Exploratory Partial Least Squares Models for use in Transcriptomic Analysis

Edward Tjörnhammar\textsuperscript{b,c}, Richard Tjörnhammar\textsuperscript{a}

\textsuperscript{a}KI Karolinska Institutet, SE-171 71 Stockholm, Sweden
\textsuperscript{b}KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden
\textsuperscript{c}FOI Swedish Defence Research Agency, SE-164 90 Stockholm, Sweden

Abstract

In this paper, we demonstrate the interpretative power of partial least square models and how robust interpretability can lead to new quantitative insights. Interpretability methods for partial least square decomposition are useful by virtue of PLS being popular in multivariate analysis, describing specificity of latent variables for corresponding effects.

We discuss the statistical properties of the PLS weights, \( p \) values associated with specific axes, as well as their alignment properties. We present hierarchical pathway enrichment results stemming from aligned \( p \) values, which are compared with results derived from enrichment analysis, as an external validation of our method. The applicability of this approach is further demonstrated by applying it to both publically available microarray data from multiple sclerosis and diabetes patients as well as RNA sequencing data from breast cancer patients.

Our method confers more global, physiologically relevant, knowledge about the datasets than traditional and more myopic methods.

Keywords: Partial Least Squares, Orthogonal Projections to Latent Structures, Statistical Methodology, Dimensionality Reduction Analysis, Descriptive Statistics, Bioinformatics

\textit{Email addresses: edwardt@kth.se, edward@foi.se} (Edward Tjörnhammar), \textit{richard.tjornhammar@ki.se} (Richard Tjörnhammar)
1. Introduction

Many fields within “Life Science” bear the suffix “omics”, such as transcriptomics (the study of transcripts), proteomics (the study of proteins), or metabolomics (the study of metabolites). These characteristically rely upon clinical trials for human dataset generation, often carrying empty, complementary or “messy” features as artifacts from their compilation. Trials are likewise expensive, creates “few samples” but “many features” and generative methods are dispatched with.

Orthogonal and ordinary Partial Least Squares (PLS) models Wold et al. (1984); Wegelin (2000) have a long history of usage in metabolomics Rothenberg et al. (2019); Eriksson et al. (2013), as well as other fields of research. The ability to decompose and align the feature space while communicating the effect of the decomposition on sample space is a well-known property of PLS models. The PLS methods have long been employed for characterisation of data features, but the interpretation often stops once the sample and feature groupings have been identified.

In this work, a pathway is a knowledge-based group descriptor of the transcripts. It should not be confused with the path diagram of the latent variables Wegelin (2000). Connecting the weights with axis p-values and calculating pathway enrichment becomes an important sanity check for this method. There are several knowledge-based pathway databases Szklarczyk et al. (2019); Kanehisa and Goto (2000); Jassal et al. (2019); Mi et al. (2018). These are curated with regards to known activity of transcripts and as such are not deduced from knowledge inherent in the data. In theory a new and properly well carried out experiment should be unknown to the knowledge database curator facilitating the group definitions. As such, if the alignment of weights to a descriptor axis is sensible then, the pathway enrichments should convey sensible enrichment for the groups that are under study.

In short for our purposes. Given the feature and sample space description matrices $(X, Y)$ are described by their corresponding weights $(w, c)$ and scores $(t = Xw, u = Yc)$. If the PLS algorithm has converged on a solution, by maximizing covariance between $t$ and $u$, then $w$ and $c$ will be aligned on the projected subspace.

In this work we propose a method to detect plausible research targets by an exploration of the sample space in high entropy domains through comparative plotting of cosine alignment on Orthogonal Projection to Latent Structure (OPLS) Trygg and Wold (2002) decomposition. Further, we
demonstrate that the PLS decomposition can be directly utilized for: (i) \( p \)-value calculations, (ii) pathway enrichment assessment, (iii) determining the suppression, or excitation state, of a pathway. We also demonstrate that weights associated with a simple binary one-hot encoded, multivariate axis are equivalent to calculating traditional fold changes between those groups.

The paper is structured as follows. Section 2 contextualizes the work. Section 3 motivates and describes the statistical viewpoint. Section 4 exemplifies analysis usage. A number of illustrating analysis examples related to the previously described method process are then presented in Section 5 in order to exemplify and validate the proposed solution. Finally, Sections 6 and 7 provide an assessment analysis, conclude from the undertaken experiment.

2. Related Work

Both Principle Component Analysis (PCA) \cite{Abdi2010} and PLS create new predictor variables often termed components. The components created by PCA maximize the variance in each without regard for the response. In PCA, we get a large number of components that aim to describe orthogonal variance. This affine transform can be further exploited to evaluate the statistical significance of weight relations. Similar to PCA, the PLS decomposes an input matrix into unique and complementary matrices. However, unlike PCA, which ultimately relies on factorising a sample centred expression matrix along the components of maximal orthogonal variance, the PLS method instead decomposes an input matrix by maximizing the mutual covariance between the feature space projection against the sample space response projection on the PLS predictor plane.

