2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

MODELLING DENSE RELATIONAL DATA

Tue Herlau, Morten Mørup, Mikkel N. Schmidt & Lars Kai Hansen

Technical University of Denmark
pCenter of Cognitive Systems
Richard Petersens Plads, b120
2800 Lyngby, Denmark

ABSTRACT

Relational modelling classically consider sparse and discrete data. Measures of influence computed pairwise between temporal sources naturally give rise to dense continuous-valued matrices, for instance p-values from Granger causality. Due to asymmetry or lack of positive definiteness they are not naturally suited for kernel K-means. We propose a generative Bayesian model for dense matrices which generalize kernel K-means to consider off-diagonal interactions in matrices of interactions, and demonstrate its ability to detect structure on both artificial data and two real data sets.

Index Terms— Relational Modelling, Non-parametrics, Infinite Relational Model, Granger Causality

1. INTRODUCTION

Consider the problem of analysing a large set of signal sources $S_i \in \mathcal{S}, i = 1, \ldots, n$ each emitting a time-dependent signal. In this work we consider discrete signals with real domain, $S_i : \{1, 2, \ldots, T\} \mapsto \mathbb{R}$, but this choice is not critical. A concrete instance of the problem could be neural activity in fMRI where each source correspond to a voxel.

An interesting problem in a temporal setting is discovering *influence* amongst the signals, for instance when activity in one group of voxels in the brain is highly informative of activity in another group of voxels at a later time but not vice versa. Notice this work only consider statistical claims of influence and we emphasize true claims of causal relation can only be done after intervention studies or under strong domain assumptions[1].

Typically the signals are first grouped according to similarity (using a standard clustering method) or background information (for instance similarity of neural tissue, and then the influence-analysis is applied on either the clusters of signals or the average of the signals within the clusters[2, 3]. By a measure of influence we consider any real or vector valued function such as Correlation or time-lagged correlation, Granger causality, transfer entropy, etc.[4, 5]. Since it is the influence-based analysis which is interesting, it is natural to ask if the first step could be done away with. One difficulty is the measure of influence $W : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$ will typically be asymmetric and difficult kerneling. Motivated by kernel K-means clustering we will show how the problem is more naturally formulated as discovering structure in matrices, and show how a natural generalization of a well-known model from relational modelling can be used to directly cluster according to the influence measure. These techniques are applied to North American temperature records and fMRI data.

2. METHODS

2.1. The kernel K-means objective and relational modelling

Consider a partition of the set $\{S_i\}_i$ into K clusters. Let $\boldsymbol{z}$ be a $n \times K$ matrix such that $z_{i\mu} = 1$ iff. data point i belong to cluster μ and 0 otherwise. Let $c_\mu = \{j : z_{j\mu} = 1\}$. The kernel K-means objective arise by considering a mapping of each point into a Hilbert space $\phi : \mathcal{S} \mapsto H$ and *minimizing* the objective:

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\phi(S_i) - \left(\frac{1}{n_\mu}\sum_{j\in c_{z_i}}\phi(S_j)\right)\right\|^2 \tag{1}$$

where $n_\mu = |c_\mu|$ is the number of objects assigned to cluster μ. The mapping ϕ is entirely characterized by the Gram-matrix $k_{ij} = \langle\phi(S_i), \phi(S_j)\rangle = k(S_i, S_j)$ provided the mapping k is positive semidefinite[6]. Ignoring constant terms, Kernel K-means become equivalent to *maximizing*[7]

$$\sum_{\mu=1}^{K}\frac{1}{n_\mu}\sum_{i,j\in c_\mu}k_{ij} = \mathrm{diag}((\boldsymbol{z}^T\boldsymbol{z})^{-1}\boldsymbol{z}^T\boldsymbol{k}\boldsymbol{z}). \tag{2}$$

For any $\boldsymbol{z}$ let I be a permutation of $\{1, \ldots, n\}$ which "reorder i according to cluster assignments" (see bottom panes of figure 1). Formally, if $z'_{h\mu} = z_{I(h)\mu}$ then for all $i < j < k$: $z'_{i\mu}p = z'_{k\mu} = 1 \Rightarrow z'_{j\mu} = 1$. Reordering k according to the same permutation give a matrix $k'_{ij} = k_{I(i)I(j)}$, and the

