

GMRF Estimation under Topological and Spectral Constraints

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

▶ **To cite this version:**

Victorin Martin, Cyril Furtlehner, Yufei Han, Jean-Marc Lasgouttes. GMRF Estimation under Topological and Spectral Constraints. 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Sep 2014, Nancy, France. pp.370-385, 10.1007/978-3-662-44851-9_24 . hal-01065607

HAL Id: hal-01065607
https://hal.science/hal-01065607

Submitted on 18 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GMRF Estimation under Topological and Spectral Constraints

Victorin Martin[1], Cyril Furtlehner[2], Yufei Han[2,3], and Jean-Marc Lasgouttes[3]

[1] Mines ParisTech, Centre for Robotics, France
[2] Inria Saclay–Île-de-France, TAO team, France
[3] Inria Paris–Rocquencourt, RITS team, France

Abstract. We investigate the problem of Gaussian Markov random field selection under a non-analytic constraint: the estimated models must be compatible with a fast inference algorithm, namely the Gaussian belief propagation algorithm. To address this question, we introduce the $\star$-IPS framework, based on iterative proportional scaling, which incrementally selects candidate links in a greedy manner. Besides its intrinsic sparsity-inducing ability, this algorithm is flexible enough to incorporate various spectral constraints, like e.g. walk summability, and topological constraints, like short loops avoidance. Experimental tests on various datasets, including traffic data from San Francisco Bay Area, indicate that this approach can deliver, with reasonable computational cost, a broad range of efficient inference models, which are not accessible through penalization with traditional sparsity-inducing norms.

Keywords: Iterative proportional scaling, Gaussian belief propagation, walk-summability, Gaussian Markov Random Field.

1 Introduction

The Gaussian belief propagation algorithm [2] (GaBP) is an efficient distributed inference algorithm, well adapted to online inference on large scale Gaussian Markov random fields (GMRF). However, since it may encounter convergence problems, especially with non-sparse structures, it can be of practical interest to construct off-line a GMRF which is compatible with GaBP. When selecting such a model from observations, we potentially face a difficult constrained problem. In the present work, we propose to solve it in an approximate but satisfactory manner, with good accuracy and limited computational cost. To achieve this, we combine various methods, which have been discussed in the context of sparse inverse covariance matrix estimation [1, 7, 15].

The GMRF distribution is naturally characterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}$ and a positive definite precision (or concentration) matrix $\mathbf{A}$, which is simply the inverse of the covariance matrix $\mathbf{C}$. Zero entries in the precision matrix $\mathbf{A}$ indicate conditionally independent pairs of variables. This gives a graphical representation of dependencies: two random variables are conditionally independent if, and only if, there is no direct edge between them. Observations are

summarized in an empirical covariance matrix $\hat{\mathbf{C}} \in \mathbb{R}^{N\times N}$ of a random vector $\mathbf{X} = (X_i)_{i\in\{1,\dots,N\}}$, and we look for a GMRF model with sparse precision matrix $\mathbf{A}$. The model estimation problem can be expressed as the maximization of the log-likelihood:

$$\mathbf{A} = \underset{\mathbf{M}\in\mathcal{S}_{++}^{\text{BP}}}{\operatorname{argmax}}\ \mathcal{L}(\mathbf{M}), \quad \mathcal{L}(\mathbf{M}) \stackrel{\text{def}}{=} \log\det(\mathbf{M}) - \operatorname{Tr}(\mathbf{M}\hat{\mathbf{C}}),$$

where $\mathcal{S}_{++}^{\text{BP}}$ formally represents the set of positive definite matrices corresponding to some GaBP-compatible GMRF.

Without any constraint on $\mathbf{M}$, the maximum likelihood estimate is trivially $\mathbf{A} = \hat{\mathbf{C}}^{-1}$. However, enforcing sparsity with simple thresholding of small magnitude entries may easily ruin the positive definiteness of the estimated precision matrix. In the context of structure learning, where meaningful interactions have to be determined, for instance among genes in genetic networks, the maximization is classically performed on the set of positive definite matrices, after adding to the log-likelihood a continuous penalty function P that imitates the L_0 norm. The Lasso penalty, a convex relaxation of the problem, uses the L_1 norm, measuring the amplitudes of off-diagonal entries in $\mathbf{A}$ [7, 9]. Various optimization schemes have been proposed to solve it efficiently [1, 7]. However, the L_1 norm penalty suffers from a modeling bias, due to excessive penalization of truly large magnitudes entries of $\mathbf{A}$. To overcome this issue, concave functions, that perform constant penalization to the large magnitudes, have been proposed. Experimental results indicate promising improvements compared to Lasso penalty by reducing bias, while conserving the sparsity-introducing capability [6, 11].

In our context, where compatibility with GaBP has to be imposed, sparsity is a desirable feature, albeit without much guarantee: specific topological properties, like the presence of short loops, are likely to damage the GaBP compatibility, even on a sparse graph. Some spectral properties, e.g. walk-summability [12], which guarantee the compatibility with GaBP based inference, might be relevant too. In order to incorporate these explicitly, we propose an efficient constrained model selection framework called $\star$-IPS, where $\star$ stands for the imposed constraints. Approaches based on the iterative proportional scaling (IPS) procedure [17] have already been discussed for tackling the original sparse inverse covariance matrix problem [10, 15]. A first contribution of this paper is to improve its performance by combining it with block update techniques used in [1, 7], along with providing some precision guarantee based on duality. Our second and main contribution is to exploit the incremental nature of the method to impose, for a reasonable cost, both topological and/or spectral constraints, to generate GMRF models compatible with GaBP, achieving a very good trade-off between computational cost and precision in inference tasks, as shown experimentally.

The paper is organized as follows. The principles of IPS are described in Section 2. Our method includes a likelihood maximization step at fixed graph structure, for which we give a stopping criterion based on duality. In Section 3, we propose several constraints improving GaBP compatibility of the estimated models, and show how to introduce them in our framework. In Section 4, we

describe ⋆-IPS as a whole, discuss its complexity and provide some implementation details. Finally Section 5 is devoted to numerical experiments, both on synthetic data and on real traffic data coming from $\approx 10^3$ fixed sensors in the San Francisco Bay Area, to illustrate the use of the method for traffic applications.

2 IPS-Based GMRF Selection

Iterative proportional scaling has been proposed for contingency table estimation [4] and extended further to MRF maximum likelihood estimation [17]. Assuming the structure of the graph is known, it appears to be less efficient than other gradient based methods [13]. Conversely, local changes based on single row-column updates have been shown to be very efficient, even in the first order setting [7]. In our work, we combine the benefits of the incremental characteristics of IPS to identify links (Section 2.1), with the efficiency of row-column update to optimize their parameters at fixed structure (Section 2.2).