The subspace plane could be any dimensionality but is most often two dimensional such as in PLS regression \cite{Rosipal2006}. Because the weights are commonly computed via an iterative procedure it does not necessarily converge to a good solution, it however often does. Once the solution has been obtained then the variance in the dataset is often better described by fewer components than in PCA. It is not strange because the goal is to model the response by directly decomposing variance onto components describing the response. For this reason, a PLS algorithm is similar to a type-2 linear regression ANalysis Of VAriance (ANOVA) \cite{Fisher1919}. In both methods a two-dimensional tensor is constructed to describe the response and as such both methods are prone to similar issues and produce
similar statistical insights. We will see that, as opposed to a type-2 ANOVA
model, the PLS model also divulge information about the directionality of
the variation along an axis.

Tapp and Kemsley make a point that OPLS regression coefficients used in
ranked analysis do not take advantage of the multivariate nature of OPLS
Tapp and Kemsley (2009). This is correct even for our work, we project the de-
composition down onto a univariate vector to increase the robustness of our
model interpretations.

3. Method

For our adaptation of the OPLS method, we employ the same nomen-
clature as in the work by Rosipal and Krämer (2006). Given the subspace
vectors \( \mathbf{w} \perp \mathbf{t} \) we assume that any symmetry operations on \( \mathbf{w} \) will commute
to \( \mathbf{t} \). Habitually for our numerical experiments, we treat a set of gene expres-
sions as the feature matrix \( \mathbf{X} \) and the patient journal, with coarse-grained
end states, as the sample response matrix \( \mathbf{Y} \). As such \( \mathbf{X} \) either represent a
microarray- or a Fragments Per Kilobase of transcript per Million mapped
reads (FPKM) Ribonucleic Acid (RNA)-sequence matrix.

The formulation of groups is done by retrieving journal value entries for
a group present in the statistical formula enclosed with parentheses and a
leading ‘\( C \)’. Example

\[
\text{Transcript} \sim C(\text{Group})
\]

would retrieve the \( \text{Group} \) journal value entries.

Since PLS maximizes the pairwise covariance score, feature weights and
sample weights are “co-linearized” in a similar way. This lets us ascribe
features to sample weights, one way is to employ tessellation. Each transcript
is simply associated with the closest sample axis. This can be either absolute
coordinate proximity on the weight projection plane or the angular proximity
between the transcript and sample axis weight. The formula for calculating
the angular distance is given by the cosine similarity:

\[
d = 1 - \cos(\alpha) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||}
\]

Here \( \mathbf{a} \), a vector, designates a transcript weight position and \( \mathbf{b} \) a sample
score weight position, also a vector. The angle \( \alpha \) is the angle between \( \mathbf{a} \) and
\( \mathbf{b} \). In summary, we perform the steps:
1. Use a statistical test formula that defines the group variations to test.

2. Device an encoding data frame from journal group descriptor instances.

3. Utilize OPLS2 to find the solution to the formula using the expression matrix and patient (sample response) journal.

4. Categorical membership is determined through either tessellation (Euclidean distance) or angle, cosine distance to a centroid.

5. Calculate characteristic length scales of the prediction plane by finding the maximum weights.

6. Calculate the projected density of the weights onto the group axes defined by the statistical test formula.

7. Calculate $p$-values for the projections by assuming that the weights are distributed normally.

8. Calculate the ranking of the weights on the projected axis, mapped in ascending order from zero to one.

   This can be relied upon to visualize how features associate with samples. As a side note. Since ReactomePA [Jassal et al. (2019)] knowledge base is hierarchical, we also employed a hierarchical correction scheme similar to the elim algorithm as suggested by Alexa et al. (2006).

4. Implementation

   We have chosen to employ the OPLS2 algorithm as implemented in the publically available python [Pedregosa et al. (2011)] package and it also exists as an R [Geladi and Kowalski (1986); Höskuldsson (1988)] package.

   All categoricals present in the patient journal are translated into a one-hot encoding matrix. Thus all and any unique descriptors included in the group under study gets translated into a unique axis. All formula entries that are not categoricals are treated as real variables. The exact routine execution is covered in the supplementary information [1].

[1]Supporting information in experiments.py
The quantitative analysis relies on projecting the feature weight distribution onto an axis and studying the point density distribution on that axis. The \( p \) values are calculated from the cumulative error function for the density along the projection axis, see supplementary code Tjörnhammar (2019a). The type-2 ANOVA comparison is performed using the statsmodels Seabold and Perktold (2010) SciPy package.

5. Results

According to the method described in Section 3 we conducted three experiments:

- Multiple Sclerosis (MS) microarray dataset analysis, testing sample variation to group interaction. Also presenting the alignment properties of angular PLS model interpretation.

- The Cancer Genome Atlas (TCGA)-BReast CAncer Gene (BRCA) type-2 ANOVA, deducing the interaction model from the OPLS plot. Also comparing the analytical strength to type-2 ANOVA.

- Diabetes MICROARRAY dataset compounded analysis with clustering enrichment for genomic impact in Type 2 Diabetes (T2D).

The datasets for the analysis, as well as the group definition files, can all be found online Tjörnhammar (2019a). Details of the processing of the MS, as well as TCGA dataset, is also deposited online Tjörnhammar (2019b).

