Constrained Principal Component Analysis

Identifying relationships with known structure using PCA and sub-selection of variables

-Manoj Kumar Gootam

Background and significance of the proposed research

Several physical phenomena have hidden structural constraints driven by the dynamics of the processes. Technologies like DNA micro-array, Utility distribution networks for water, steam, current, air, Data communication network etc. have an underlying lower dimensional dynamics driving the higher dimensional data observed through metering. But while identifying the relationships between the inputs and the outputs of these processes, this connectivity information is generally ignored. This approach usually leads to absurd results with no practical sense since the governing physical laws are violated. To avoid such nonsensical results in real life datasets, connectivity constraints should be honored while estimating the model equations.

In this research study, I propose a novel and an innovative approach to solve the problem of reconstruction of the underlying structure (constraint matrix) and the input signals entirely from the outputs and the network connectivity information between the input signals. The proposed approach imposes the connectivity constraints and uses an iterative procedure to estimate the constraint matrix and the input signals recursively until convergence, while the estimation is done using the sub-selection of variables and the principal component analysis. The proposed approach obtains solutions unique upto a scaling factor if they are NCA (Network Component Analysis) compliant, otherwise a unique solution may not be identifiable.

Research objectives

In this research, I address the problem of reconstruction of the underlying structure (constraint matrix) and the input signals entirely from the outputs and the network connectivity information among the input signals. In a parallel viewpoint, the problem is also the decomposition of a high dimensional output data matrix into the product of constraint matrix and the low dimensional input data matrix, where constraint matrix contains the underlying connectivity between the input variables. In addition to the connectivity information, if there is partial information available on some of the linear relationships between variables, then it is possible to reconcile the data and identify the remaining linear relationships using the partially known linear relationships.

Research plan

Mathematical Framework: Let Y[NxM] be the output matrix(rows as features and columns as samples), A[NxP] be the constraint matrix and X[PxM] be the input matrix(rows as features and columns as samples) with the following relationship

$$Y_{NxM} = A_{NxP} * X_{PxM} \quad \text{---------(1)}$$

Let I[NxN] be an identity matrix with principal diagonal elements as 1 and the off-diagonal elements as 0.

Pre-multiplying Equation 1 with the I[NxN], identity matrix, we get the following equation,

$$I_{NxN} * Y_{NxM} = A_{NxP} * X_{PxM}$$

$$[\,I \quad A\,]\begin{bmatrix} Y \\ -X \end{bmatrix} = 0$$

$$B * Z = 0, \quad \text{------------(2)} \qquad here\ B\ is\ [\,I \quad A\,]\ and\ Z\ is\ \begin{bmatrix} Y \\ -X \end{bmatrix}$$

Constrained Principal Component Analysis

Identifying relationships with known structure using PCA and sub-selection of variables

-Manoj Kumar Gootam

Equation 2 is in the typical regression format, with B being the coefficient matrix and Z being the augmented data matrix

Let C[N x N+P] be the connectivity matrix, representing the structure of the matrix B, by having 0s in the positions where B has 0s and 1s where B has non-zero values.

Coefficient Matrix (B)			
2	0	-3	0
0	5.1	0	4
0	3	2.6	0
4	0	0	7

Connectivity Matrix (C)			
1	0	1	0
0	1	0	1
0	1	1	0
1	0	0	1

If an entry $C_{ij} = 1$, then it indicates the presence of the corresponding jth input feature in the ith linear equation, when $C_{ij} = 0$, then that corresponding jth input feature is ignored in the ith linear equation.

Algorithm:

1. Guess the input matrix(X_0)
2. Using the input matrix, transform equation 1 into regression form as in equation 2
3. Estimate B_i using the model estimation technique described in the following section and then compute A_i and rescale it
4. Update the input matrix, X_i, using the estimated A and the output matrix Y as shown below
$$X_i = \left(A_i^T A_i\right)^{-1} A_i^T Y$$
5. Repeat steps 2, 3 and 4 until the frobenius norm of the error matrix is lower than a predefined threshold, ϵ $\|Y - A_i X_i\| < \epsilon$

Model Estimation:

1. Sub-selection of variables that participate in a model equation as given in the connectivity matrix, C, and subset the input data accordingly.
2. Identify the principal component corresponding to the least eigenvalue in the reduced dimensions and transform it into the original dimensions by adding zeros in the other absent variables.
3. Retain this principal component as the model equation, if the rank of the model equations matrix is equal to the number of the equations, otherwise go for the principal component corresponding to the next least eigen value
4. Continue this procedure for every equation in the structure matrix (from step 1 to 3) till all the equations are obtained.

Since these PCs have only a subset of the original dimensions, the rest of the variables that don’t participate are automatically set to zeros.

The proposed algorithm attempts to utilize the knowledge of the structure of the constraint to apply PCA on the given data set and find a constraint matrix that has the same structure as what we have started with. The variables with non-zero coefficients in a particular constraint are selected as a sub-problem and solved for constraint on the variables independently. The eigen vectors corresponding to the least eigen value of the covariance matrix is chosen to be the constraint matrix. This is brought into the original dimension by padding zeroes at the corresponding positions of zero coefficients. This is done for all the constraints.

Conclusion:

Estimating the relationships using the proposed approach is better than naïve regression techniques like PCR as observed on some test datasets, where the solution is closer to the actual model than solution of PCR is to the actual model. Also in the presence of noise, this approach performs better than PCR with lower subspace angles between the estimated models and the true model. Also, ignoring a structure or using a wrong structure leads to very bad estimations of the model equations. So, imposing the connectivity constraints on the model equations is a good way to identify the true relationships between the input output variables of a process. In scenarios where additional information is available in the form of partially known model equations, modelling can be improved further using the proposed approach.

One of the major limitations of this approach is that non-linear relationships cannot be estimated although there are indirect ways of modelling such cases. Also, for large and highly sparse constraint matrices with bad condition number, computation of the inverse may be a bit difficult impacting the efficiency of the algorithm, but this can also be avoided using efficient algorithms for computing matrix inverse.

Availability of the data

Any physical phenomenon with underlying network dynamics, that needs to be modeled using data with the help of the connectivity information or partially known relationships between variables, is a potential application for the proposed algorithm. This can be applied to a wide class of network dynamics reconstruction problems including, but not limited to, the following

- Reconstruction of regulatory signals in biological systems: This can be directly applied to Gene Expression Regulation data, DNA microarray data. Many other types of large scale data, such as, include neuronal signals, signal transduction data, metabolic fluxes, protein-protein interaction information, may potentially be modeled as the output of underlying functional networks that are driven by regulatory signals. There is a lot of data available online related to the Gene Expression Regulation data in the link here, https://www.ncbi.nlm.nih.gov/gds
- Utility Distribution networks: Network topology for a water pipeline distribution system, can be modeled using the proposed algorithm. Data collected from pressure sensors at the valves can be used to identify the flow patterns through the network. For that matter, any utility distribution network would be a potential application for the proposed algorithm. There are online resources for data related to the utility distribution networks (water, gas, current, etc) and this information can also be collected from the installed systems as they are usually well equipped with sensors for diagnostic purposes.
- Data Communication Networks: Communication networks (for example, Internet) that provide digital data transfer between resources can be modeled for performance optimization. This data is also available online in this link, https://snap.stanford.edu/data/