2.1 Optimal 1-Link Perturbation

Suppose that we are given a set of single and pairwise empirical marginals $\hat{p}_i$ and $\hat{p}_{ij}$ from a real-valued random vector $\mathbf{X} = (X_i)_{i\in\{1\dots N\}}$, and a candidate distribution $\mathcal{P}^{(n)}$, based on the dependency graph $\mathcal{G}^{(n)}$. Let us first describe optimal link addition in terms of likelihood. Let $\mathcal{P}^{(n)}$ be the reference distribution, to which we want to add one factor ψ_{ij} to produce the distribution

$$\mathcal{P}^{(n+1)}(\mathbf{x}) = \mathcal{P}^{(n)}(\mathbf{x}) \times \psi_{ij}(x_i, x_j).$$

This is a special case of IPS and the optimal perturbation is

$$\psi_{ij}(x_i, x_j) = \frac{\hat{p}_{ij}(x_i, x_j)}{p^{(n)}_{ij}(x_i, x_j)}, \tag{1}$$

where $p^{(n)}_{ij}$ is the (i, j) pairwise marginal of $\mathcal{P}^{(n)}$. The correction to the log-likelihood can then be written as a Kullback-Leibler divergence:

$$\Delta\mathcal{L} = D_{KL}(\hat{p}_{ij} \| p^{(n)}_{ij}) = \iint \hat{p}_{ij}(u, v) \log \frac{\hat{p}_{ij}(u, v)}{p^{(n)}_{ij}(u, v)} \mathrm{d}u\mathrm{d}v.$$

Sorting all the links w.r.t. this quantity yields the optimal 1-link correction to be made. Hence, the best candidate is the one for which the current model yields the joint marginal $p^{(n)}_{ij}$ that is most divergent from $\hat{p}_{ij}$. Note that the update mechanism can in fact also be applied if the link is already present.

In the general case, computing the pairwise marginals $\{p_{ij}, (ij) \notin \mathcal{G}^{(n)}\}$ is expensive. However, in the GMRF family, these marginals depend only on the covariance matrix associated to $\mathcal{P}^{(n)}$. The correction factor (1) reads in that case

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

where $\mathbf{C}_{\{ij\}}$ (resp. $\hat{\mathbf{C}}_{\{ij\}}$) represents the restricted 2×2 covariance matrix corresponding to the pair (X_i, X_j) of the current model $\mathcal{P}^{(n)}$ (resp. of the empirical distribution $\hat{\mathcal{P}}$) specified by precision matrix $\mathbf{A} = \mathbf{C}^{-1}$ (resp. $\hat{\mathbf{A}} = \hat{\mathbf{C}}^{-1}$). Let $[\mathbf{C}_{\{ij\}}]$ denote the $N \times N$ matrix formed by completing $\mathbf{C}_{\{ij\}}$ with zeros. The new model obtained after adding or changing link (i, j) reads

$$\mathbf{A}' = \mathbf{A} + [\hat{\mathbf{C}}^{-1}_{\{ij\}}] - [\mathbf{C}^{-1}_{\{ij\}}] \stackrel{\text{def}}{=} \mathbf{A} + [\mathbf{V}], \tag{2}$$

with a log-likelihood variation given by:

$$\Delta\mathcal{L} = \frac{C_{ii}\hat{C}_{jj} + C_{jj}\hat{C}_{ii} - 2C_{ij}\hat{C}_{ij}}{\det(\mathbf{C}_{\{ij\}})} - 2 - \log\frac{\det(\hat{\mathbf{C}}_{\{ij\}})}{\det(\mathbf{C}_{\{ij\}})}. \tag{3}$$

For a 2×2 perturbation matrix $\mathbf{V} = \mathbf{V}_{\{ij\}}$, the Sherman–Morrison–Woodbury (SMW) formula allows us to efficiently compute the new covariance matrix as

$$\mathbf{C}' = \mathbf{A}'^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}[\mathbf{C}^{-1}_{\{ij\}}]\big(\mathbf{I} - [\hat{\mathbf{C}}_{\{ij\}}][\mathbf{C}^{-1}_{\{ij\}}]\big)\mathbf{A}^{-1}. \tag{4}$$

The number of operations needed to maintain the covariance matrix – and to keep track of all pairwise marginals – after each addition is therefore $\mathcal{O}(N^2)$. This technical point is determinant to the usefulness of our approach. The identity $\det(A') = \det(A) \times \det(\hat{\mathbf{C}}_{\{ij\}})/\det(\mathbf{C}_{\{ij\}})$ ensures that the new precision matrix remains definite positive when both $\mathbf{C}_{\{ij\}}$ and $\hat{\mathbf{C}}_{\{ij\}}$ are non-degenerate.

It is also possible to remove links, so that, with help of a penalty coefficient per link, the model can be optimized with a desired connectivity level. For a GMRF with precision matrix $\mathbf{A}$, removing the link (i, j) amounts to setting the entry A_{ij} to zero, and thus $\psi_{ij}(x_i, x_j) = \exp\big(A_{ij}x_ix_j\big)$. The corresponding change of log-likelihood is then

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

and, using again the SMW formula, we get the new covariance matrix

$$\mathbf{C}' = \mathbf{C} - \frac{A_{ij}}{1 - 2A_{ij}C_{ij} - A_{ij}^2\det(\mathbf{C}_{\{ij\}})}\,\mathbf{C}[\mathbf{B}_{\{ij\}}]\mathbf{C},$$

with

$$\mathbf{B}_{\{ij\}} \stackrel{\text{def}}{=} \begin{bmatrix} A_{ij}C_{jj} & 1 - A_{ij}C_{ij} \\ 1 - A_{ij}C_{ij} & A_{ij}C_{ii} \end{bmatrix}.$$

In this case, the positive-definiteness of $\mathbf{A}'$ needs to be checked and we have

$$\det(\mathbf{A}') = \det(\mathbf{A})\Big[1 - \alpha\big(C_{ij} - \sqrt{C_{ii}C_{jj}}\big)\Big]\Big[1 - \alpha\big(C_{ij} + \sqrt{C_{ii}C_{jj}}\big)\Big],$$

so that $\mathbf{A}'$ is definite positive if the following condition is verified:

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

2.2 Block Updates

When a new link is added, existing links become detuned by a slight amount. As pointed out, the optimal update given in Section 2.1 is actually indifferent to whether the considered link exists or not. This means that, after a while, detuned links may be automatically updated if the likelihood gain exceeds the one obtained by adding a new link. We observe in practice that, when many links have been added, all the existing links are slightly detuned, which eventually causes suboptimal or bad decisions for the next links, resulting in a significant departure of the learning curve from the optimal one (see Fig. 2-left in Section 5). However, correcting existing links can become very time consuming, the update of one single link having the same computational cost $\mathcal{O}(N^2)$ as the addition of one link. There are various options to address this problem. To keep the algorithm fast, robust and simple, we choose to stay with the logic of coordinate descent, by remarking that local updates are still possible via a single row-column update of the precision matrix, as originally proposed in [1] and refined in [7]. In our context, the method is based on the following expression of the log determinant of the precision matrix $\mathbf{A}$:

$$\log\det(\mathbf{A}) = \log\det(\mathbf{A}_{\setminus i\setminus i}) + \log(\mathbf{A}_{ii} - \mathbf{A}_i^T \mathbf{A}_{\setminus i\setminus i}^{-1} \mathbf{A}_i),$$

where $\mathbf{A}_{\setminus i\setminus i}$ is the block matrix obtained after taking aside the i^{th} row and column and $\mathbf{A}_i$ is the i^{th} column vector of $\mathbf{A}$ without A_{ii}. The direct optimization of the log-likelihood w.r.t. $\mathbf{A}_i$ and A_{ii} yields the following updated values:

$$A'_{ii} = \frac{1}{\hat{C}_{ii}} + \mathbf{A}_i^T \mathbf{A}_{\setminus i\setminus i}^{-1} \mathbf{A}_i \quad \text{and} \quad \mathbf{A}'_i = \left[\mathbf{I}_{V(i)} \mathbf{A}_{\setminus i\setminus i}^{-1} \mathbf{I}_{V(i)}\right]^{-1} \mathbf{I}_{V(i)} \frac{\hat{\mathbf{C}}_i}{\hat{C}_{ii}}, \qquad (5)$$

where $\hat{\mathbf{C}}_i$ represents the i^{th} column vector of $\hat{\mathbf{C}}$, $V(i)$ the set of neighbors of i in the current graph, and $\mathbf{I}_{V(i)}$ the identity restricted to entries $j \in \{i\} \cup V(i)$. Note that this solution involve the inverse $\mathbf{A}_{\setminus i\setminus i}^{-1}$ of a matrix of size $N-1$. It is related to $\mathbf{C} = \mathbf{A}^{-1}$ as follows:

$$\mathbf{A}_{\setminus i\setminus i}^{-1} = \mathbf{C}_{\setminus i\setminus i} - \frac{\mathbf{C}_i \mathbf{C}_i^T}{C_{ii}}.$$

The overall cost for updating column (and row) i is thus $\mathcal{O}(|V(i)|^3)$ for the inversion of $\left[\mathbf{I}_{V(i)} \mathbf{A}_{\setminus i\setminus i} \mathbf{I}_{V(i)}\right]$ and $\mathcal{O}(N^2)$ to update the covariance matrix $\mathbf{C}$ after this change. The log-likelihood gain $\Delta\mathcal{L}$ reads

$$-\log\left[\hat{C}_{ii}(A_{ii} - \mathbf{A}_i^T \mathbf{A}_{\setminus i\setminus i}^{-1} \mathbf{A}_i)\right] - 2(\mathbf{A}'_i - \mathbf{A}_i)^T \hat{\mathbf{C}}_i - \left[\frac{1}{\hat{C}_{ii}} + \mathbf{A}_i'^T \mathbf{A}_{\setminus i\setminus i}^{-1} \mathbf{A}'_i - A_{ii}\right] \hat{C}_{ii}.$$

2.3 Stopping Criterion

If the set of links to be optimized is given by some graph $\mathcal{G}$, the likelihood optimization is a convex problem. Let us investigate its dual properties. Let $\mathbf{A}$

denote the precision matrix, and $\mathbf{\Pi}$ a Lagrange matrix multiplier, that imposes the structure given by $\mathcal{G}$. The support of $\mathbf{\Pi}$ is the complementary graph of $\mathcal{G}$: $\forall (i,j) \in \mathcal{G}, \mathit{\Pi}_{ij} = 0$, $\forall i, \mathit{\Pi}_{ii} = 0$ and $\mathbf{\Pi}$ is symmetric. Then, given $\mathbf{\Pi}$, we want to optimize

$$\mathbf{A}_{\mathbf{\Pi}} = \underset{\mathbf{M}}{\operatorname{argmin}}\ \operatorname{Tr}(\mathbf{M}\mathbf{\Pi}) + f(\mathbf{M}),$$

with $f(\mathbf{M}) \stackrel{\text{def}}{=} \operatorname{Tr}(\mathbf{M}\hat{\mathbf{C}}) - \log\det(\mathbf{M})$ being convex for given support $\mathcal{G}$. The explicit solution is

$$\mathbf{A}_{\mathbf{\Pi}} = (\hat{\mathbf{C}} + \mathbf{\Pi})^{-1}. \tag{6}$$

We assume that $\mathbf{\Pi}$ is such that $\hat{\mathbf{C}}+\mathbf{\Pi}$ is positive definite, so the dual optimization problem reads $\mathbf{Y} = \operatorname{argmax}_{\mathbf{\Pi}}\ g(\mathbf{\Pi})$, with $g(\mathbf{\Pi}) \stackrel{\text{def}}{=} N + \log\det(\hat{\mathbf{C}} + \mathbf{\Pi})$. The problem is now concave and, because of the barrier resulting from the log term, we are certain to have a positive definite solution. Thus, for any matrix $\mathbf{\Pi}$, such that $\hat{\mathbf{C}} + \mathbf{\Pi}$ is definite positive, $g(\mathbf{\Pi})$ is a lower bound of the log-likelihood. The support of $\mathbf{\Pi}$ represents the set of links to be removed from the precision matrix. Once optimality is reached for $\mathbf{\Pi}$, all non-zero entries $\mathit{\Pi}_{ij}$ correspond to vanishing coefficients A_{ij} in (6). We may proceed as before, by computing the potential log-likelihood gain $\mathit{\Delta}\mathcal{L}$ for such local transformations of the covariance matrix. Local moves in the dual formulation deal with the covariance matrix instead of the precision matrix in the primal one. Let $\mathbf{C}$ and $\mathbf{C}'$ be two covariance matrices differing by a single modification on $\mathbf{\Pi}$ with $\mathbf{A} = \mathbf{C}^{-1}$ and $\mathbf{A}' = \mathbf{C}'^{-1}$. We have

$$\det(\mathbf{C}') = \det(\mathbf{C})\big(1 + 2\mathit{\Pi}_{ij}A_{ij} - \mathit{\Pi}_{ij}^2 \det(\mathbf{A}_{\{ij\}})\big),$$

with $\det(\mathbf{A}_{\{ij\}}) > 0$, since $\mathbf{A}$ is definite positive. Maximizing the log-likelihood variation yields the optimal values

$$\mathit{\Pi}_{ij} = \frac{A_{ij}}{\det(\mathbf{A}_{\{ij\}})} \quad \text{and} \quad \mathit{\Delta}\mathcal{L} = \log\left(1 + \frac{A_{ij}^2}{\det(\mathbf{A}_{\{ij\}})}\right).$$

In practice, we will not use this backward scheme: its computational cost is prohibitive, since the complementary graph, composed of links to be removed, is dense. However, this dual formulation will help us to build a confidence interval. During the greedy procedure, we always have to maintain $\mathbf{C} = \mathbf{A}^{-1}$ but $\mathbf{C}$ cannot be used directly to get a dual cost because, except at convergence, it does not fulfill the dual constraints $C_{ij} = \hat{C}_{ij}$, $\forall (i,j) \in \mathcal{G}$.