5.1. Multiple Sclerosis and the alignment properties of PLS models

The Multiple Sclerosis microarray data was obtained via the Gene Expression Omnibus (GEO) accession number for the microarray dataset GSE21942. 15 ‘Healthy’ controls and 14 ‘MS’ patients were processed and used in the analysis Kemppinen et al. (2011). The PLS model is conducting the test whether a transcript sample variation is well described by the between-group variation. We choose to express this as a statistical formula on the form

\[
\text{Transcript} \sim C(\text{Status})
\]

Here the Status label can take on the values Healthy or MS.

\(^2\text{http://www.ncbi.nlm.nih.gov/geo/}\)
(a) The OPLS transcript weight plane with weights coloured by alignment to MS state

(b) The OPLS sample scores with samples coloured by MS state

Figure 1: OPLS model of the MS state.

In Figure 1a, we can see how the transcripts become aligned to the disease state. A large number of transcripts are positioned close to origo and those that are farther away than 99% of all transcripts are highlighted with a higher opacity. From Figure 1b, it is clear that the OPLS model separates the cohort well. By comparing the left and right graph we see that the transcript weights are aligned so that we only have significant weights in the same quadrants as samples. The Y matrix, describing the MS and Healthy subjects sample score weights are also in the same quadrants. We can calculate the angular proximity between the sample score weights and the transcripts feature weights and assign what transcripts are closest to a particular score weight. This allows us to colour the transcripts according to the descriptor that best captures their variation in the data. We can also define new vectors corresponding to different directions on the transcript weight plane.

In Figure 2a, we depict the negative logarithm of the OPLS p-values on the y-axis and fold change values on the x-axis. In this case, the OPLS model describes the separation between Healthy and MS patients and the p values are thus analogue to t-test comparison based p values resulting in a volcano plot appearance. The axis describing the Healthy-Multiple Sclerosis state variation p values can then be employed to calculate pathway enrichments.
(a) The gene transcripts associated OPLS $p$ values and fold changes.

(b) Hierarchically corrected pathway enrichment.

Figure 2: Sizes corresponds to gene ratios and the significance levels depicted are raw $p$ values.

In Figure 2b, the pathway enrichment has been calculated using a Fisher exact test together with the Reactome pathway database. For each pathway, significant genes are deduced based on a $p$-value cutoff equal to 0.05 and a contingency table is produced. The resulting contingency table is then evaluated using a two-sided Fisher exact test. We choose a two-sided test because we employ a variational model. As such the enrichment does not correspond to an up or down-regulated accumulation in the group but rather correspond to if the group is important in describing the PLS axis.

In Figure 3 we study the gene associations calculated directly from the PLS feature weights. The feature associations can be calculated from the angle between two transcripts, $\beta$, on the transcript PLS weight plane. All pairs of $\cos(\beta)$ make up an association map for the PLS model. The values range from +1, for fully aligned, through 0, for orthogonal, to −1 for antialigned. The PLS solution constructs a model surface that maximizes the variance between the $Y$ axes. This is the reason why the association map represents an over accentuated correlation map of the genes depicted.

By selecting a smaller subset of transcripts, here with $q < 0.16$ values, and calculating the association map we see that the significant transcripts...
(a) The gene transcripts association map. Colors correspond to the cosine of the angle between the feature weights.

(b) Traditional quantification of the transcripts with most significant group differences.

Figure 3: Gene transcription association exploration.

are roughly divided into two anti-associated clusters. Under the assumption that our coding genes produce proteins, we may utilize string-db Szklarczyk et al. (2019) to analyse them. We find that they enrich for immune system activity in the specific and tertiary granule lumens. The largest cluster enriches for antifungal humoral response and innate immune response in mucosa while the second cluster enriches for publications on long non-coding RNA activity. In the second cluster, the gene with the highest intersection among such publications is the PSMD5 transcript. The protein of this gene has also been reported for proteomic quantification of differentially expressed genes between MS and Healthy patients Berge et al. (2019).

5.2. TCGA-BRCA and the comparison with a type-2 ANOVA

In this section we study the TCGA Breast cancer dataset Koboldt et al. (2012) under OPLS and a type-2 ANOVA. Clinical parameters Tao et al. (2019), as well as messenger Ribonucleic Acid (mRNA) expression data from human female patients. Theses were obtained from the TCGA web-site Koboldt et al. (2012). All of the downloaded samples originated from
(a) Transcript weight colored by their hormone alignment owners.  

(b) The sample scores show that triple negative patients are diametrically opposed to HER2- & PR+, ER+ patients.

Figure 4: OPLS model of the TCGA breast cancer data set.

Illumina RNA-Seq data using the High Throughput Sequencing (HTS) 3 FPKM workflow [Anders et al. (2015)]. In the case where more than one expression profile belonged to one sample id we randomly selected one of the profiles. Sample receptor statuses [Ross-Innes et al. (2012)] were stored as ER+, ER-, PR+, PR-, HER2+, HER2- or ND in a sample journal file. This resulted in 1184 breast cancer samples [Tjörnhammar (2019b)].