978-1-4673-1026-0/12/$31.00

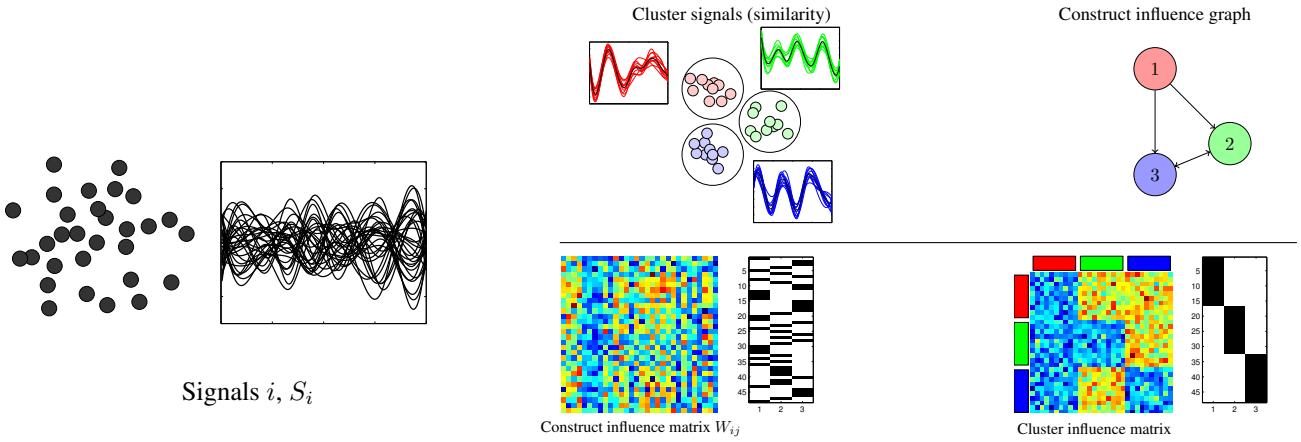


Fig. 1. Concept for analyzing influence amongst signals (left). Classical way (top) depend on clustering signals (top-middle) and then discovering influence amongst the clusters (top-right). Proposed method (bottom) depend on constructing full matrix of influence measure (bottom-middle) and discovering clusters within influence matrix (bottom-right).

objective (2) can now be seen as dividing k' into K^2 submatrices, each of size $n_\mu n_\nu$, and taking the sum of each of the diagonal submatrices divided by n_μ. In other words, the objective only consider what happens *inside* the diagonal blocks induced by $\boldsymbol{z}$, see figure 1 top.

Returning to the general case, consider a generic measure of influence $W : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$, for instance Granger causality[4]. Such relationships are usually asymmetric, and more importantly, the interesting structure in the relationship is exactly off diagonal (see figure 1 bottom): If a group of signals are very similar, they do not provide any extra information in order to forecast each other. Hence even if the mapping $\boldsymbol{W}$ could be kernelized by appropriate symmetrization and eg. squared exponential mapping, one will loose the interesting structure. Furthermore kernel K-means require prespecification of the number of components K.

To overcome these difficulties we propose explicitly modelling the pmatrix of interactions $\boldsymbol{W}$ as a dense relational matrix using a Bayesian generative model inspired by the network litterature[8]. Given $\boldsymbol{z}$, $\boldsymbol{W}$ is naturally divided into K^2 submatrices (figure 1, bottom). The elements of each $n_\mu n_\nu$ submatrix is assumed to be independently drawn from a normal distribution parameterized with mean $m_{\mu\nu}$ and precision $\lambda_{\mu\nu} = \sigma_{\mu\nu}^{-2}$. Using a Chinese Restaurant Process[8] for $\boldsymbol{z}$ and a Normal-gamma prior on the parameters of the multivariate normal distribution we obtain the following generative process (Normal-IRM,nIRM):

$$
\begin{aligned}
\boldsymbol{z} &\sim \mathrm{CRP}(\gamma_0), && \textit{Cluster assignment,}\\
\lambda_{\mu\nu} &\sim \mathrm{Gamma}(\alpha_0, \text{rate} = \beta_0) && \textit{precision,}\\
m_{\mu\nu} &\sim \mathrm{Normal}\left(m_0, (\kappa_0\lambda_{\mu\nu})^{-1}\right) && \textit{mean,}\\
W_{ij} &\sim \mathrm{Normal}(m_{z_i z_j}, \lambda_{z_i z_j}) && \textit{Observed data.}
\end{aligned}
$$

