Computational Molecular Biology and Bioinformatics

DLVPM

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

August, 2025

What is DLVPM?

The DLVPM is a deep latent variable path modeling technique [1], which combines the representational power of deep learning with the capacity of path modeling to identify relationships between interacting elements in a complex system.

DLVPM is a method for modeling dependencies between different data types. This method stands out for its ability to uncover complex, nonlinear interactions among both structured and unstructured data types, overcoming the limitations of traditional path-modelling techniques [2].

DLVPM is conceptually a generalization of Partial Least Squares approach to Path Modeling (PLS-PM) [2], which is conceptually a generalization of canonical correlation [3].

Canonical correlation

Recall that for a pair of random variables X and Y with finite second moments, the covariance is defined as the expected value (i.e., mean) of the product of their deviations from their individual expected values given by

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

For a pair of sets of random variables X' and Y' (i.e., random vectors each containing random elements whose expected value and variance exist), the cross-covariance matrix (Σ) is the matrix whose (i, j) entry is the covariance

$$\Sigma_{X_i Y_j} = cov[X_i, Y_j] = E[(X_i - E[X_i])(Y_i - E[Y_i])].$$

Canonical correlation

Patient	BP		SHAPE	
	Systolic BP	Diastolic BP	Height	Weight
1	120	76	165	60
2	109	80	180	80
3	130	82	170	70
4	121	78	185	85
5	135	85	180	90
6	140	87	187	87

Is there an association between **BP** (say X) and **SHAPE** (say Y)?

Canonical correlation

Let us calculate

$$\Sigma_x = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

The eigenvalue decomposition of Σ_x provides the weights to combine the variables in X.

Let us calculate

$$\Sigma_y = \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}.$$

The eigenvalue decomposition of Σ_y provides the weights to combine the variables in Y.

Based on this, we can find out the association between X and Y.

Canonical correlation

To state formally, the objective of canonical correlation can be written as follows:

$$max_{w_1,w_2} {w_1}^T X^T Y w_2$$

subject to the constraints $||Xw_1||_2^2 = 1$ and $||Yw_2||_2^2 = 1$.

It is possible to find multiple modes of variation using this method. Here the correlation between subsequent canonical variates is maximized subject to their being uncorrelated with other canonical variates. This can also be written as follows:

$$max_{W_1,W_2} \mathrm{tr}((XW_1)^T YW_2)$$

subject to the orthogonalization constraints $(XW_1)^T XW_1 = 1$ and $(YW_2)^T YW_2 = 1$, where W_1 and W_2 are $p_1 \times n_{dims}$ and $p_2 \times n_{dims}$ matrices, respectively; and I is an $n_{dims} \times n_{dims}$ identity matrix.

Generalized canonical correlation

We can optimize the sum of correlations between different data views. This involves maximizing the following criteria:

$$max_{W_1,W_2,\ldots,W_k} \sum_{i,j,i\neq j}^{K} \mathrm{tr}((X_i W_i)^T X_j W_j).$$

The generalized canonical correlation can be used to identify latent variables that are highly correlated between multiple data types. It may happen that we wish to identify associations between some—but not all—data types. Any model that attempts to link these data types may end up highlighting spurious effects.

Partial Least Squares approach to Path Modeling (PLS-PM)

Path-modeling/structural-equation-modeling techniques help to map dependencies between different data types. These methods are able to model arbitrarily many data types simultaneously, providing a holistic view of a system of interacting elements.

PLS-PM is designed to construct sets of latent variables that are optimally correlated between data types connected by a path model.

There are two major types of PLS-PM algorithm, namely mode A and mode B.

Partial Least Squares approach to Path Modeling (PLS-PM)

- **Mode A PLS-PM:** It involves optimizing the association between different data types. This approach requires the calculation of the matrix inverse of within-modality covariance matrices. This is not applicable when the number of examples in the data modality is smaller than the number of features.
- **Mode B PLS-PM:** It solves this issue by replacing within-modality covariance matrices with identity matrices.

Why deep learning?

Classical path modeling techniques (e.g., PLS-PM) are used to derive latent variables that exhibit optimal correlation among datasets linked by the path model. However, such techniques are limited to modeling linear effects.

Deep neural networks excel in their ability to model nonlinear effects, and to process structured and unstructured data.

Notations

Let us symbolize a neural network in the general form as follows:

$$\bar{Y}(X, U),$$

where $\bar{Y}$ is the network output, X is some data input and U is the set of network parameters (e.g., weights, biases, etc.).