In Figure 4 we employed an OPLS model of the form

\[ \text{Transcript} \sim C(\text{HER2}) : C(\text{ERPR}) + C(\text{HER2}) + C(\text{ERPR}) \]

and coloured all the group instances with unique colours. The HER2 group describes the HER2 receptor state while the Estrogen Receptor Progesterone Receptor (ERPR) group describes whether or not the ER and PR receptors are in a particular state. Thus triple-negative patients will belong to the HER2- & PR-, ER- intergenetion and triple positive to the HER2+ & PR+, ER+. It is well known that triple-negative patients have a poor survival outcome for breast cancer and are interesting to study in this context. Figure 4b shows that triple-negative patients amass in the most southern parts of the lower quadrants. Diametrically opposed we find the patients bearing the interac-
(a) Association map of transcripts describing the variation along the triple negative axis

(b) Box plots of most significant transcripts HER2-, PR+, ER+ patients exhibit starkest differences towards triple negatives

Figure 5: OPLS model association map of the TCGA breast cancer data set.

... continuation combination belonging to HER2- & PR+, ER+. We note that the triple-positive group form an angle of roughly 110 degrees to the triple-negative group. From here on the call the triple-negative towards HER2-, PR+, ER+ axis the high-risk axis.

Among the genes with the strongest association to the high-risk axis, we find ESR1, CXXC5, GREB1, FOXA1 and GATA3. They enrich Szklarczyk et al. (2019) for estrogen-dependent gene expression in Reactome as well as the prostate gland and uterus development in Biological processes Mi et al. (2018). We only have women in the data so we safely assume that the prostate gland enrichment is due to the interaction of these genes with the progesterone receptor in men.

In Figure 5b, we find the traditional quantification of the transcript data. Triple-negative patients are remarkably different from the other groups. We also show both triple-positive as well as HER2-, PR+ and ER+ patient groupings. The PLS model in Figure 4b conveyed the information that HER2-, PR+, ER+ combination are anti associated to the triple negatives and we can also
(a) The \( p \) values produced by the PLS (b) OPLS model with all transcripts having type-2 ANOVA \( q < \frac{0.05}{10455} \) shown with full opacity.

see that the biggest difference in the boxplots of Figure 5b is found between the triple-negative and HER2-, PR+, ER+ group. It is also becomes apparent that the gene cluster associated with being up in triple-negative patients belong to PSAT1, TTL4, FOXC1, USB1, PDSS1, YBX1, KCMF1, YBX1P10 and DSC2. It constitutes a novel cluster for describing triple-negative patients and it is only when evaluated along the high-risk axis that the full picture emerges.

These results would not be complete without mentioning the statistical properties of the weights. Since there is no warranty that PLS models converge on any specific residual distribution this can become convoluted. The maximisation of variance on the common subspace plane implies that a solution would minimize the PLS models squared residual. In using our formalism we can study the same formula using both an ANOVA as well as an OPLS model. In Figure 6b we show the same statistical model,

\[
\text{Transcript} \sim C(\text{HER2}) : C(\text{PR}) + C(\text{HER2}) + C(\text{ER})
\]

produces qualitatively similar results.

In Figure 6a depicts the OPLS model where all the interaction categories of

\[
\text{Transcript} \sim C(\text{HER2}) : C(\text{PR})
\]

are shown. The transcripts are also modelled using a type-2 ANOVA and all
transcripts having an interaction with a \( q < \frac{0.05}{10^{77}} \) are depicted in full opacity. Two realizations are immediate. The first is that the ANOVA results are qualitatively similar to those of the OPLS, with a large density of significant ANOVA transcripts being found on the rim of the OPLS weight plane. Our OPLS model, however, offers an improvement in that we may determine not just the specific interaction pair associated with which transcript but also the direction. From our model, we immediately know that \textit{TTL4} will be scaling in magnitude with the triple-negative state as well as being significantly different from the other groupings. The traditional box plot quantification in Figure 5b is important but not essential. We further explore our PLS \( p \)-values and it is clear from Figure 6a that we have a large accumulation of weights near the centre. This also means that we are more prone to false discoveries near the centre. One can directly correct for this in the \( q \) value calculation by assuming that the False Discovery Rate (FDR) scale linearly with the Fractional Rank (FR) of the list of \( p \) values, which is quantified by:

\[
\text{fr}(p) = \frac{\text{ranks}(p)}{\text{length}(p)}
\]

Instead of assuming a constant FDR [Storey and Tibshirani (2003)] across \( p \) values, often set equal to 1, for the \( q \) value calculation we set it to scale using the FR formula. These scaled \( p \) values are here dubbed \( o \) values and are only used to show the effect of this scaling. The final \( q \) values are showed in Figure 6b. We recognize that the OPLS transcripts with significant interaction all belong to the rim of the categories defined by the axis studied (here the far edge of the HER2+, PR-, ER- axis).

5.3. Diabetes Microarray Analysis

Type 2 Diabetes is a metabolic disease on the rise in the general populous. It is furthermore influenced by both lifestyle choices as well as genetic factors. Understanding the detailed balance required for this disease phenotype to emerge is important to achieve better therapies and quality of life of those affected by it. The understanding of altered metabolism is also of great importance since many diseases are associated with altered cell and tissue metabolism [Mootha et al. (2003)].

There is an increasing amount of evidence that type 2 diabetes T2D, or Diabetes Mellitus 2 (DM2), disease phenotype constitutes several underlying metabolic disease phenotypes [Ahlqvist et al. (2018)].
Here we have chosen to study human DM2 data from the Broad Institute [Mootha et al. (2003)]. Suspecting that we have more than a single type of DM2 patients we first segmented the data. This was done by employing approximative density clustering as implemented in our package [Tjörnhammar (2019a)]. Because of the limited number of samples (18) we stopped once we had obtained the two most separated sample groups. We thereby obtain two well-separated DM2 cluster types that we name $C_{1,DM2}$ and $C_{2,DM2}$ where the cluster index is the inverted cluster size rank.

