Partial Least Squares to Identify Functional Dynamics of Proteins

Tatyana Krivobokova, Marco Singer, Bert de Groot, Axel Munk
Georg-August-Universität Göttingen, Germany,
Corresponding author: Tatyana Krivobokova, email: tkrivob@gwdg.de

Abstract

It has been recently demonstrated on several examples that the (multivariate) partial least squares algorithm can be successfully used to study certain functional mechanisms of proteins. This is achieved by identification of collective modes of internal protein dynamics that maximally correlate to an external order parameter(s) of functional interest. Thereby, the standard partial least squares algorithm need to be adjusted to the specific goals and data types. In this talk we discuss a new multivariate partial least squares algorithm and its statistical properties, as well as performance of (multivariate) partial least squares in presence of serial correlation.

Key Words: collective modes, serially correlated data

1. Statistical methods for protein dynamics

Protein function frequently requires dynamics. Ranging from transporters to enzymes, from motors to signaling proteins, conformational transitions are usually at the heart of protein function. Consequently, a key step in understanding protein function is detailed knowledge of the underlying dynamics. Molecular dynamics (MD) simulations and related techniques are routinely used to study the dynamics of biomolecular systems at atomic detail at timescales of typically nanoseconds to microseconds. Although in principle allowing to directly address function-dynamics relationships, such analyzes are frequently hampered by the large dimensionality of a proteins configuration space, rendering it non-trivial to identify collective modes of motion that are directly related to a functional property of interest.
In statistical terms, a linear regression model needs to be estimated

$$f = X\beta + \varepsilon,$$

with $f \in \mathbb{R}^n$ containing n samples of the unidimensional property to be described in terms of a $n \times p$ matrix X of p cartesian coordinates of protein atoms at n time points, obtained from MD simulations. Typically, both n and p are very large (but still $p \ll n$) and many columns of X are collinear. Moreover, observations in columns are serially correlated.
Application of the Principal Component Analysis (PCA) to effectively reduce the dimensionality of a proteins configuration space p is not always successful, since PCA sorts the collective modes (eigenvectors) according to their contribution (eigenvalue) to the total mean-square fluctuation, but completely ignores f. Therefore, Krivobokova et al. (2012) applied Partial Least Squares (PLS) to identify a hidden relation between atom coordinates of a protein and a functional parameter of interest. The PLS algorithm proves to yield robust and parsimonious solutions. Several examples illustrate that the PLS-based analysis successfully reveals the collective dynamics underlying the fluctuations in selected functional order parameters, delivering completely new insights. For

example, the aquaporin channel Aqy1 case reveals a gating mechanism that connects the inner channel gating residues with the protein surface, thereby providing an explanation of how the membrane may affect the channel.
However, the proteins are complicated structures, that can be seen as a collections of certain units. In particular, it is extremely important to understand the relationship of these units dynamics to the atom movements within the units, which naturally leads to the application of a multivariate version of the PLS algorithm.

2. PLS algorithms

In univariate PLS k ($k \ll p$) new regressors T_k are defined iteratively such that each coordinate is a linear combination of the original coordinates X ($T_k = XW_k$) with maximal covariance with $f \in \mathbb{R}$, while being uncorrelated to each previous coordinate in T_k, see Helland (1988). Subsequently, the regression problem $f = XW_k\alpha_k + \varepsilon$ is solved using XW_k as basis. This has as an advantage that both the variance in f and X as well as the correlation between f and X is taken into account, and therefore a basis W_k is generated such that by construction includes only components of X that are correlated to f and have sufficient variance to contribute to f.
More precisely, the idea is to find k orthogonal components $T_k = (t_1, \ldots, t_k)$, such that $t_i = Xw_i$ for some p-dimensional w_i. Thereby, weights w_i are chosen to maximize the empirical covariance between the data f and t_i. The first component $t_1 = Xw_1$ is found solving

$$w_1 = \arg\max_w \frac{\mathrm{cov}^2(Xw, f)}{w^t w} = \arg\max_w \frac{w^t X^t f f^t X w}{w^t w}, \tag{1}$$

which gives, up to a scalar, $w_1 = X^t f$. Further components t_i are found from the equation (1), subject to mutual orthogonality to all t_j, $j = 1, \ldots, i-1$, e.g.

$$t_i = Xw_i = XX^t \left\{ f - T_{i-1}(T_{i-1}^t T_{i-1})^{-1} T_{i-1} f \right\}, \ \ i = 2, \ldots, k.$$