In DLVPM, we define a set of submodels (indexed by i, termed as measurement models), each corresponding to one data type as follows:

$$\bar{Y}_i(X_i, U_i, W_i),$$

where $\bar{Y}_i$ is the network output (a set of deep latent variables, abbreviated as DLVs), X_i is some data input, U_i is the set of network parameters, and W_i corresponds to the network weights on the last layer of the neural network.

The DLVPM algorithm

The DLVPM algorithm is trained to construct DLVs from each measurement model, which are optimized to be maximally associated with DLVs from other measurement models, connected by the path model. These optimization criteria can be written as follows:

$$\max_{W_1,W2,\ldots,W_k,U_1,U_2,\ldots,U_k} \sum_{i,j,i\neq j} c_{ij}\mathrm{tr}((\bar{Y}_i(X_i,U_i,W_i))^T(\bar{Y}_j(X_j,U_j,W_j))),$$

where c_{ij} denotes the association matrix input from data type i to data type j. The DLVs derived from each data type are constrained to be orthogonal to one another as shown below:

$$\bar{Y}_i^T\bar{Y}_i = I, \forall i,$$

where I is the identity matrix.

The DLVPM algorithm

DLVPM can be formulated using different orthogonalization procedures. During training, the orthogonalization constraint is achieved via whitening or iterative orthogonalization.

- **Whitening:** It is a widely used approach to make the DLVs uncorrelated and of unit variance. It maintains a zero mean and identity covariance matrix of the DLVs.
- **Iterative orthogonalization:** It has the advantage that it prioritizes DLVs by their importance – a feature of considerable importance in the biological application presented here.

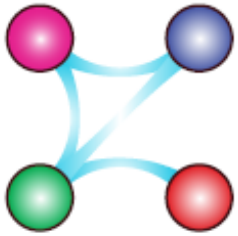
Removing the confounding effect

For removing the effect of confounding variables, a custom neural network layer that uses the Moore–Penrose pseudo-inverse of a matrix of nuisance covariates was incorporated.

This versatile layer represents a separate contribution from the main DLVPM method, and can be used in any neural network model.

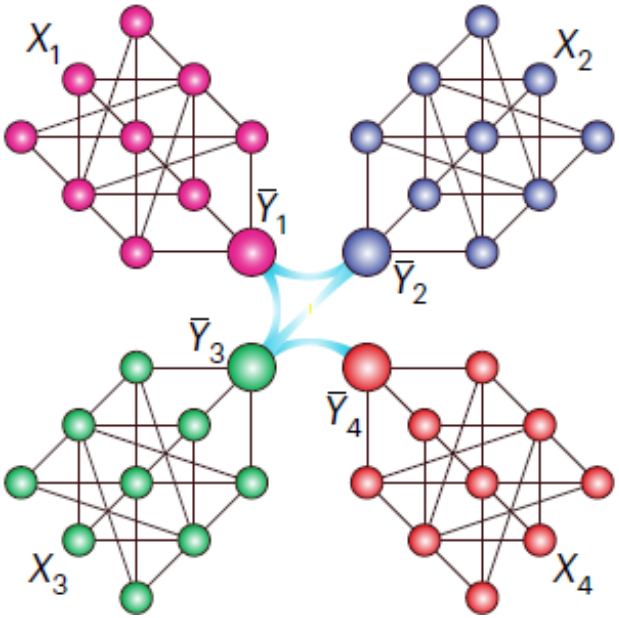
Schematic illustration of DLVPM

Path model



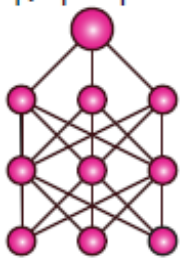
$$C = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Full model