The joint likelihood of the generative model is given by:

$$
\begin{aligned}
p(\boldsymbol{W}, \boldsymbol{m}, \boldsymbol{\lambda}, \boldsymbol{z} \mid m_0, \alpha_0, \beta_0, \kappa_0, \gamma_0) = \\
p(\boldsymbol{W}|\boldsymbol{m}, \boldsymbol{\lambda}, \boldsymbol{z}) p(\boldsymbol{m}|\boldsymbol{\lambda}, m_0, \kappa_0) p(\boldsymbol{\lambda}|\alpha_0, \beta_0) p(\boldsymbol{z}|\gamma_0)
\end{aligned}
$$

An attractive feature of the joint likelihood is that the prior is conjugate so that $\boldsymbol{m}$ and $\boldsymbol{\lambda}$ can be integrated out[9] giving:

$$
\begin{aligned}
p(\boldsymbol{W}, \boldsymbol{z} \mid m_0, \alpha_0, \beta_0, \kappa_0, \gamma_0) = \\
\iint d\boldsymbol{m} d\boldsymbol{\lambda} \quad p(\boldsymbol{W}|\boldsymbol{m}, \boldsymbol{\lambda}, \boldsymbol{z}) \\
\times p(\boldsymbol{m}|\boldsymbol{\lambda}, m_0, \kappa_0) p(\boldsymbol{\lambda}|\alpha_0, \beta_0) p(\boldsymbol{z}|\gamma_0) = \\
\left[\prod_{\mu\nu} \frac{\Gamma(\alpha_{\mu\nu})}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_{\mu\nu}^{\alpha_{\mu\nu}}} \sqrt{\frac{\kappa_0}{\kappa_{\mu\nu}}} (2\pi)^{-\frac{n_\mu n_\nu}{2}} \right] \\
\times \left[\frac{\gamma_0^K}{\Gamma(n + \gamma_0)} \prod_{\mu=1}^{K} \Gamma(n_\mu) \right]
\end{aligned} \tag{3}
$$

with the definitions[9]:

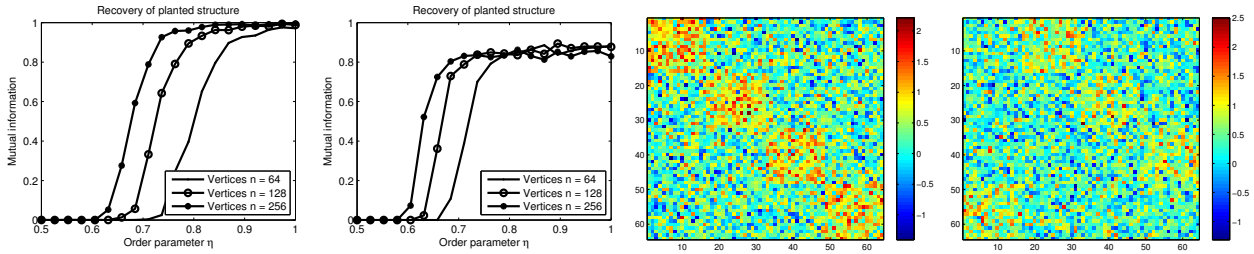
$$\alpha_{\mu\nu} = \alpha_0 + \frac{n_\mu n_\nu}{2} \tag{4a}$$

$$\kappa_{\mu\nu} = \kappa_0 + n_\mu n_\nu \tag{4b}$$

$$\beta_{\mu\nu} = \beta_0 + \frac{C_{\mu\nu}^{(2)}}{2} - \frac{\left(C_{\mu\nu}^{(1)}\right)^2}{2 n_\mu n_\nu} + \frac{\kappa_0 (C_{\mu\nu}^{(1)} - n_\mu n_\nu m_0)^2}{2(\kappa_0 + n_\mu n_\nu) n_\mu n_\nu} \tag{4c}$$

and pseudo-counts $C_{\mu\nu}^{(a)} = \sum_{i \in c_\mu, j \in c_\nu} W_{ij}^a$. As the name implies, the model can be seen as the Infinite Relational Model[8] with normal observations and normal-gamma priors instead of bernoulli observations and beta priors.