Finally, for $W_k = (w_1, \ldots, w_k)$ and $T_k = XW_k$ the PLS estimator of order k for f is given by

$$\widehat{f}_{PLS}^k = XW_k\widehat{\alpha}_k = T_k(T_k^t T_k)^{-1} T_k f$$

Generalization of univariate PLS algorithm to a multivariate version for $f \in \mathbb{R}^{m \times n}$, $m < p$, has been carried out in Manne (1987) and is completely analogous to the univariate case, but (1) is replaced by

$$w_1 = \arg\max_{w,\, c} \frac{\mathrm{cov}^2(Xw, fc)}{w^t w c^t c},$$

that is, one is looking for the maximal empirical covariance between some linear combination fc and t_i. In the literature this algorithm is referred to as PLS2 algorithm.
Alternatively, we suggest to choose as a next direction the one, which has the maximal covariance to t_i, that is

$$w_1 = \arg \max_{w,\, i=1,\ldots,m} \frac{\mathrm{cov}^2(Xw, fe_i)}{w^t w},$$

where $e_i = (0, \ldots, 0, 1, 0, \ldots 0)$ is the zero vector with 1 at ith position. This multivariate algorithm will be referred to as PLS-Max algorithm.

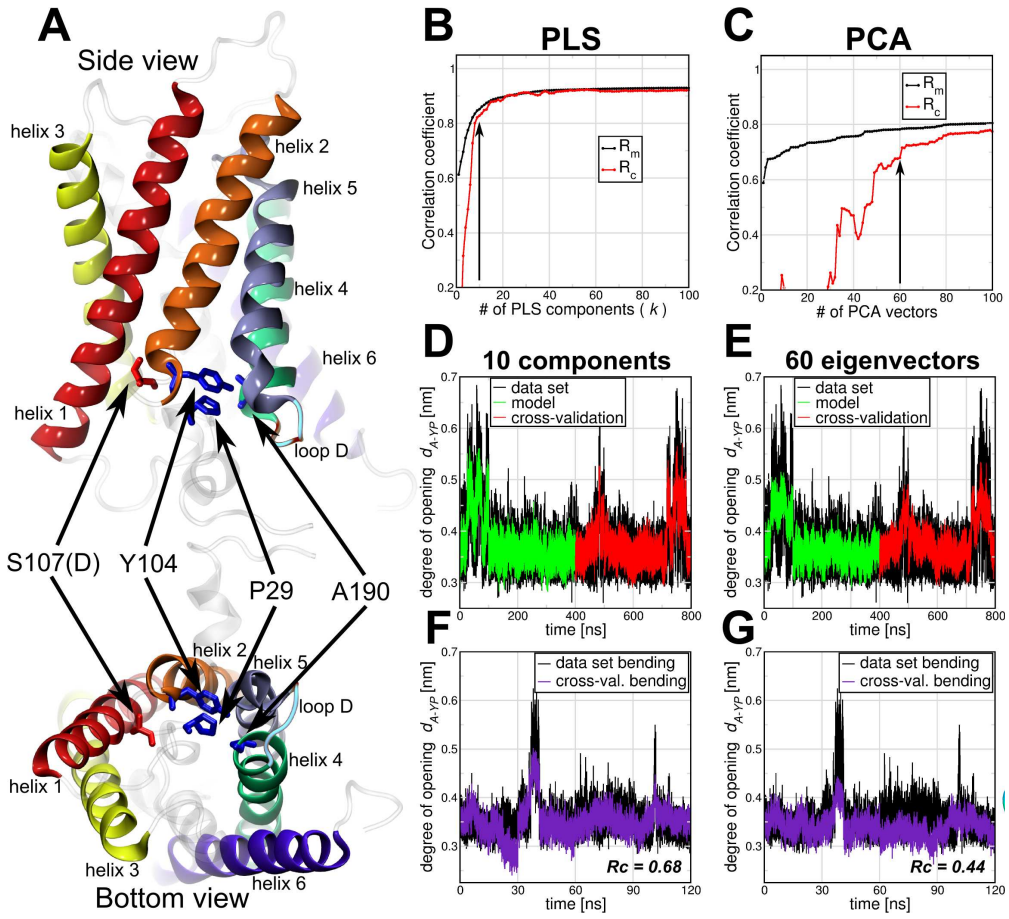

Figure 1: Comparison of PCA and PLS for Aquaporin data

It turns out, that the PLS-Max algorithm not only allows for a simpler (and sometimes more meaningful) interpretation of the results. The theoretical properties of PLS-Max appear to be more tractable.
Finally, it has been observed, that the columns in Y and f are serially correlated, namely possessing an AR(1) structure (autoregressive of order 1), which needs to be taken into account in the implementation of the algorithms.

