AMIDST Model Testing

July 2, 2014

1 Data Set Description

By simplicity, we assume that our data set can be describe by a matrix D with three dimensions $N \times M \times T$:

- **M** is the number of variables:
  - Daimler: Velocity, Relative Distance, etc.
  - CajaMar: N. of annotations, account balance, etc.
  - Verdande: BIT, Pressure, ROP, etc.
- **T** is the length of the sequence[1]:
  - Daimler: Number of samples during one maneuver.
  - CajaMar: N. of days since a costumer opened an account in the bank.
  - Verdande: Number of samples during a drilling operation or a drilling day.
- **N** is the number of data sequences
  - Daimler: Number of maneuvers.
  - CajaMar: Number of customers.
  - Verdande: Number of drilling days or drilling operations.

In that way, our data set is composed by a set of data sequences. Each data sequence is composed by a sequence of vector-valued observations. Then, an observation is a M-dimensional vector; a data sequence is a $M \times T$ matrix; and the set of N data sequences defines our data set.

[1] It might actually happen that the data sequences have different length

2 Empirical Distribution Based Analysis

2.1 Univariate Marginals

This analysis is based on a visual inspection, using a histogram, of the empirical distribution of a single variable. The aim is to test distributional assumptions. As our data is composed by a set of data sequences, there could be several ways to make this analysis:

- **Aggregate over the T dimension:** We obviate the temporal index and assume all the observations across time are i.i.d., in addition to the assumption that the observations across N are also i.i.d.
- **Mean over the T dimension:** For each data sequence we compute the mean and we end up with a $M \times N$ matrix from which we compute our histogram.
- **Mean over the N dimension:** For each variable we compute its mean across N and we end up with a $M \times T$ matrix from which we compute our histogram.
- **Local Marginals:** Plot the histogram for the marginal over a single data sequence (i.e. obviate the rest of data sequences). Build this same histogram over different data sequences and see whether it is or not the same.
- **Class Variable:** The above analysis can be made by conditioning over the class variable (if applicable). I.e. take only the data sequences associated to the same single class.

2.2 Bivariate Temporal Marginals

Let us denote x_t the observation of a single variable X at time t. This analysis is focused on analyzing the empirical bivariate distribution of the observations $\{(x_t, x_{t-1})\}_{t\in\{1,\ldots,T\}}$. The employment of a contour plot is a way to accomplish this. In the case of a single (and large) data sequence this analysis is straightforward. Here we have a set of data sequences, so this analysis could be done in different ways as happened with the univariate marginals.

3 Auto-correlation Based Analysis

3.1 The Correlogram

The distinguishing characteristic of a time series is that it can exhibit serial correlation; that is, correlation over time. Let $x_1, ..., x_T$ be a univariate time series. The *sample autocorrelation coefficient at lag v* is given by

$$\hat{\rho}_v = \frac{\sum_{t=1}^{T-v}(x_t - \bar{x})(x_{t+v} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

$\bar{x}$ is the sample mean. The plot of $\hat{\rho}_v$ versus v for $v = 1, ..., M$ for some maximum M is called the **correlogram** of the data.

However, in our case we do not have for a single variable a single time series as stated above. For a single variable we have a set of N time series. If we denote by $\hat{p}_v^i$ to the sample autocorrelation coefficient at lag v for the i-th time series associated of the variable X, we could compute this autocorrelation for all $i = 1, ..., N$ and we would have a set of sample autocorrelation coefficients: $\{\hat{p}_v^1, ..., \hat{p}_v^N\}$. Using these coefficients we could build a correlogram by plotting, instead of $\hat{p}_v$ versus v, a R “box plot” summarizing the coefficients $\{\hat{p}_v^1, ..., \hat{p}_v^N\}$ versus v.

If the variable X is not continuous, all the above analysis could be extended by using an association measure such as the mutual information.

3.2 The Partial Correlogram

Let us denote by X_t to the random variable associated to X taking values at time t. We can build the following regression problem:

$$X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + ...a_{v-1} X_{t-v-1}$$

Let us also denote $e_{i,v}$ to the residuals of this regression problem (i.e. the error when estimating X_t using a linear combination of the $v - 1$ previous observations) . The *sample partial autocorrelation coefficient of lag v*, denoted by $\hat{\theta}_v$, is the the standard sample autocorrelation between the variable X_{t-v} and these residuals. Intuitively, the sample partial autocorrelation coefficient of lag v can be seen as the correlation between X_t and X_{t+v} after having removed the common **linear** effect of the data in between.

When plotting $\hat{\theta}_v$ versus v for $v = 1, ..., M$ for some maximum M, we get the **partial correlogram** of the data. As above, this discussion only applies to one time series, but this plot can be extended for N different time series using R box plots as commented before.

If the variable X is not continuous, the above analysis could be extended by using an association measure such as the mutual conditional information.

4 Static Bayesian Network Model

With this analysis, the idea is to obviate the temporal dimension of the data and build from our original three-dimensional $N \times M \times T$ matrix a bidimensional matrix $(N \cdot T) \times M$ by aggregating the points of all the data sequences and removing the temporal dimension. Continuous variables will be discretized and the application of some automatic structural learning algorithm for Bayesian networks can be applied. The obtained BN structure could show us some relevant (conditional) independences between the variables of the domain problem.

5 Vector Auto-regressive (VAR) Models

Let us denoted by $\{X^1, ..., X^M\}$ to the M variables of our domain problem. We also denote by X_t^m to the random variable associated to the observations of the variable X^m at time t. If X^m is a continuous random variable we can define the following regression problem:

$$\begin{aligned} X_t^m &= b_m + a_{m,1}X_{t-1}^m + a_{m,2}X_{t-2}^m + ... + a_{m,v}X_{t-v}^m && (1) \\ &+ b_1 + a_{1,0}X_t^1 + a_{1,1}X_{t-1}^1 + a_{1,2}X_{t-2}^1 + ... + a_{1,v}X_{t-v}^1 && (2) \\ &+ .... && (3) \\ &+ b_M + a_{M,0}X_t^M + a_{M,1}X_{t-1}^M + a_{M,2}X_{t-2}^M + ... + a_{M,v}X_{t-v}^M && (4) \\ & && (5) \end{aligned}$$

I.e., we try to predict X_t^m using the same variable at previous time stamps (with lag v) and the rest of the variables at previous time stamps (with lag v) but also at the same time stamp. The key point here is to learn this regression model by using a L1-regularizer which shrinks the a_{ij} coefficients to zero. With this analysis we could approximately infer which are the variables which directly predict X_t^m and have some idea about how the structure of the dynamic BN should look like. The book "Bayesian Networks in R" explains how to implement this analysis in R.

In the case that X_t^m is a discrete variable, the same analysis can be applied by using logistic regression with a L1-regularizer.