Let $\mathbf{\Pi}^{\|}$ be the correction matrix with coefficients $\mathit{\Pi}_{ij}^{\|} \stackrel{\text{def}}{=} (\hat{C}_{ij} - C_{ij})\mathbb{1}_{\{(i,j)\in G\}}$. Provided that $\tilde{\mathbf{C}} \stackrel{\text{def}}{=} \mathbf{A}^{-1} + \mathbf{\Pi}^{\|}$ is definite positive, which happens when $\mathbf{A}$ is close enough to the optimum $\mathbf{A}^\star$, it satisfies the dual constraints yielding the confidence bound

$$\log\det(\tilde{\mathbf{C}}) + N \leq -\mathcal{L}(\mathbf{A}^\star) \leq \operatorname{Tr}(\mathbf{A}\hat{\mathbf{C}}) - \log\det(\mathbf{A}).$$

We have

$$\log\det(\tilde{\mathbf{C}}) = -\log\det(\mathbf{A}) + \log\det\left(\mathbf{I} + \mathbf{A}\mathbf{\Pi}^{\|}\right),$$

with both $\mathbf{A}$ and $\mathbf{\Pi}^{\|}$ sparse matrices, so the determinant can be estimated in $\mathcal{O}(N^2K)$ operations by expanding the logarithm at order 2 in $\mathbf{A}\mathbf{\Pi}^{\|}$. It leads to the following bound

$$\Delta\mathcal{L} \leq \frac{1}{2}\operatorname{Tr}\big(\mathbf{A}\mathbf{\Pi}^{\|}\mathbf{A}\mathbf{\Pi}^{\|}\big),$$

which will be used in practice as a stopping criterion for the link updates.

3 Introducing Constraints for GaBP Compatibility

Usually, GMRF estimation intends to describe a dependency structure. We pursue here another aim: finding a model suitable for fast inference. While inference in GMRF models can always be performed exactly in $\mathcal{O}(N^3)$ through matrix inversion, this may not be fast enough for some "real-time" applications on large networks. The GaBP algorithm [2] is a fast alternative to matrix inversion for sparse GMRF, which uses message passing along links in the graph $\mathcal{G}$, and thus, assuming it converges, will perform the inference in $\mathcal{O}(mKN)$, where K is the mean connectivity of $\mathcal{G}$ and m the maximum number of iterations before convergence. An important property of the GaBP algorithm is that, whenever it converges, it provides the exact mean values for all variables [18]. Variances are however generally incorrect [12]. Having a sparse GMRF gives no guarantee about its compatibility with GaBP, so we need to impose more precise constraints on the precision matrix and to the graph structure. In this section, we make such constraints explicit and show how to impose them in the framework of Section 2.

3.1 Spectral Constraints

The most precise condition known for convergence of GaBP is walk-summability (WS) [12]. Let $\mathbf{R}(\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{A} - \mathbf{diag}(\mathbf{A})$ contain the off-diagonal terms of $\mathbf{A}$, and let $\rho(\cdot)$ denote the spectral radius of a matrix, that is, the maximal modulus of its eigenvalues. The two equivalent necessary and sufficient conditions for WS that we will use are:

(i) The matrix $\mathbf{W}(\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{diag}(\mathbf{A}) - |\mathbf{R}(\mathbf{A})|$ is definite positive;
(ii) $\rho(|\mathbf{R}'(\mathbf{A})|) < 1$, with $\mathbf{R}'(\mathbf{A})_{ij} \stackrel{\text{def}}{=} \frac{R(\mathbf{A})_{ij}}{\sqrt{A_{ii}A_{jj}}}$.

Let us consider a GMRF with WS precision matrix $\mathbf{A}$ and investigate under which conditions the model remains WS after a perturbation of a link (i, j). The following proposition gives a sufficient condition:

Proposition 1. *Let* $\mathbf{A}$ *be a WS precision matrix and denote* $\mathbf{W} \stackrel{\text{def}}{=} \mathbf{W}(\mathbf{A})$. *The matrix* $\mathbf{A}' = \mathbf{A} + [\mathbf{V}_{\{ij\}}]$ *is WS if*

$$\Theta(\alpha) > 0, \forall\alpha \in [0, 1], \tag{7}$$

where Θ is the following function

$$\Theta(\alpha) \stackrel{\text{def}}{=} \det(W^{-1}_{\{ij\}}) \left(\alpha^2 V_{ii}V_{jj} - (|A_{ij}| - |\alpha V_{ij} + A_{ij}|)^2\right) + \alpha \left(W^{-1}_{ii}V_{ii} + W^{-1}_{jj}V_{jj}\right) + 2\left(|A_{ij}| - |\alpha V_{ij} + A_{ij}|\right) W_{ij}. \quad (8)$$

In order to check condition (7), it is necessary to solve two quadratic equations. Note however that knowledge of the matrix $\mathbf{W}(\mathbf{A})^{-1}$ is also mandatory. We will discuss this point at the end of this section.

Proof. The sufficient condition is obtained as follows: for $\alpha \in [0, 1]$ we have

$$\mathbf{W}(\mathbf{A} + \alpha\mathbf{V}) = \mathbf{W}(\mathbf{A}) + [\boldsymbol{\phi}(\alpha\mathbf{V}, \mathbf{A})],$$

with

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

$\mathbf{W}(\mathbf{A})$ being invertible, the determinant of $\mathbf{W}(\mathbf{A} + \alpha\mathbf{V})$ is expressed as

$$\det(\mathbf{W}(\mathbf{A} + \alpha\mathbf{V})) = \det\left(\mathbf{W}(\mathbf{A})\right) \det\left(\mathbf{I} + \mathbf{W}(\mathbf{A})^{-1}[\boldsymbol{\phi}(\alpha\mathbf{V}, \mathbf{A})]\right),$$

and we can check that $\Theta(\alpha) = \det(\mathbf{I} + \mathbf{W}(\mathbf{A})^{-1}[\boldsymbol{\phi}(\alpha\mathbf{V}, \mathbf{A})])$, with Θ defined in (8). The spectrum of $\mathbf{W}(\mathbf{A} + \alpha\mathbf{V})$ being a continuous function of α, as the roots of a polynomial, the condition (7) follows.