3. Application

We demonstrate the application of PCA, PLS, PLS2 and PLS-Max to Aquaporin data. Figure 1 compares PCA and PLS for (A) the degree of channel opening of yeast Aquaporin (Aqy1). (B/C) Pearson correlation coefficients between data and model for PLS/PCA-based algorithm as a function of the number of PLS components/PCA vectors calculated for the model training subset (black, R_m) and the cross-validation subset (red, R_c). (D/E) Overlay of data and model for the calculated channel opening distance as function of time. The black lines correspond to the MD data, the green to the model training subset and red to the cross-validation subset. (F/G) 120 ns of a membrane bending simulation

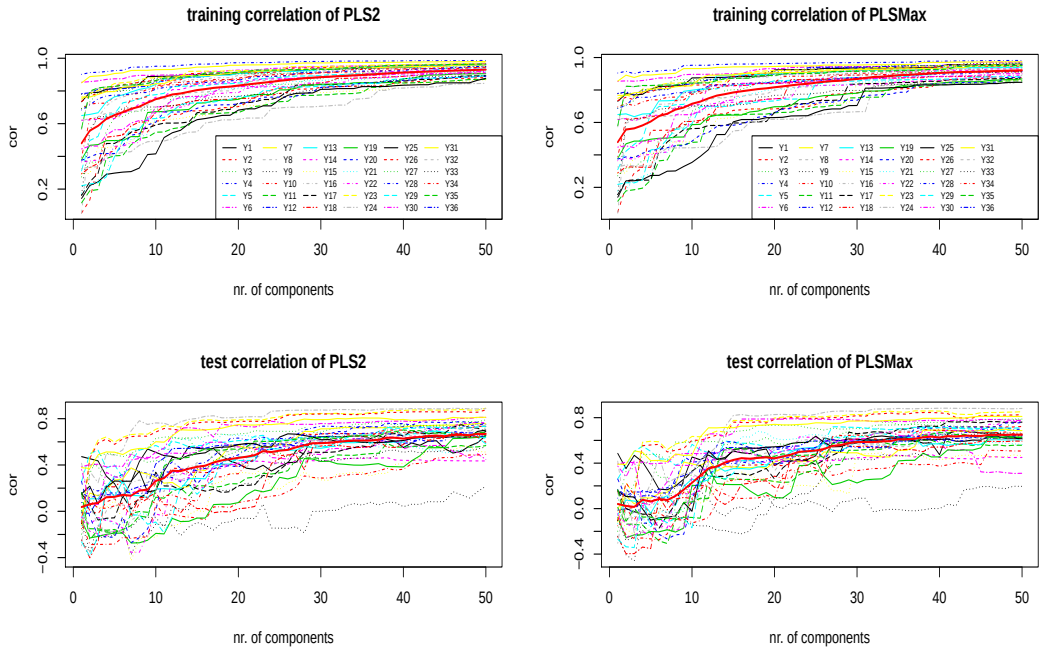

Figure 2: Comparison of PLS2 and PLS-Max for Aquaporin data

were used as an extra cross-validation sets (violet line).
Figure 2 compares two algorithms: PLS2 and PLS-Max, that relate 4 units of the Aquaporin protein to its atom dynamics ($f \in \mathbb{R}^{36}$). The upper plots show the Pearson correlation coefficients between data and model for PLS2/PLS-Max based algorithms as a function of the number of PLS components calculated for the model training subset, while bottom plots show Pearson correlation coefficients between data and the cross-validation subset.

References

Helland, I. S., 1988. On the structure of partial least squares regression. *Commun. Stat. Simulat.* 17:581607.

Krivobokova, T., Briones, R., Hub, J., Munk, A., de Groot, B. (2012) Partial least squares functional mode analysis: application to the membrane proteins AQP1, Aqy1 and CLC-ec1. *Biophys. J.*, 103(4): 786-796.

Manne, R. (1987) Analysis of two partial-least-squares algorithms for multi- variate calibration. *Chemometr. Intell. Lab. Syst.* 2: 187–197