In Figure 7 we see that the second cluster consists of only three samples. They form a group well-separated from the rest but lying close to orthogonal to the NGT group. The first cluster forms an axis with the NGT group with weights being positioned opposed to one another. Using our OPLS model we would believe that they constitute a smaller DM2 phenotype subset with distinct metabolic alterations. By studying the $p$-values associations on the rim of the entire OPLS graph we find that $FAM35A$ have $q < \frac{0.05}{N_{transcripts}}$ and aligning with the $C_{1,DM2}$ axis. $FAM35A$ is known to be a favourable prognostic marker of glioma and colo-rectal cancer [Uhlén et al. (2015)]. By allowing ourself to study genes with $q < 0.005$ we obtain the association map in Figure 8a.

In Figure 8 we can clearly discern four transcript clusters. The first two correspond to transcripts that are positively associating with being healthy.
(a) All the edge weight associations below $q < 0.005$.

(b) The association map of the largest DM2 group and NGT variation, $q < 0.0001$.

Figure 8: Complete OPLS association maps.

Those transcripts enrich for Cyclin C activity, which is known for controlling nuclear cell division Galderisi et al. (2003). The third cluster is positioned near the centre of the association map on the left and is close to orthogonal to the other clusters.

This cluster consists of transcripts that enrich Szklarczyk et al. (2019) for striated muscle contraction Jassal et al. (2019) and organ morphogenesis Mi et al. (2018). Most of the enrichment signal is, however, coming from the MYL2, ACTA1, CLDN1, POU4F1, CRX and TNF group. This third cluster belongs to the C2.DM2 group indicating that these three individuals, albeit being diabetic, might have dramatically different muscle metabolism.

The first and biggest cluster belong is the C1.DM2 samples visible in Figure 8. Four transcripts reach high significance and together enrich for ATP binding activity. The 9 high significance transcripts aligning with this axis enrich Szklarczyk et al. (2019) for DeoxyriboNucleic Acid (DNA) repair and DNA metabolism. Taken together with the full C1.DM2-NGT axis transcripts enrich for activity in the condensed nuclear chromosome. By allowing ourselves to be less strict and studying enrichments at lower confidence corresponding to the C1.DM2-NGT transcripts in Figure 8a.
6. Discussion

As a general remark on both our results from Section 5.1 and Section 5.2 is that our \( p \) values are larger than the typical two-way ANOVA \( p < 0.05 \), which could indicate that our results are conservative as compared to a two way ANOVA discovery for the same gene.

6.1. Multiple Sclerosis

From Figure 2 in Section 5.1 we see that the major pathways deduced are Complex I biogenesis and Neutrophil granulation, pathways which are affiliated with increased metabolism and immune system activity. Compartment enrichment analysis, also utilizing a two-sided Fisher exact test, of our \( p \) values tells us that the dataset is mainly active in the cytosol, nucleoplasm and ficolin-1-rich granule membrane. This conveys the picture that we are dealing with neutrophil degranulation and metabolism of surface membranes. This is aligned with the common view that MS is associated with demyelination of axon myelin. Roughly translating into altered lipid metabolism by the immune system.

Our modelling scheme shows weaker statistical power than other methods but recovers fundamental and broad descriptive knowledge about the dataset. We also identify a small subset of significant genes that are differentially expressed in the cohort and share an intersection with proteomic analysis. Both the proteomic derived set as well as our PLS transcriptome analysis does not share a common intersection with the GWAS study published together with the dataset that we employed. This analysis method should as such be viewed as a complementary method for producing a data-driven story of the cohort as well as targets of interest for use in further biological validation. In this section, we have shown how the \( Y \) matrix alignments translate into the feature space of PLS models.

Looking at Figure 3a and allowing ourself down into the first 100 significant PLS genes, with a \( p < 0.0002 \) value, we see a common intersection set between our results and the Berge et al. (2019) proteins. These are TES, CLTC, SIN3A as well as PSMD5 and known to influence RUNX1 transcription pathways.

6.2. Breast Cancer

The Forkhead box (FOX) C1 transcript is thought to be a possible target in nasopharyngeal carcinoma Ou-Yang et al. (2015). FOXA1 is similarly known
to be correlated with \textit{GATA3}, Progesteron receptor protein expression and clinical outcomes in breast cancer \cite{Ross-Innes2012} as well as prostate cancer \cite{Barbieri2012}. However, some of the associations in Figure 5a are new and might be interesting targets for further biological validation. It is worth further emphasis that these gene transcripts correspond to the outliers lying on the axis connecting the triple-negative and the HER2-, PR+, ER+ in the OPLS feature weight plane.

6.3. \textit{Type 2 Diabetes}

Hierarchical modelling or agglomerative hierarchical clustering are well-known methods of partitioning one’s data. Traditionally, building hierarchical models are done via curating knowledge and manually constructing a hierarchy \cite{Jassal2019, Ahlqvist2018}. One can also choose to computationally construct hierarchical data structures using agglomerative hierarchical clustering \cite{Mehta2019} or simply try to find the grouping that will separate the density the most \cite{Tjornhammar2019}. Approximate density clustering is similar to agglomerative hierarchical clustering but instead of building a distance hierarchy we try to find the best sample grouping describing the sample space.