Note that the special case of removing one link of the graph always preserves the WS property. Indeed, it will change the matrix $\mathbf{A}$ in $\mathbf{A}'$ such as $|\mathbf{R}'(\mathbf{A}')| \leq |\mathbf{R}'(\mathbf{A})|$ where $\leq$ denotes element-wise comparison. Then, using elementary results on positive matrices [16, p. 22], $\rho(|\mathbf{R}'(\mathbf{A}')|) \leq \rho(|\mathbf{R}'(\mathbf{A})|)$ and thus $\mathbf{A}'$ is WS whenever $\mathbf{A}$ is WS.

As we shall see in the numerical experiments, imposing WS is generally too restrictive. It is easy to find non WS models which are still GaBP compatible. The above principle allows us however to impose a weaker spectral constraint: imposing that the matrix $\mathbf{diag}(\mathbf{A}) - \mathbf{R}(\mathbf{A})$ remains definite positive. This is equivalent to constrain the spectral radius $\rho(\mathbf{R}'(\mathbf{A}))$ to be strictly lower than 1 and it is a necessary condition for GaBP convergence [12]. We call that condition "weak walk-summability" (WWS) as a relaxation of the WS condition. We obtain the following condition

Proposition 2. *Let $\mathbf{A}$ be a WWS precision matrix, i.e. such as $\rho(\mathbf{R}'(\mathbf{A})) < 1$, and $\mathbf{S}(\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{diag}(\mathbf{A}) - \mathbf{R}(\mathbf{A})$. The matrix $\mathbf{A}' = \mathbf{A} + [\mathbf{V}_{ij}]$ is WWS if*

$$\Gamma(\alpha) > 0, \forall\alpha \in [0, 1], \quad (9)$$

with Γ the following degree 2 polynomial and $\mathbf{S} \stackrel{\text{def}}{=} \mathbf{S}(\mathbf{A})$

$$\Gamma(\alpha) \stackrel{\text{def}}{=} \alpha^2 \det(\mathbf{V}\mathbf{S}^{-1}_{\{i,j\}}) + \alpha(V_{ii}S^{-1}_{ii} + V_{jj}S^{-1}_{jj} - 2V_{ij}S^{-1}_{ij}) + 1.$$

In order to check condition (9), we have to solve a quadratic equation. As in Proposition 1, we need to keep track of an inverse matrix, in this case $\mathbf{S}(\mathbf{A})^{-1}$.

Proof. Mimicking the proof of Proposition 2, we define, for $\alpha \in [0,1]$,

$$\mathbf{M}(\alpha) \stackrel{\text{def}}{=} \mathbf{diag}(\mathbf{A} + \alpha[\mathbf{V}]) - \mathbf{R}(\mathbf{A} + \alpha[\mathbf{V}]) = \mathbf{S}(\mathbf{A}) + \alpha\ \mathbf{diag}([\mathbf{V}]) - \mathbf{R}([\mathbf{V}]),$$

and we have

$$\det(\mathbf{M}(\alpha)) = \det(\mathbf{S}(\mathbf{A}))\det\left(\mathbf{I} + \alpha\mathbf{S}^{-1}\left[\mathbf{diag}(\mathbf{V}) - \mathbf{R}(\mathbf{V})\right]\right) = \det(\mathbf{S}(\mathbf{A}))\Gamma(\alpha).$$

The spectrum of $\mathbf{M}(\alpha)$ being a continuous function of α, condition (9) follows.

Both (7) and (9) are only sufficient conditions for spectral constraint conservation after a pairwise perturbation. However, there are only a few cases where they lead to rejection of a valid perturbation. Indeed, it means that at least one eigenvalue goes to zero for some $\alpha \in]0,1[$ and is positive again for $\alpha = 1$.

We have pointed out that checking sufficient condition (7) (resp. (9)) imposes to keep track of the inverse matrix $\mathbf{W}(\mathbf{A})^{-1}$ (resp. $\mathbf{S}(\mathbf{A})^{-1}$). This will not impact the overall complexity of the algorithm since, using the SMW formula, it can be done in $\mathcal{O}(N^2)$ operations, like for keeping track of the covariance matrix.

Note that, if we want to maintain these spectral constraints, we will not be able in practice to use the column updates described in Section 2.2. Indeed, computing the optimal perturbation is in this case costly, and we have no easy way to check whether it leads to a new admissible model, since our method would involve higher order characteristic polynomials.

3.2 Topological Constraints

We present in this section another approach, based mainly on empirical knowledge about the belief propagation algorithm. Belief propagation has been designed as an exact procedure on trees [14] and short loops are usually believed to cause convergence troubles. In the extreme case, where we forbid the addition of any loop, the best precision matrix estimate based on likelihood is known to be the max-spanning tree w.r.t. mutual information [3]. Since this is usually not enough, we propose here that the estimated precision matrix contains no loops of size smaller than ℓ. This is quite easy to impose: when adding a link (i,j), we have to search if i is in the neighborhood of j of depth $\ell - 1$. The computational cost is $\mathcal{O}(K^\ell)$, with K the connectivity of $\mathcal{G}$.

We can impose a more precise condition using the fact that, in the absence of frustrated loops, the GaBP algorithm is always convergent [12]. A frustrated loop is a loop along which the product of partial correlations is negative. Preventing the formation of frustrated loops is very similar to the previous loop constraint; the search cost is the same, the only difference is that we will avoid only this kind of loops. This last constraint cannot be imposed with guarantees during the block updates since the sign of partial correlations along edges may change. Prohibiting frustrated loops would require to store all the loops in the graph,

which is by far too costly. However, experimental results show that a change of sign usually corresponds to small partial correlations, which are less likely to cause convergence issues.

4 Algorithm Description and Complexity

In this section, we give an overview of $\star$-IPS, leaving aside the backtracking option. A formal implementation[4] is given in Algorithm 1. Note that we suppose that the initial point of the algorithm corresponds to an empty graph. We may as well start from any precision matrix $\mathbf{A}$, provided that we have computed $\mathbf{C} = \mathbf{A}^{-1}$ and $\mathbf{W}(\mathbf{A})^{-1}$ or $\mathbf{S}(\mathbf{A})^{-1}$ if we want to impose spectral constraints.

Algorithm 1 $\star$-constrained Iterative Proportional Scaling ($\star$-IPS). The function *Check_constraint* returns true if the perturbation of $\widehat{ij}$ leads to a model compatible with the given constraint $\star$.