In case the measure of influence is symmetric, one can easily restrict the model to the upper triangular part $W_{ij}, i \leq j$ and thus the products are restricted to only cover $\mu \leq \nu$. The



(a) Symmetric recovery (b) Asymmetric recovery (c) Symmetric relation, $\eta = 0.8$ (d) Asymmetric relation, $\eta = 0.7$

Fig. 2. Artificial data results. Results obtained by varying η for the two different types of structure (c)-(d) induced by $\boldsymbol{B}$ using $\sigma = \frac{1}{2}$ (see text). Simulations performed with the split-merge sampler described in the text. For each η-value the NMI between planted/recovered structure is obtained by averaging over 40 runs using 20 Gibbs sweeps per run and 4 restarts from random configurations with priors $\gamma_0 = \alpha_0 = \kappa_0, \beta_0 = m_0 = 1$. Notice the change in detection treshold between the symmetric/asymmetric networks.

model was also extended to model multiple influence matrices $\{\boldsymbol{W}_r\}_{r=1}^R$ by sharing the assignments $\boldsymbol{z}$, ie.

$$p(\{\boldsymbol{W}_r\}_{r=1}^R \mid \boldsymbol{z}) = p(\boldsymbol{z}) \prod_{r=1}^R p(\boldsymbol{W}_r \mid z)$$

and use different sets of hyperparameters $(\alpha_0, \beta_0, m_0, \kappa_0)$ for the diagonal blocks ($\mu = \nu$) and off-diagonal ($\mu \neq \nu$), all of these extensions are trivial and details will not be given here[8].

2.2. Efficient Implementation

Inference in the nIRM require some considerations both in terms of speed and reasonable mixing. A key issue for larger problems is memory consumption. Consider a single Gibbs update of variable pi to initial assignment matrix $\boldsymbol{z}$. Writing $\boldsymbol{z}^* = \boldsymbol{z}_0 + \boldsymbol{\Delta}$ with $\Delta_{lm} = \delta_{il}\delta_{\mu m}$ the Gibbs update equation for $\boldsymbol{z}^*$ become:

$$p(\boldsymbol{z}^* \mid \boldsymbol{W}) \propto p(\boldsymbol{W}, \boldsymbol{z}_0 + \boldsymbol{\Delta}).$$

From the form of the joint likelihood (3) and update equations (4) it is only necessary to keep track of changes to the pseudo-count matrices $\boldsymbol{C}^{(a)} = \boldsymbol{z}^T \boldsymbol{W}^a \boldsymbol{z},\ a = 1, 2.$

$$\begin{aligned}(\boldsymbol{z}_0 + \boldsymbol{\Delta})^T \boldsymbol{W}^a (\boldsymbol{z}_0 + \boldsymbol{\Delta}) = \left[\boldsymbol{z}_0^T \boldsymbol{W}^a \boldsymbol{z}_0 + W_{ii}^a\right] \\ + \boldsymbol{z}_0^T W_{:,i}^a \boldsymbol{e}_\mu^T + \boldsymbol{e}_\mu W_{i,:}^a \boldsymbol{z}_0\end{aligned}$$

where notation $\boldsymbol{W}^a$ denote the element-wise exponential, $w_{:,i}$ is the ith column and $\boldsymbol{e}_\mu$ the μth canonical basis vector. Thus considering each of the $K+1$ possible assignments of z_i (the last assignment denoting a new cluster) is equivalent to only computing $2(K+1)^2$ changed entries of $\boldsymbol{C}$ and the normalization term in the product in equation (3). Furthermore, since $\boldsymbol{W}$ only enters as products of the form $\boldsymbol{z}_0 \boldsymbol{W}^a$.