All other things being equal we can compare Figure 7 to Figure 9 and see that OPLS, by its very nature, better separates the data. In Figure 9
we see that the three samples who have their own cluster in Figure 7 also correspond to outliers when employing PCA.

Guided by Figure 8 we see an enrichment Szklarczyk et al. (2019) for Guanosine TriPhosphate (GTP) metabolism in the mitochondria. We know that mitochondrial GTP metabolism is essential for glucose-dependent insulin secretion Kibbey et al. (2007). It is known that this will lead to disrupted oxidative phosphorylation and an altered mitochondrial membrane potential Mehta et al. (2019). Furthermore, mitochondria maintain oxidative phosphorylation by creating a membrane potential, also implying an altered membrane potential will lead to altered oxidative phosphorylation Momcilovic et al. (2019). This together is indicative that Adenosine TriPhosphate (ATP) synthesis might be inhibited, for the diabetics, in the mitochondria Sivitz and Yorek (2010). These findings are qualitatively similar to those reported by Mootha et al. (2003). This is indeed good news since we are studying people whom, by definition, have inhibited metabolic regulatory control of glucose and insulin.

7. Conclusions

In and for this article, we have developed a new analysis method, in Section 3 and code library, as described in Section 4, for use in bioinformatic analysis. We have demonstrated the use of OPLS models for transcriptomic analysis on three different mRNA data sets, MS (Section 5.1), TCGA-BRCA (Section 5.2) and T2D (Section 5.3). Our models produce unbiased sensible descriptions of the endpoint endogen states where we can relate what we see in the data to known properties of the human patient samples that we studied, as discussed in Section 6. We have shown how the alignment of the sample to feature scores translate into the ability to interpret feature weights in the context of the sample weights. As such the realization that the weights of the feature and sample descriptor matrices are aligned facilitates the use of the feature weights directly in traditional pathway analysis methods. This also means that a high weight in the direction of a predictor axis will equate to large, between-group, difference quantifications lying on that axis.

From this, several new and sensible interpretations of the data emerge. We can also confirm previously known properties of the samples. We find a gene cluster which follows the triple-negative breast cancer patients which have interesting suggestion targets for continued validation.
PLS models are employed in such diverse research fields as financial forecasting Huang and Wu (2010), brain imaging Abdi (2010) as well as environmental spatial characterization and forecasting Thioulouse et al. (1995). We believe as such that the method is suitable as decision support for providing global suggestions of future study in high entropy domains. To our knowledge, looking at other PLS model applications, no one explores the orientation, or vector alignment, properties of such models.
References

H. Abdi. Partial least squares regression and projection on latent structure regression (PLS Regression): PLS REGRESSION. Wiley Interdisciplinary Reviews: Computational Statistics, 2(1):97–106, jan 2010. ISSN 19395108. doi: 10.1002/wics.51.

H. Abdi and L. J. Williams. Principal component analysis: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4):433–459, jul 2010. ISSN 19395108. doi: 10.1002/wics.101.

E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, Å. Lernmark, K. Lahti, T. Forsén, T. Tuomi, A. H. Rosengren, and L. Groop. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology, 6(5):361–369, may 2018. ISSN 22138587. doi: 10.1016/S2213-8587(18)30051-2.

A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 22(13):1600–1607, jul 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl140.

S. Anders, P. T. Pyl, and W. Huber. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2):166–169, jan 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu638.

C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, N. Stransky, E. Nickerson, S.-S. Chae, G. Boysen, D. Auclair, R. C. Onofrio, K. Park, N. Kitabayashi, T. Y. MacDonald, K. Sheikh, T. Vuong, C. Guiducci, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, W. M. Hussain, A. H. Ramos, W. Winckler, M. C. Redman, K. Ardlie, A. K. Tewari, J. M. Mosquera, N. Rupp, P. J. Wild, H. Moch, C. Morrissey, P. S. Nelson, P. W. Kantoff, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, G. Getz, M. A. Rubin, and L. A. Garraway. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate...
cancer. *Nature Genetics*, 44(6):685–689, jun 2012. ISSN 1546-1718. doi: 10.1038/ng.2279.

T. Berge, A. Eriksson, I. S. Brorson, E. A. Høgestøl, P. Berg-Hansen, A. Døskeland, O. Mjaavatten, S. D. Bos, H. F. Harbo, and F. Berven. Quantitative proteomic analyses of CD4(+) and CD8(+) T cells reveal differentially expressed proteins in multiple sclerosis patients and healthy controls. *Clinical proteomics*, 16:19–19, may 2019. ISSN 1542-6416. doi: 10.1186/s12014-019-9241-5.

L. Eriksson, T. Byrne, E. Johansson, J. Trygg, and C. Vikström. *Multi- and Megavariate Data Analysis Basic Principles and Applications, Third Revised Edition*. Umetrics Academy, 2013. ISBN 91-973730-5-2.

R. A. Fisher. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. ISSN 0080-4568, 2053-5945. doi: 10.1017/S0080456800012163.