Inputs: $\star$: constraint type ($\varnothing$, WS, WWS, ℓ-LOOP, ℓ-FLOOP);
$\hat{\mathbf{C}}$: empirical covariance matrix.
Parameters: ϵ and ϵ_u: stopping criteria on the log-likelihood and for the update step;
δK: connectivity increment after which an update step is performed.

$\mathbf{A} = \mathbf{C} = \mathbf{W}^{-1} = \mathbf{S}^{-1} = \mathbf{diag}(\mathbf{1})$.
while $\Delta\mathcal{L}_{\max} > \epsilon$ and $n_{\text{iter}} < n_{\max}$ **do**
  $\Delta\mathcal{L}_{\max} \leftarrow 0$, $n_{\text{iter}} \leftarrow n_{\text{iter}} + 1$
  for all pairs of nodes $\widehat{ij}$ **do**
    compute $\Delta\mathcal{L}^{ij}$ using (3)
    if $\Delta\mathcal{L}^{ij} > \Delta\mathcal{L}_{\max}$ **then**
      if Check_constraint$(\star, \widehat{ij}, \mathbf{A})$ **then**
        $\Delta\mathcal{L}_{\max} \leftarrow \Delta\mathcal{L}^{ij}$ and $\mathbf{V} \leftarrow [\mathbf{V}_{\{i,j\}}]$ defined in (2)
  $\mathbf{A} \leftarrow \mathbf{A} + \mathbf{V}$, update $\mathbf{C}$ using (4).
  if $\star =$ WS or WWS **then**
    update $\mathbf{W}^{-1}$ or $\mathbf{S}^{-1}$ using the SMW formula.
  if connectivity has increased by δK **then**
    while $\mathrm{Tr}\left(\mathbf{A}\mathbf{\Pi}^{\parallel}\mathbf{A}\mathbf{\Pi}^{\parallel}\right) > \epsilon_u$ **do**
      for $\widehat{ij} \mid A_{ij} \neq 0$ **do**
        $\Delta\mathcal{L}_{\max} \leftarrow 0$, compute $\Delta\mathcal{L}^{ij}$ using (3)
        if $\Delta\mathcal{L}^{ij} > \Delta\mathcal{L}_{\max}$ **then**
          if Check_constraint$(\star, \widehat{ij}, \mathbf{A})$ **then**
            $\Delta\mathcal{L}_{\max} \leftarrow \Delta\mathcal{L}^{ij}$ and $\mathbf{V} \leftarrow [\mathbf{V}_{\{i,j\}}]$ defined in (2)
      if $\star =$ WS or WWS **then**
        $\mathbf{A} \leftarrow \mathbf{A} + \mathbf{V}$, update $\mathbf{C}$ and $\mathbf{W}^{-1}$ or $\mathbf{S}^{-1}$ using the SMW formula.
      else
        Block updates (5) of i and j corresponding to $\Delta\mathcal{L}_{\max}$. Update $\mathbf{C}$ using the SMW formula.

[4] The source code is available at `http://www.rocq.inria.fr/~lasgoutt/star-ips/`.

Let us clarify the complexity of this algorithm. Each link addition or update has a cost $\mathcal{O}(N^2)$ to update the covariance matrix. If spectral constraints are imposed, it is necessary to keep track of another inverse matrix, which requires as well $\mathcal{O}(N^2)$ operations and does not change the complexity. Adding M links will therefore require at least $\mathcal{O}(MN^2)$ operations. This means that our algorithm complexity is in $\mathcal{O}(N^3)$ in the sparse regions, whereas it becomes $\mathcal{O}(N^4)$ in the dense ones. Note that this complexity does not take into account the time spent in link updates. As pointed out in Section 2.2, this update step is useful to avoid departing from the optimal learning curve. For a given bound ϵ_u, we observe numerically that the number of updates is $\mathcal{O}(N)$, regardless of the mean connectivity in the sparse regime, so this adds up another $\mathcal{O}(N^3)$ computational cost. Note that the critical parts of the algorithm, which are the update of the covariance matrix and the search for the perturbation that maximizes likelihood increase, can easily be parallelized.

Let us anticipate on the application to emphasize the usefulness of our algorithm, which complexity is comparable to a direct covariance matrix inversion. Suppose that the workflow is:

1. select off-line a GMRF model based on an empirical covariance matrix;
2. use the above model to perform inference for a "real-time" application (which here means at most a few minutes).

We may allow the first task to take a few hours, and thus matrix inversion is acceptable for quite large networks. However, the resultant model will not be suited to GaBP in the inference task and we will have to resort to exact inference, through a matrix inversion of complexity $\mathcal{O}((N-n_o)^3)$, where n_o is the number of observed variables. Using our sparse GaBP-compatible model instead, with a mean connectivity K, the approximate inference complexity $\mathcal{O}(mK(N-n_o))$ will then typically scale down from a few hours to a few seconds or a few minutes depending on the needed precision.

5 Experimental Results

To have some elements of comparison, let us first quickly describe traditional ways to tackle the maximum likelihood estimation problem with penalty-induced sparsity constraints. It involves a maximization of the form

$$\mathbf{A} = \underset{\mathbf{M}\in\mathcal{S}_{++}}{\operatorname{argmax}} \ \log\det(\mathbf{M}) - \operatorname{Tr}(\mathbf{M}\hat{\mathbf{C}}) + \lambda P(\mathbf{M}),$$

where $\mathcal{S}_{++}$ is the set of positive definite matrices. A classical penalization function P is a continuous approximation to the discrete L_0 norm like the "seamless L_0 penalty" (SELO) proposed in [11]

$$P(x) = \log\left(\frac{2|x|+\tau}{|x|+\tau}\right).$$

In the following, we set $\tau = 5.10^{-3}$, which is empirically good enough. We propose to use the Doubly Augmented Lagrange method [5] to solve this penalized log-determinant programming.

The second method used for comparison is QUIC [9], which uses the L_1 norm as penalty. This is a second order optimization method, leading to superlinear convergence rates. We perform it directly on the empirical covariance matrix with different values for the regularization coefficient λ. Once the structure has been found, it is necessary to maximize the likelihood. According to our experiments, L_1 norm penalty leads to poor precision matrices in terms of likelihood, even if it may be very efficient to find an existing sparse graph structure.

We can now compare the performance of $\star$-IPS and sparsity penalized likelihood optimization. For generating one single GMRF with a given designed sparsity level, both methods are comparable in terms of computational cost, while $\star$-IPS is faster in very sparse regime. Due to its incremental nature, it also has the advantage of generating a full Pareto set of approximate solutions. To assess the quality of the $\star$-IPS model selection, we first look into data fitting accuracy through log-likelihood, and then investigate its compatibility with GaBP inference.