Caching these products lower the memory requirement from $\mathcal{O}(n^2)$ to $\mathcal{O}(nK)$. The cost of this procedure is that the cached products $\boldsymbol{z}_0 \boldsymbol{W}^a$ need to be updated *but only when* $z_i^* \neq z_i$, and when this occur one need to re-compute row/column i of $\boldsymbol{W}$. However, if these can be calculated efficiently from the data, for instance in the case they can be written as vector products (such as time-lagged correlation), the savings can be very substantial since typically less than 10% of the assignments in $\boldsymbol{z}$ is updates in each sweep after a few iterations. Using these techniques it was possible to sample problems of up to $65'000$ vertices where storage of $\boldsymbol{W}$ would be infeasible.

2.2.1. *Split-merge sampling*

In practice Gibbs sampling works well for determining how vertices should be assigned between existing clusters, but has difficulties discovering new clusters since this often require simultaneous change of multiple assignments.

To overcome this difficulty the split-merge sampling framework proposed by S. Jain[10] was considered. The key to the method is using the Gibbs sampler to produce favorable split configurations. Draw two distinct vertices i, j at random with assignments $z_i = \mu, z_j = \nu$. Construct a new *launch state* $\boldsymbol{z}^\ell$ by first letting $z_i^\ell = \mu$, $z_j^\ell = \nu' \notin \{z_h\}_h$ and for all other $h \in S = \{h | z_h = \mu \text{ or } z_h = \nu\}$ assign z_i^ℓ to either μ or ν' at random. All other elements are not reassigned. Secondly perform q Gibbs sweeps on assignments $z_i^\ell, i \in S$ *but constrained to only considering assignment to clusters* μ *or* ν' to obtain the final launch state.

If $z_i = z_j$ propose a *split* by performing one final restricted gibbs sweep on $\boldsymbol{z}^\ell$ to obtain $\boldsymbol{z}^*$, store the Gibbs transition probability for the final sweep as $T^{\text{split}} = 1$ and let $T^{\text{merge}} = 1$.

Alternatively propose a *merge* move by computing the Gibbs transition probability of a single sweep on $\boldsymbol{z}^\ell$ where each $i \in S$ is constrained to be assigned to its original configuration, ie. $\boldsymbol{z}_i^* = z_i$, and let T^{merge} denote the corresponding probability and T^{split}. The proposal is accepted with metropolis-hastings