U. Galderisi, F. P. Jori, and A. Giordano. Cell cycle regulation and neural differentiation. *Oncogene*, 22(33):5208–5219, aug 2003. ISSN 1476-5594. doi: 10.1038/sj.onc.1206558.

P. Geladi and B. R. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, jan 1986. ISSN 0003-2670. doi: 10.1016/0003-2670(86)80028-9.

A. Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2(3):211–228, jun 1988. ISSN 0886-9383. doi: 10.1002/cem.1180020306.

S.-C. Huang and T.-K. Wu. Integrating recurrent SOM with wavelet-based kernel partial least square regressions for financial forecasting. *Expert Systems with Applications*, 37(8):5698–5705, aug 2010. ISSN 09574174. doi: 10.1016/j.eswa.2010.02.040.

B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D’Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, (gkz1031), nov 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz1031.
M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, jan 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.27.

A. K. Kemppinen, J. Kaprio, A. Palotie, and J. Saarela. Systematic review of genome-wide expression studies in multiple sclerosis. *BMJ Open*, 1(1), 2011. ISSN 2044-6055. doi: 10.1136/bmjopen-2011-000053.

R. G. Kibbey, R. L. Pongratz, A. J. Romanelli, C. B. Wollheim, G. W. Cline, and G. I. Shulman. Mitochondrial GTP Regulates Glucose-Stimulated Insulin Secretion. *Cell Metabolism*, 5(4):253–264, apr 2007. ISSN 1550-4131. doi: 10.1016/j.cmet.2007.02.008.

D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael, L. L. Fulton, D. J. Dooling, L. Ding, E. R. Mardis, R. K. Wilson, A. Ally, M. Balasundaram, Y. S. N. Butterfield, R. Carlsen, C. Carter, A. Chu, E. Chuah, H.-J. E. Chun, R. J. N. Coope, N. Dhall, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, A. J. Mungall, E. Pleasance, A. Gordon Robertson, J. E. Schein, A. Shafiei, P. Sipahimalani, J. R. Slobodan, D. Stoll, A. Tam, N. Thiessen, R. J. Varhol, N. Wye, T. Zeng, Y. Zhao, I. Birol, S. J. M. Jones, M. A. Marra, A. D. Cherniack, G. Saksena, R. C. Onofrio, N. H. Pho, S. L. Carter, S. E. Schumacher, B. Tabak, B. Hernandez, J. Gentry, H. Nguyen, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, S. B. Gabriel, M. Meyerson, L. Chin, P. J. Park, R. Kucherlapati, K. A. Hoadley, J. Todd Auman, C. Fan, Y. J. Turman, Y. Shi, L. Li, M. D. Topal, X. He, H.-H. Chao, A. Prat, G. O. Silva, M. D. Iglesias, W. Zhao, J. Usary, J. S. Berg, M. Adams, J. S. Parker, D. Neil Hayes, C. M. Perou, S. Malik, S. Mahurkar, H. Shen, D. J. Weisenberger, T. Triche Jr, P. H. Lai, M. S. Bootwalla, D. T. Maglinte, B. P. Berman, D. J. Van Den Berg, S. B. Baylin, P. W. Laird, C. J. Creighton, L. A. Donehower, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C.-J. Wu, S. Yingchun Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, R. Sinha, R. W. Park, M.-D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, S. Reynolds, R. B. Kreisberg, B. Bernard, R. Bressler, T. Erkkila, J. Lin, V. Thorsson, W. Zhang, I. Shmulevich,
G. Ciriello, N. Weinhold, N. Schultz, J. Gao, E. Cerami, B. Gross, A. Jacobsen, R. Sinha, B. Arman Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, M. Ladanyi, C. Sander, P. Anur, P. T. Spellman, Y. Lu, W. Liu, R. R. G. Verhaak, G. B. Mills, R. Akbani, N. Zhang, B. M. Broom, T. D. Casasent, C. Wakefield, A. K. Unruh, K. Baggerly, K. Coombes, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Zhu, C. C. Szeto, G. K. Scott, C. Yau, E. O. Paull, D. Carlin, C. Wong, A. Sokolov, J. Thusberg, S. Mooney, S. Ng, T. C. Goldstein, K. Ellrott, M. Grifford, C. Wilks, S. Ma, B. Craft, C. Yan, Y. Hu, D. Meerzaman, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. D. Black, R. E. Pyatt, P. White, E. J. Zmuda, J. Frick, T. M. Lichtenberg, R. Brookens, M. M. George, M. A. Gerken, H. A. Harper, K. M. Leraas, L. J. Wise, T. R. Tabler, C. McAllister, T. Barr, M. Hart-Kothari, K. Tarvin, C. Saller, G. Sandusky, C. Mitchell, M. V. Iacooca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, O. Dolzhansky, M. Abramov, O. Voronina, O. Potapova, J. R. Marks, W. M. Suchorska, D. Murawa, W. Kycler, M. Iibs, K. Korski, A. Spychala, P. Murawa, J. J. Brzeziński, H. Perz, R. Łaźniak, M. Teresiak, H. Tatka, E. Leporowska, M. Bogusz-Czerniewicz, J. Malicki, A. Mackiewicz, M. Wiznerowicz, X. Van Le, B. Kohl, N. Viet Tien, R. Thorp, N. Van Bang, H. Sussman, B. Duc Phu, R. Hajek, N. Phi Hung, T. Viet The Phuong, H. Quyet Thang, K. Zaki Khan, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, J. N. Ingle, F. J. Couch, W. L. Lingle, T. A. King, A. Maria Gonzalez-Angulo, G. B. Mills, M. D. Dyer, S. Liu, X. Meng, M. Patangan, The Cancer Genome Atlas Network, Genome sequencing centres: Washington University in St Louis, Genome characterization centres: BC Cancer Agency, Broad Institute, Brigham & Women’s Hospital & Harvard Medical School, C. H. University of North Carolina, University of Southern California/Johns Hopkins, Genome data analysis: Baylor College of Medicine, Institute for Systems Biology, Memorial Sloan-Kettering Cancer Center, Oregon Health & Science University, The University of Texas MD Anderson Cancer Center, S. C. I. University of California, NCI, Biospecimen core resource: Nationwide Children’s Hospital Biospecimen Core Resource, Tissue source sites: ABS-IUPUI, Christiana, Cureline, Duke University Medical Center, The Greater Poland Cancer Centre, ILS-Bio, International Genomics Consortium, Mayo Clinic, MSKCC, and MD Anderson Cancer Center. Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61–70, oct 2012. ISSN 1476-4687. doi: 10.1038/nature11412.
P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *A high-bias, low-variance introduction to Machine Learning for physicists*, 810:1–124, may 2019. ISSN 0370-1573. doi: 10.1016/j.physrep.2019.03.001.