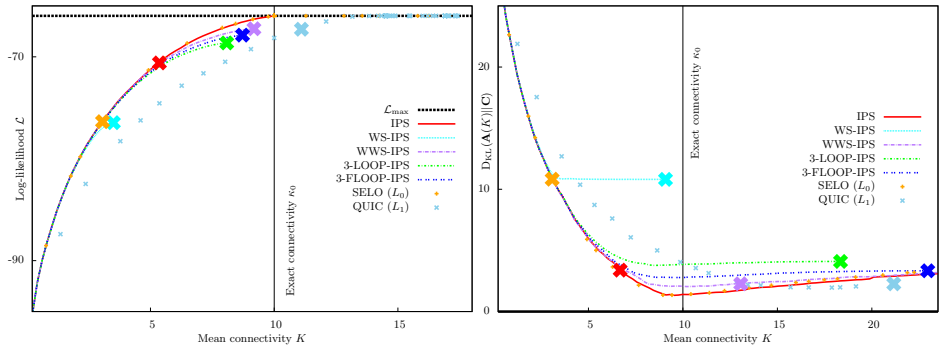


Fig. 1. Left: Log-likelihood $\mathcal{L}$ as a function of mean connectivity K for $\star$-IPS with different constraints, SELO and QUIC, all computed from the exact covariance matrix $\mathbf{C}$. Right: Kullback-Leibler divergence to the actual distribution as a function of mean connectivity (estimations based on an empirical covariance matrix $\hat{\mathbf{C}}$, generated with 1000 samples). The end of GaBP compatibility for each algorithm is indicated by $\times$'s.

Likelihood and GaBP Compatibility Trade-off. The first test is performed on a randomly generated GMRF of 100 variables. The structure of its precision matrix is an Erdős-Rényi random graph, where each link is assigned a value with random sign and magnitude (between 0.1 and 0.8). A diagonal term is added to make it definite positive. The results of the different algorithms are shown in Fig. 1. Both SELO and IPS algorithm are able to find the true graph with the exact covariance matrix. As expected, WS is a very strict constraint and yields low

connectivity models. Using WWS instead yields better GaBP compatibility, but provides no guarantee about it. However, this constraint can be enough to get a GaBP-compatible model with (almost) maximal likelihood (Fig. 1-right). For both models, QUIC is clearly sub-optimal regarding sparsity/likelihood trade-off. Fig. 1-right illustrates $\star$-IPS performance in terms of overfitting. This overfitting starts after the Kullback-Leibler divergence reaches a minimum. This can happen as well as before as after the end of GaBP compatibility. Detecting this point is a classical but difficult statistical problem and further investigations are needed to find the best criterion adapted to our case. The second test (Fig. 2) is performed

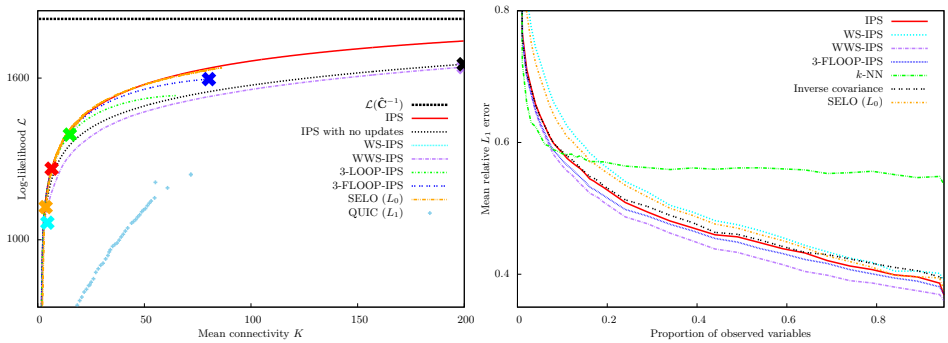


Fig. 2. Left: log-likelihood $\mathcal{L}$ as a function of mean connectivity K. Right: mean relative L_1 reconstruction error as a function of the fraction of observed stations on the San Francisco Bay Area network for various methods; results are averaged over 100 sample test experiments and normalized by the score obtained with daytime moving average predictor.

on traffic data from the San Francisco Bay Area[5]. Each sample data is a N-dimensional vector of observed speeds $\{\hat{V}_i, i = 1 \dots N\}$, giving a snapshot of the network at a given time of the day, as measured by a set of fixed sensors. After filtering out inactive sensors, we finally kept 1020 variables, for which we had data from January to June 2013. The travel time distribution at each link, being bounded with heavy tail, is far from being Gaussian. In order to work with normal variables, we make the following transformation

$$Y_i = \Phi^{-1}\Big(\frac{1 + \hat{F}_i(V_i)}{2}\Big), \quad \forall i = 1 \dots N, \tag{10}$$

which maps the speed V_i to the positive domain of a standard Gaussian variable Y_i, where $\hat{F}_i$ and Φ are respectively the cumulative distribution functions of the speeds and of a normal distribution. The input of the algorithms we are comparing is the covariance matrix of the vector $\mathbf{Y}$. This mapping is important to use the selected GMRF for the inference tasks in the next section.

[5] Available at the Caltrans PeMS website: `http://pems.dot.ca.gov/`.

Fig. 2-left compares the performance of some $\star$-IPS variants with methods based on penalized norms. IPS with update but no constraint is comparable to SELO optimization, albeit much faster to generate the Pareto set, while QUIC is by far the weakest contender. In fact, QUIC is not performing well because, in this case, there is no true underlying sparse dependency graph. Both IPS and SELO loose the compatibility with GaBP at low likelihood and mean connectivity (respectively at $K < 6.5$ and $K < 4$). By contrast, imposing the "no frustrated loops of size 3" constraint (3-FLOOP-IPS) yields a nearly optimal $\mathcal{L}(K)$ path, up to some flat regime, which endpoint is still GaBP compatible. This is the best trade-off which can be found among all constraints that we have tested. While the WS constraint is again too restrictive, we notice that WWS yields models which are all the way compatible, but with a suboptimal $\mathcal{L}(K)$ path. This is partly due to the absence of block updates, replaced by less efficient local updates. Actually, we see that the WWS $\mathcal{L}(K)$ roughly follows the one obtained with $\varnothing$-IPS (no update at all), which also delivers only GaBP-compatible models. Our interpretation is that updating the links has the effect of reducing some uncorrelated noise otherwise present in the approximate model. At some point, it may spoil the GaBP compatibility because of stronger correlations being taken into account.

Inverse Models for GaBP Inference. Our original motivation for this work is to provide models for travel time inference for large scale traffic network in real time [8]. In this application, from an historical data set, we have to build a GMRF reflecting the mutual information between traffic levels among different segments of the traffic network. Then, in real time, GaBP runs on this GMRF to propagate the information given by observed segments to the other ones.

