identification projected in the 3-d main PCA space.



(a) (b) (c)

Fig. 1.2 Comparing an heuristic method and a variational one for inserting real-time information (a) and (b). for the Siouxfalls network data. Figure (b) indicates that the optimal value of α for traffic reconstruction is coherent with the best clustering value in the variational case. Reconstruction experiment on the same data using the mapping based on the cumulative and compared with a K-NN predictor.

A first set of data has been generated with the traffic simulator "METROPOLIS" [5], for the benchmark network called Siouxfalls shown on Figure 1.1.a. An example of the automatic clustering of spatial configurations with the corresponding BP fixed points associated to free flow and congestion is shown on Figure 1.1.b. On Figure 1.2.a and b, two ways of inserting information are compared, the variational one mentioned in Section 1.4.3 with an heuristic one based on local fields. In both cases the optimal tuning of α yields two BP fixed points, but the variational method gives better results on reconstruction test and is more consistent with clustering results (same optimal value of α). On Figure 1.2.c a reconstruction experiment on simulated Siouxfalls Metropolis data is shown, using the cumulative distribution for both the encoding (1.24) and inverse decoding (1.25) of traffic indexes. The performance is similar to a K-NN predictor, although much more economical. To set a scale of comparison predictors obtained from historical mean and median values are also shown on the same plot.

To perform tests on real data, we have also considered a dataset consisting of travel times measured every 3 minutes over 2 years of a highway segmented

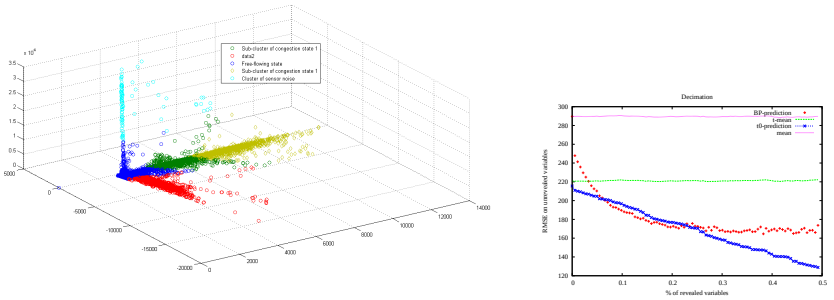


Fig. 1.3 (a) Automatic segmentation of Highway data projected on 3-d dominant PCA space. (b) Error on travel time for a BP prediction of 3 time layers in future as a function of the fraction of observed variables at t_0. Comparison is made with a predictor combining recent available observations with historical time dependent mean.

into 62 segments. We use for each segment $i = 1 \dots 62$ a weighted cumulative travel time distribution based on the automatic segmentation for the traffic index encoding. The automatic segmentation using non-negative matrix factorization techniques [11] is displayed on the 3-d Figure 1.3.a. Results of a short term horizon prediction test is displayed on Figure 1.3.b. showing reasonable performance even though highway data do not correspond to the situation for which the model was designed.

1.7 Conclusion

The work concerning the application of belief propagation and related Boltzmann machine to traffic data is related to some ongoing projects [33, 34]. It is based on mean-field concepts in physics and basically related to the linear response theory. When combined with machine learning techniques like the automatic segmentation methods it can lead to efficient models able to cope with real-time constraints on large scale networks. We advocate for an Ising model for traffic statistical modelling and propose a proper way for defining traffic indexes which could be also useful for traffic management systems. Still a natural concurrent approach not exposed here can be build analogously using a multivariate model, for which Gaussian belief propagation would apply. More real data will help to decipher from these two possibilities. The main hypothesis underlying our Ising based approach assumes that traffic congestion is well represented by multiple distant pattern superposition. This needs validation with real data on networks. Our reconstruction schema seems to work already with simple underlying binary indexes, but more work is needed for the dynamical part to be able to perform prediction.