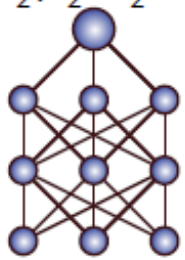
Measurement models

$\bar{Y}_1(X_1, U_1, W_1)$



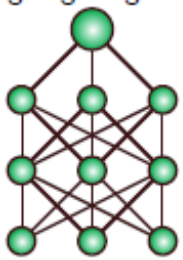
X_1

$\bar{Y}_2(X_2, U_2, W_2)$



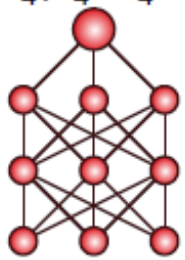
X_2

$\bar{Y}_3(X_3, U_3, W_3)$



X_3

$\bar{Y}_4(X_4, U_4, W_4)$

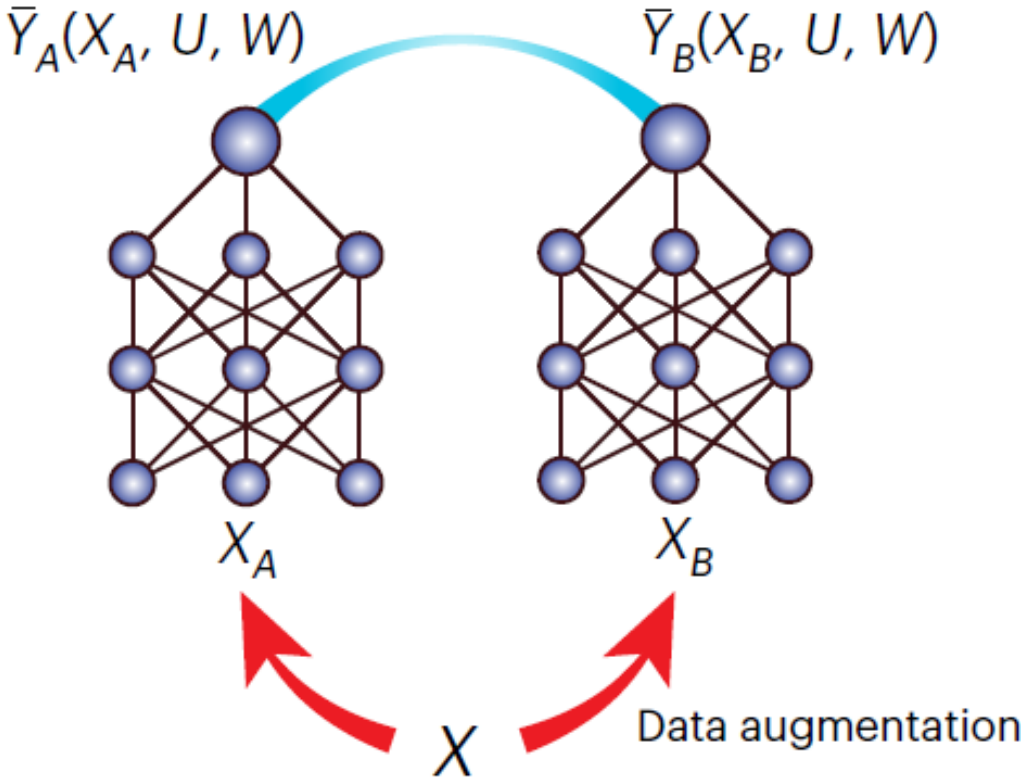


X_4

$$\max_{W_1,W_2\ldots W_K,U_1,U_2\ldots U_K} \sum_{i,j,i\neq j}^{K} c_{ij}\,\mathrm{tr}\left(\bar{Y}_i(X_i, U_i, W_i)^T \bar{Y}_j(X_j, U_j, W_j)\right)$$