In our experiment with PeMS data, both the historical and test data sets contain samples of a N-dimensional vector of 5-minutes averaged speeds, obtained from fixed sensors. Given a sample, for which a proportion ρ of the variables is observed, we want to infer the states of the $(1-\rho)N$ unobserved variables. In practice, we proceed gradually on each test sample, by revealing the variables in a random order, and plot the relative L_1 error $|\hat{v} - v|/v$ made by the inference model on the unobserved variables as a function of ρ, aggregated over 100 different test samples. The error is measured on the speed, after inference has been done in the space defined by (10).

In principle, $\star$-IPS does not require complete samples and even knowledge of the whole covariance matrix is actually not mandatory. But, for this highway dataset, the samples have no missing values, which allows us to use a brute force k-NN predictor for the sake of comparison. The setting for k-NN is as follows: k samples out of the whole training set are selected according to their mean L_1 distance on the observed variables. Then, for each unobserved variable, the median value is extracted from the k selected samples as a predictor. In the experiments, the value $k = 70$ has been determined to yield the best k-NN predictions.

Fig. 2-right compares the prediction made by $\star$-IPS with the respective results of SELO, full inverse covariance matrix model and k-NN. GMRF models

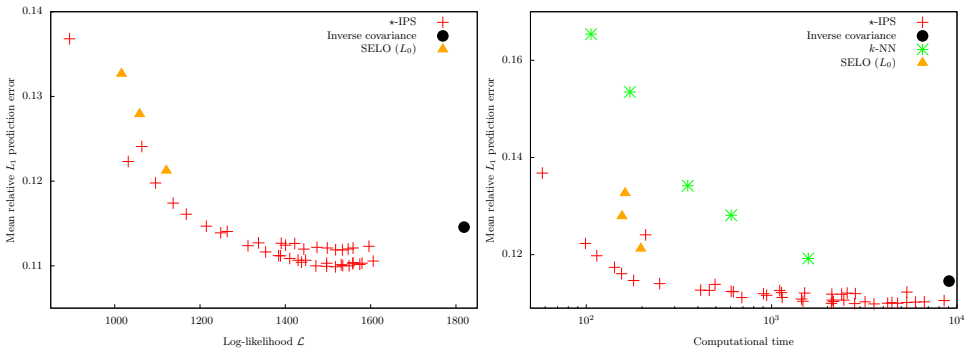


Fig. 3. Mean relative L_1 reconstruction error as a function of log-likelihood (left) and of computational time cumulated over 100 experiments (right). The different results for k-NN are obtained with historical datasets of sizes 10^3, 2.10^3, 5.10^3, 10^4 and 3.10^4.

and k-NN behave very differently. k-NN seems to capture rapidly ($\rho \leq 0.1$) the global network behavior, but remains flat after that, despite the additional information. By contrast, GMRF models performance always improves with new information, because of their local nature. Moreover, while constraints applied to IPS offer us more precise models, the role of $\mathcal{L}$ as proxy is not completely respected. This is due to overfitting problems: as we can see, the full inverse covariance matrix model behaves worse than the $K = 6.5$ model obtained by simple IPS. This appears clearly in Fig. 3-left, where the prediction error reaches a minimum before increasing again with $\mathcal{L}$. Finally, Fig. 3-right shows various trade-offs between precision and efficiency of the models. We clearly see that ⋆-IPS extracts more precise and less time-consuming models for traffic states reconstruction. For instance, the highest precision is obtained with WWS-IPS for $K \approx 50$, leading to a 10-fold time reduction of the computational time w.r.t. the full inverse covariance model, with a gain of 5% in precision.

6 Conclusion

In this paper, we revisit IPS and show that it provides an efficient framework to find GMRF models with constraints more specific than sparsity. Comparisons show the merits of the proposed ⋆-IPS in terms of flexibility, likelihood values reached and diversity of solutions, since a Pareto set can be delivered for the computational cost of one estimation.

In terms of trade-off between sparsity and likelihood, ⋆-IPS is comparable to the SELO approach, with less computational cost. In contrast, L_1 based methods do not provide satisfying results in our problem setting.

In addition, the flexibility of ⋆-IPS allows one to embed additional and rather exotic but useful constraints for GaBP compatibility, which is not simple to do with traditional penalized likelihood approaches. Experiments show that both topological and spectral constraints are useful. While the walk-summability

constraint seems too strict to be useful in practice, relaxing it to weak walk-summability leads to very good models. At the same time, avoiding only frustrated triangles give very satisfactory results in our experiments.

Still in this context, the overfitting problem seems completely open to us. Classical information-theoretic criteria failed in our tests to locate properly where to stop in the incremental link addition process. In fact, we observe that both the link-addition and the link-update procedure can lead to overfitting, and the design of a specific criterion able to avoid this deserves further investigation.

References

1. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. JMLR 9, 485–516 (2008)
2. Bickson, D.: Gaussian Belief Propagation: Theory and Application. Ph.D. thesis, Hebrew University of Jerusalem (2008)
3. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on (1968)
4. Darroch, J., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Statistics 43, 1470–1480 (1972)
5. Dong, B., Zhang, Y.: An efficient algorithm for l_0 minimization in wavelet frame based image restoration. Journal of Scientific Computing 54(2-3) (2012)
6. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association (2001)
7. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441 (2008)
8. Furtlehner, C., Han, Y., Lasgouttes, J.M., Martin, V., Marchal, F., Moutarde, F.: Spatial and temporal analysis of traffic states on large scale networks. In: ITSC. pp. 1215–1220 (2010)
9. Hsieh, C., Sustik, M.A., Dhillon, I.S., Ravikumar, K.: Sparse inverse covariance matrix estimation using quadratic approximation. In: NIPS (2011)
10. Jalali, A., Johnson, C.C., Ravikumar, P.D.: On learning discrete graphical models using greedy methods. In: NIPS. pp. 1935–1943 (2011)
11. Lee Dicker, B.H., Lin, X.: Variable selection and estimation with the seamless-L_0 penalty. Statistica Sinica 23(2), 929–962 (2012)
12. Malioutov, D., Johnson, J., Willsky, A.: Walk-sums and Belief Propagation in Gaussian graphical models. JMLR 7, 2031–2064 (2006)
13. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: COLING. pp. 49–55 (2002)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference. Morgan Kaufmann (1988)
15. Scheinberg, K., Rish, I.: Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. In: ECML-PKDD (2010)
16. Seneta, E.: Non-negative matrices and Markov chains. Springer (2006)
17. Speed, T., Kiiveri, H.: Gaussian Markov distributions over finite graphs. The Annals of Statistics 14(1), 138–150 (1986)
18. Weiss, Y., Freeman, W.T.: Correctness of belief propagation in Gaussian graphical models of arbitrary topology. Neural Computation (2001)