Acknowledgements: This gives me the occasion to express my warm thanks to my colleagues Jean-Marc Lasgouttes and Victorin Martin with whom it is a pleasure to collaborate on the main subjects discussed in this review. I am also grateful to Anne Auger, Yufei Han, Fabrice Marchal and Fabien Moutarde for many aspects mentioned in this work concerning ongoing projects. This work was supported by the grant ANR-08-SYSC-017 from the French National Research Agency.

References

1. Amit, D. J., Gutfreund, H., and Sompolinsky, H. Statistical mechanics of neural networks near saturation. *Annals of Physics 173*, 1 (1987), 30–67.
2. Chau Nguyen, H., and Berg, J. Bethe-peierls approximation and the inverse ising model. *ArXiv e-prints*, 1112.3501 (2011).
3. Cocco, S., and Monasson, R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. arXiv:1110.5416, 2011.
4. 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.
5. 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.
6. Frey, B., and Dueck, D. Clustering by passing messages between data points. *Science 315* (2007), 972–976.
7. 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.
8. 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.
9. 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.
10. 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.
11. Han, Y., and Moutarde, F. Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization. In *Proc. of ITSC* (2011).
12. Hansen, N., and Ostermeier, A. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation 9*, 2 (2001), 159–195.
13. Heskes, T. On the uniqueness of loopy belief propagation fixed points. *Neural Computation 16* (2004), 2379–2413.
14. Hopfield, J. J. Neural network and physical systems with emergent collective computational abilities. *Proc. of Natl. Acad. Sci. USA 79* (1982), 2554–2558.
15. Jaynes, E. T. *Probability Theory: The Logic of Science (Vol 1)*. Cambridge University Press, 2003.
16. Kabashima, Y., and Saad, D. Belief propagation vs. tap for decoding corrupted messages. *Europhys. Lett. 44* (1998), 668.
17. Kappen, H., and RodrÃ\DGuez, F. Efficient learning in boltzmann machines using linear response theory. *Neural Computation 10*, 5 (1998), 1137–1156.

18. Kschischang, F. R., Frey, B. J., and Loeliger, H. A. Factor graphs and the sum-product algorithm. *IEEE Trans. on Inf. Th. 47*, 2 (2001), 498–519.
19. Martin, V., Lasgouttes, J., and Furtlehner, C. Encoding dependencies between real-valued observables with a binary latent MRF. to be published, 2011.
20. Mezard, M., and Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris 103*, 1-2 (2009), 107 – 113.
21. Mézard, M., Parisi, G., and Virasoro, M. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.
22. Mézard, M., and Zecchina, R. The random K-satisfiability problem: from an analytic solution to an efficient algorithm. *Phys.Rev.E 66* (2002), 56126.
23. Minka, T. Expectation propagation for approximate bayesian inference. In *Proceedings UAI* (2001), pp. 362–369.
24. Mooij, J. M., and Kappen, H. J. On the properties of the Bethe approximation and loopy belief propagation on binary network. *J. Stat. Mech.* (2005), P11012.
25. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann, 1988.
26. 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.
27. Wainwright, M. J. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, MIT, Jan. 2002.
28. Watanabe, Y., and Fukumizu, K. Graph zeta function in the bethe free energy and loopy belief propagation. In *Advances in Neural Information Processing Systems* (2009), vol. 22, pp. 2017–2025.
29. Weiss, Y., and Freeman, W. T. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Comput. 13*, 10 (2001), 2173–2200.
30. Welling, M., and Y.W., T. Approximate inference in boltzmann machines. *Artif. Intell. 143*, 1 (2003), 19–50.
31. Yasuda, M., and Tanaka, K. Approximate learning algorithm in boltzmann machines. *Neural Comput. 21* (2009), 3130–3178.
32. Yedidia, J. S., Freeman, W. T., and Weiss, Y. Generalized belief propagation. *Advances in Neural Information Processing Systems* (2001), 689–695.
33. TRAVESTI project, (2009-2012). http://travesti.gforge.inria.fr/.
34. PUMAS project, (2010-2013). http://pumas.inria.fr/public/document.