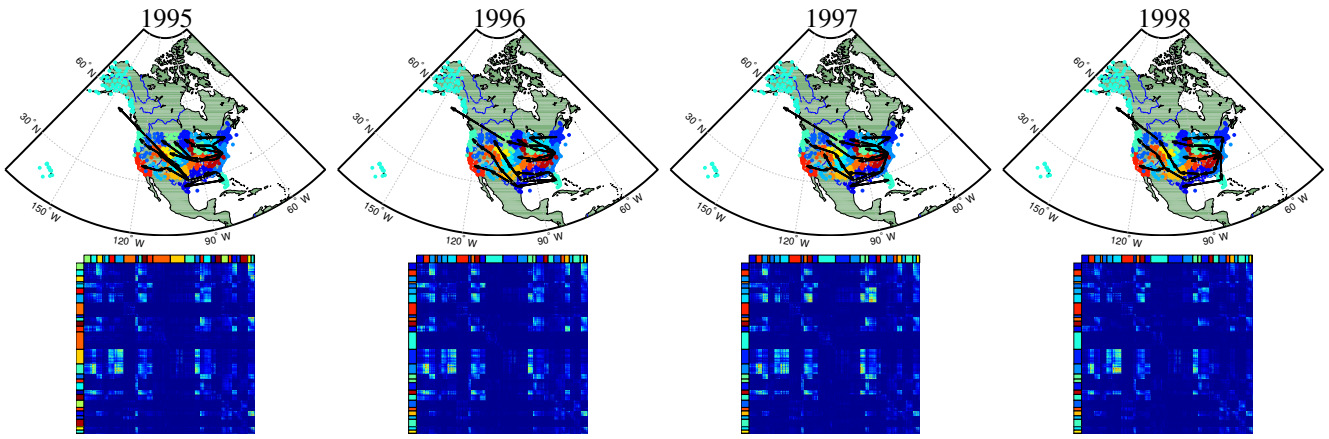


Fig. 3. Results of NCDC data set. Results obtained from the maximum-likelihood configuration after 100 gibbs sweep of the assymetric formulation of the nIRM. Colors indicate the different clusters ($K = 10$), arrows indicate prominent directions of interaction (in the sense of large $m_{\mu\nu}values$, notice there is some difference across the years since these values are not shared). Hyperparameters was choosen as $\gamma_0 = \alpha_0 = \beta_0 = \kappa_0 = 1$, however we used different values of m_0 on the diagonal: $m_0 = 0$ if $\mu = \nu$ and $m_0 = 20$ if $\mu \neq \nu$ to account for the strong off-diagonal nature of $\boldsymbol{W}_r$, see also 4 for the year 1999.

ratio:

$$\min\left\{1, \frac{T^{\text{split}}}{T^{\text{merge}}} \frac{p(\boldsymbol{W}, \boldsymbol{z}^*)}{p(\boldsymbol{W}, \boldsymbol{z})}\right\}$$

which ensure detailed balance[10]. The number of intermediate gibbs sweeps q is a parameter of the model, here $q = 1$. While the above method significantly enhance the ability to discover new clusters by splitting, the probability of merge moves tend to be low, and so a variation of the above method where split-configurations are proposed randomly was also implemented[10].

3. SIMULATIONS

We tested the feasibility of the method on both artificial and real data. Constructing the artificial data was done by choosing n and K and a random assignment matrix $\boldsymbol{z}$ assigning $\frac{n}{K}$ vertices to each cluster.

Secondly the structure of the data was constructing by choosing an underlying $K \times K$ structure matrix indicating which clusters interact, here we choose a symmetric interaction matrix by letting $\boldsymbol{B}$ be the $K \times K$ identity matrix and simulated an off-diagonal interaction by letting $\boldsymbol{B}$ be the identity with the columns circularly permuted to the right (see figure 2).

Finally the difficulty of the problem is determined by an *order parameter* $\eta \in [0, 1]$ such that $\eta = 0$ denote an impossible task and $\eta = 1$ the easiest case. The matrix $\boldsymbol{W}$ is constructed as

$$W_{ij} = \epsilon_{ij} + \eta B_{z_i z_j} \text{ and } \epsilon_{ij} \sim \text{Normal}(0, \sigma)$$

and two examples can be seen in figure 2 (c)-(d), both for the symmetric and asymmetric matrices.

As a measure of performance we use the *normalized mutual information* (NMI) which can be interpreted as the fraction of the total amount of information in the recovered structure $\boldsymbol{z}$ which can be learned from knowing the planted structure $\tilde{\boldsymbol{z}}$.[11]. The NMI is defined as

$$\text{NMI}(\boldsymbol{z}, \widetilde{\boldsymbol{z}}) = \frac{2I(\boldsymbol{z}, \widetilde{\boldsymbol{z}})}{H(\boldsymbol{z}) + H(\widetilde{\boldsymbol{z}})}$$

where $I(\boldsymbol{z}, \widetilde{\boldsymbol{z}}) = \sum_{\mu\nu} p(\mu, \nu) \log \frac{p(\mu,\nu)}{p(\mu)p(\nu)}$, $H(\boldsymbol{z}) = I(\boldsymbol{z}, \boldsymbol{z})$ is the entropy and the distribution $p(\mu, \nu)$ is the probability a randomly selected observation i in cluster μ in the planted cluster structure $\boldsymbol{z}$ is in cluster ν in the inferred structure $\widetilde{\boldsymbol{z}}$[11].

Notice if the two variables are independent the NMI is zero and if they are identical the NMI becomes 1. NMI has been shown to be an efficient measure of partitions in the type of problem considered here [11]. It has become a standard in assessing the quality of communities in artificial relational data [12, 13, 14].

In line with [14] we choose $K = 4$ and let n take values $64, 128, 256$ from random initial configurations and using multiple restarts, the results can be seen in figure 2. As expected the model undergoes a phase transition as η is increased for both the symmetric (where only $W_{ij}, i \leq j$ is modelled) and the asymmetric nIRM. Surprisingly, the results show different threshold values for η (asymmetric graphs detected earlier) but slightly lower total detection rate for high values of η. This may be due to the asymmetric graphs containing twice as many observations and hence obtain lower local minima.