H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47(D1):D419–D426, nov 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1038.

M. Momcilovic, A. Jones, S. T. Bailey, C. M. Waldmann, R. Li, J. T. Lee, G. Abdelhady, A. Gomez, T. Holloway, E. Schmid, D. Stout, M. C. Fishbein, L. Stiles, D. V. Dabir, S. M. Dubinett, H. Christofk, O. Shirihihai, C. M. Koehler, S. Sadeghi, and D. B. Shackelford. In vivo imaging of mitochondrial membrane potential in non-small-cell lung cancer. *Nature*, 575(7782):380–384, nov 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1715-0.

V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstäis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, jul 2003. ISSN 1546-1718. doi: 10.1038/ng1180.

L. Ou-Yang, S.-J. Xiao, P. Liu, S.-J. Yi, X.-L. Zhang, S. Ou-Yang, and S.-K. Tan. Forkhead box C1 induces epithelial-mesenchymal transition and is a potential therapeutic target in nasopharyngeal carcinoma. *Molecular Medicine Reports*, 12(6):8003–8009, oct 2015. doi: 10.3892/mmr.2015.4427.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
R. Rosipal and N. Krämer. Overview and Recent Advances in Partial Least Squares. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, pages 34–51, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34138-3.

C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, S. Ali, S.-F. Chin, C. Palmieri, C. Caldas, and J. S. Carroll. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481 (7381):389–393, jan 2012. ISSN 1476-4687. doi: 10.1038/nature10730.

D. O. Rothenberg, H. Yang, M. Chen, W. Zhang, and L. Zhang. Metabolome and Transcriptome Sequencing Analysis Reveals Anthocyanin Metabolism in Pink Flowers of Anthocyanin-Rich Tea (Camellia sinensis). *Molecules*, 24(6), 2019. ISSN 1420-3049. doi: 10.3390/molecules24061064.

S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

W. I. Sivitz and M. A. Yorek. Mitochondrial dysfunction in diabetes: From molecular mechanisms to functional significance and therapeutic opportunities. *Antioxidants & redox signaling*, 12(4):537–577, apr 2010. ISSN 1557-7716. doi: 10.1089/ars.2009.2531.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. ISSN 0027-8424. doi: 10.1073/pnas.1530509100.

D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. von Mering. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, jan 2019. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gky1131.

M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, Y. Wang, and Z. Yang. Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. *Genes*, 10(3), mar 2019. ISSN 2073-4425 2073-4425 2073-4425. doi: 10.3390/genes10030200.
H. S. Tapp and E. K. Kemsley. Notes on the practical utility of OPLS. *TrAC Trends in Analytical Chemistry*, 28(11):1322–1327, dec 2009. ISSN 01659936. doi: 10.1016/j.trac.2009.08.006.

J. Thioulouse, D. Chessel, and S. Champely. Multivariate analysis of spatial patterns: A unified approach to local and global structures. *Environmental and Ecological Statistics*, 2(1):1–14, mar 1995. ISSN 1573-3009. doi: 10.1007/BF00452928.

R. Tjörnhammar. Impetuous-gfa, 2019a.

R. Tjörnhammar. MS and TCGA dataset, sep 2019b.

J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, mar 2002. ISSN 0886-9383, 1099-128X. doi: 10.1002/cem.695.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szegyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hoer, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, jan 2015. doi: 10.1126/science.1260419.

J. A. Wegelin. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case, mar 2000.

S. Wold, C. Albano, W. J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjöström. Multivariate Data Analysis in Chemistry. In B. R. Kowalski, editor, *Chemometrics: Mathematics and Statistics in Chemistry*, pages 17–95. Springer Netherlands, Dordrecht, 1984. ISBN 978-94-017-1026-8. doi: 10.1007/978-94-017-1026-8_2.