$$\bar{Y}_i^T \bar{Y}_i = I \;\forall\, i$$

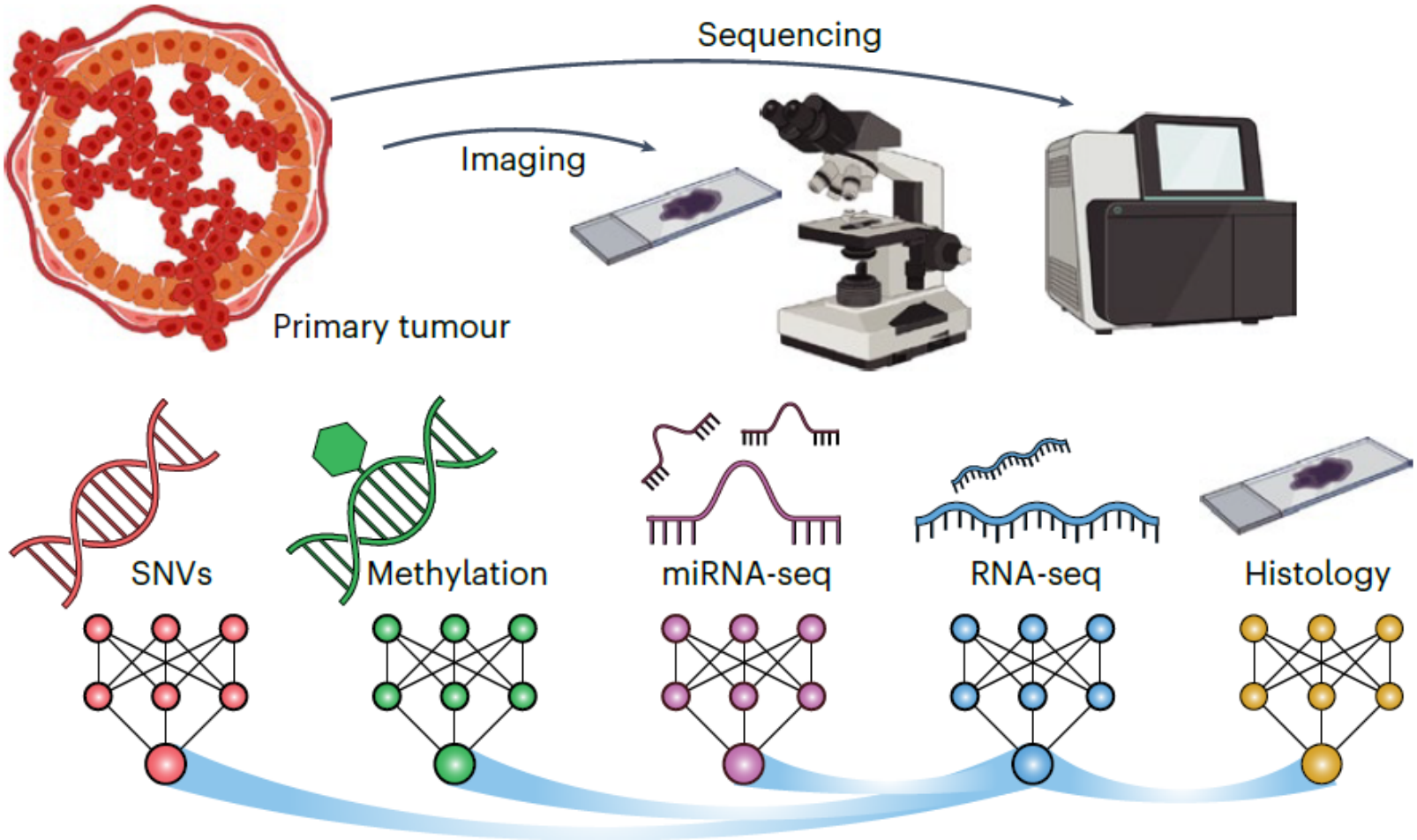
DLVPM in Siamese/twin network



$$\max_{U,W} \ \mathrm{tr}\left(\bar{Y}_A(X_A, U, W)^{\mathrm{T}} \bar{Y}_B(X_B, U, W)\right)$$

$$\bar{Y}_A^{\mathrm{T}} \bar{Y}_A = I \text{ and } \bar{Y}_B^{\mathrm{T}} \bar{Y}_B = I$$

Empirical Analysis



Graph representation of the path model

Empirical Analysis

DLVPM's training process is both iterative and end to end, enabling the model to learn directly from the raw data to the final output without the requirement for manual feature engineering.

Among the 758 TCGA breast cancer datasets:

- 80% ($n = 606$) were used for training.
- 20% ($n = 152$) were used for testing.

Concluding comments

- The DLVPM method is extremely versatile.
- As the measurement model formula hides a high level of generality and complexity, any kind of neural network can be used here.
- It can be used to create embeddings shared by feed-forward networks, convolutional networks, transformers, etc.

References

1. Ing, A., Andrades, A., Cosenza, M. R. and Korbel, J. O., Integrating multimodal cancer data using deep latent variable path modelling. Nature Machine Intelligence, 7:1053-1075, 2025.
   All code is publicly available on:
   `https://github.com/alexjamesing/Deep_LVPM`
2. Tenenhaus, M., Vinzi, V.E., Chatelin, Y.M. and Lauro, C., PLS path modeling. Computational Statistics Data Analysis, 48(1):159-205, 2005.
3. Hotelling, H., Relations between two sets of variates. In Breakthroughs in statistics: methodology and distribution (pp. 162-190). New York, NY: Springer New York, 1992.