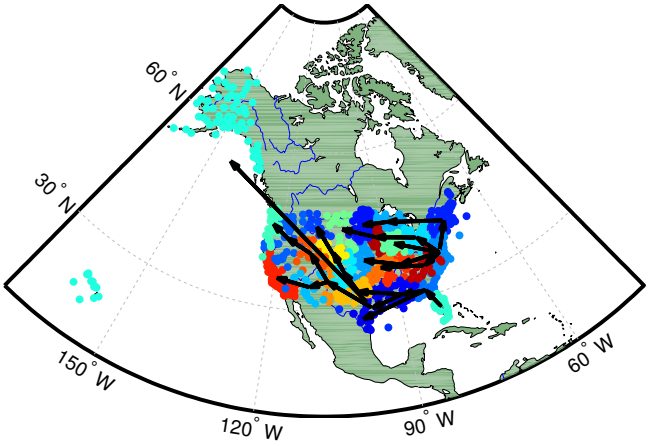


Fig. 4. Enlargement of NCDC weather results for the year 1999, see figure 3 for description

3.1. NCDC Weather Data

We attempted the model on weather data from the National Climate Data Center (NCDC)[1]. We used temperature data from the years $1995 - 1999$ for stations in USA. Since some of the stations did not have data for all days, stations with less than 300 observations for any given year was removed and missing observations imputed by linear interpolations using only the stations own time series. This exclude artificial spatial correlations. After this 879 stations remained.

For each pair of stations a Granger causal analysis[4] was performed using the *Econometrics Toolbox*[15] on the lagged signal $\hat{S}_i(t) = S_i(t) - S_{t-1}$. Granger causality fit a model of the type

$$\hat{S}_i(t) = \sum_{\tau=1}^{L} a_{ij,l}\hat{S}_i(t-\tau) + \sum_{\tau=1}^{L} a'_{ij,l}\hat{S}_j(t-\tau) + E_{ij}(t)$$

and test if the primed coefficients (obtained by OLS) are jointly different from zero using an F-test of the null hypothesis $a_{ij,\tau} = 0, \tau = 1, \ldots, L$ [16]. The corresponding p-values for year r was recorded as W_{ijr}. It is easy to see from the definition $W_{iir} = 0$ and the corresponding matrices are strongly non-negative and asymmetric.

Results of running the model for 1000 Gibbs sweeps can be seen in figure 3 and 4. The clustering with the largest likelihood is illustrated by different colors (around 10 clusters are discovered), and the interactions are illustrated using the posterior mean of the R $K \times K$ matrices $\boldsymbol{m}$. For each ν, $\mu = \text{argmax}_\kappa\, m_{\kappa\nu}$ was computed and an arrow drawn from μ to ν, ie. connecting each region to the other region which *"influence"* it the most.

Four of the $\boldsymbol{W}$-matrices are permuted according to $\boldsymbol{z}$ and can be seen in figure 3, notice the clusters are completely dominated by the off-diagonal interactions, in other words, when spatially proximate weather stations are clustered together, it is not a-priori because their signals are similar, but because they has the same interaction profile towards other weather stations. In this light it is interesting the connectivity profile (see arrows in figure 4) tend to clusters which are spatially close, since their within-cluster densities $m_{\mu\mu}$ are often very low (see matrices in figure 3). This indicate the method is able to discover the spatial resolution (size of clusters) where Granger causality is informative at a cluster-level.

Another interesting feature is the large degree of similarity between the recovered patterns across the years, and the presence of an "overall direction of flow" from the north-eastern states and directly west towards North/South Dakota, and a West-North direction from Florida towards California and Alaska.

[1]NCDC data freely available from
`http://www.ncdc.noaa.gov/cgi-bin/res40.pl`

3.2. Functional MRI

We applied our method to resting state fMRI data obtained from the Berlin dataset of `FCON1000`, see reference for details on pre-processing[17]. For each of the 26 subjects we extracted a single slice containing 2180 voxels and for each voxel the time series was normalized to zero mean and unit variance. Due to the noisy nature of the fMRI time series we applied linear time-lagged correlation to obtain 26 asymmetric relations.

The nIRM typically produced ~ 60 clusters. In figure 5 is the maximum-likelihood clustering obtained after 100 Gibbs sweeps. Arrows are drawn using similar method as in the NCDC data using the averaged correlation across all subjects

The method parcellate the brain into spatially homogeneous regions with some degree of left/right symmetry, but a more in depth study is required to reliably interpret the result in terms of clinical relevant structure due to the issues surrounding causal discovery in fMRI [18]. In particular temporal variations (both within the same subject and across populations of subjects) of the hemeodynamic response to neural activity is bound to pose a problem and may explain variations in time-lagged correlation across the subjects.[18]

3.3. Discussion

Simulations on real data sets indicate the proposed method is able to discover seemingly relevant structure. Comparing the method to kernel K-means, the method is able to model much more general matrices and in this sense offer a generalization. Furthermore, being based on the CRP, it allow automatic inference of the number of clusters as well as predicting the value of unobserved entries in $\boldsymbol{W}$. However, a point of criticism is a group of signals which interact *in a similar manner* to another group of signals necessarily must share some common property. Provided it is possible to construct a kernel function which capture this property kernel K-means will apply well to the data.

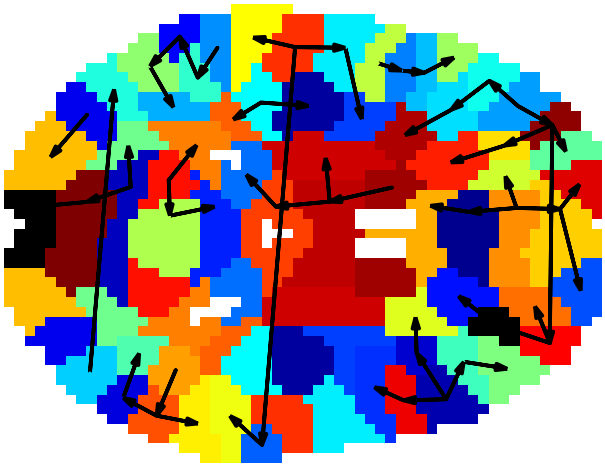


Fig. 5. Results on fMRI data, see text for details.

Finally it is possible to extend to the method by replacing the normal-gamma prior with a Normal-Wishart prior. Doing this W_{ij} is a vector rather than a scalar, and the clusters will capture non-trivial covariance structure, but we defer this extension to future work.

4. CONCLUSION

We have proposed a method for inferring influence structure amongst signal sources. This was done by constructing a relational model for dense networks. We demonstrated the feasibility of the model in terms of recovering both symmetric and asymmetric structure on artificial data, and demonstrated such structure is present in real data.

5. REFERENCES

[1] J. Pearl, *Causality: Models, Reasning and Inference*, Cambridge University Press, 2000.

[2] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol, "Exploring the brain network: a review on resting-state fmri functional connectivity.," *European neuropsychopharmacology the journal of the European College of Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[3] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[4] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. pp. 424–438, 1969.

[5] Thomas Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[6] A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[7] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2004, KDD '04, pp. 551–556, ACM.

[8] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, 2006.

[9] Kevin P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," Tech. Rep., 2007.

[10] Sonia Jain and Radford M. Neal, "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, pp. 158–182, 2004.

[11] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, pp. P09008, 2005.

[12] A. Lancichinetti, "Community detection algorithms: a comparative analysis," *Physical Review E*, vol. 80, no. 5, pp. 056117, 2009.

[13] G. Orman and V. Labatut, "A comparison of community detection algorithms on artificial networks," in *Discovery Science*. Springer, 2009, pp. 242–256.

[14] J. Reichardt and J. Reichardt, *Structure in complex networks*, Springer Verlag, 2009.

[15] J.P. LeSage and R.K. Pace, *Introduction to Spatial Econometrics*, Statistics, Textbooks and Monographs. CRC Press, 2009.

[16] John F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. pp. 907–915, 1984.

[17] Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, and et al., "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[18] J D Ramsey, S J Hanson, C Hanson, Y O Halchenko, R A Poldrack, and C Glymour, "Six problems for causal inference from fmri.," *NeuroImage*, vol. 49, no. 2, pp. 1545–1558, 2010.