Using **HDTD** to Analyze High-Dimensional Transposable Data: An Application in Genetics

Anestis Touloumis*, John C. Marioni and Simon Tavaré

1 Introduction

The R/Bioconductor package **HDTD** is designed to analyze high-dimensional transposable data. The term transposable data implies the following structural information in the dataset:

- the data for each sampling unit (e.g., subject/patient) can be written in a matrix,
- the rows and the columns in each matrix correspond to two distinct sets of variables,
- dependencies might occur among and/or between the row and column variables.

The term high-dimensional implies that the sample size (e.g., the number of subjects/patients) is a lot smaller than the total number of row and column variables.
Since the statistical methods implemented in **HDTD** were primarly motivated by studies in genetics, a microarray dataset is utilized to illustrate the functionality of the package. However, the use of **HDTD** is not limited to gene-expression experiments and we emphasize that it is suitable for analyzing datasets that satisfy the high-dimensional transposable data definition.

2 Mouse Dataset

HDTD includes a subset of the tissue study described in Zahn et al. (2007). This dataset contains expression levels for 40 mice. For each mouse, the expression levels of 46 genes that belong to the vascular endothelial growth factor signalling pathway were measured across 9 tissues (adrenal gland, cerebrum, hippocampus, kidney, liver, muscle, spinal cord, spleen and thymus). The experimental design satisfies the definition of transposable data because: i) the data for each mouse can be written in a matrix form, ii) the rows correspond to genes and the columns to multiple tissues, and iii) we do not expect the gene expression levels across the multiple tissues to vary independently.
The dataset is formatted as a single matrix with rows the 46 genes and columns the $9 \times 40 = 360$ tissues.

```
> library("HDTD")
> data(VEGFmouse)
> dim(VEGFmouse)

[1]  46 360

> rownames(VEGFmouse)

 [1] "Akt1"     "Akt2"     "Akt3"     "Arnt"     "Casp9"    "Cdc42"
 [7] "Grb2"     "Hif1a"    "Hras1"    "Hsp90aa1" "Hspb1"    "Map2k1"
[13] "Map2k2"   "Mapk1"    "Mapk13"   "Mapk14"   "Mapk3"    "Mapkapk2"
[19] "Nfat5"    "Nfatc3"   "Nfatc4"   "Nos3"     "Nras"     "Nrp1"
[25] "Pdgfc"    "Pik3ca"   "Pik3cb"   "Pik3cd"   "Pik3r1"   "Pik3r3"
```

*Anestis.Touloumis@cruk.cam.ac.uk

```
[31] "Pla2g12b" "Pla2g4a"  "Pla2g4c"  "Pla2g5"   "Pla2g6"   "Plcg1"
[37] "Plcg2"    "Ppp3ca"   "Ppp3cb"   "Ppp3r1"   "Prkca"    "Ptk2"
[43] "Rac1"     "Rac2"     "Raf1"     "Sphk2"
```

Further, every 9 consecutive columns belong to the same mouse and the tissues are ordered in the same way for each mouse. For example, we can check the column variables for the first two subjects:

```
> colnames(VEGFmouse)[1:18]
```

```
 [1] "adrenal.1"     "cerebrum.1"    "hippocampus.1" "kidney.1"
 [5] "lung.1"        "muscle.1"      "spinal.1"      "spleen.1"
 [9] "thymus.1"      "adrenal.2"     "cerebrum.2"    "hippocampus.2"
[13] "kidney.2"      "lung.2"        "muscle.2"      "spinal.2"
[17] "spleen.2"      "thymus.2"
```

It is extremely important to provide datasets in this particular format when using **HDTD**. To accomplish this, write the data for each subject (mouse) in a matrix form while preserving the order of the row (genes) and column (tissues) variables. The final step is to create a single matrix by stacking column-wise the subject-specific matrices the one after the other.

3 Mean Relationship of the Genes Across the Tissues

The user can determine the mean relationship of the genes across tissues by testing and estimating the mean matrix. One intersting hypothesis to be tested is the conservation of the gene expression levels across the tissues, i.e., if the mean gene expression levels vector in the VEGF signaling pathway changes across the 9 tissues:

```
> meanmat.ts(VEGFmouse,40,group.sizes=9,voi="columns")
```

```
MEAN MATRIX TEST
Sample Size       = 40
Row Variables     = 46
Column Variables  = 9

Hypothesis Test
H_0: 1 prespecified group(s) of columns with the same mean vector
vs.
H_1: not H_0

Test Statistic = 373.5277 , p-value = <0.0001
```

Since p-value < 0.001, we have strong evidence against the null hypothesis that there is no tissue effect in the gene expression levels. To explore the mean gene expression level pattern across tissues in detail, additional tests can be carried out. We illustrate a more complicated hypothesis that requires data manipulation. Consider testing the hypothesis that the mean gene expression levels vector is constant only across the adrenal glands, the spleen, the kidney and the liver. To do this, we first need to place these 4 tissues in a successive order in the dataset,

```
> colnames(VEGFmouse)[1:9]
```

```
[1] "adrenal.1"     "cerebrum.1"    "hippocampus.1" "kidney.1"
[5] "lung.1"        "muscle.1"      "spinal.1"      "spleen.1"
[9] "thymus.1"
```

```
> VEGForder <- orderdata(VEGFmouse,40,order.cols=c(1,4,5,8,2,3,6,7,9))
> colnames(VEGForder)[1:9]
```

```
[1] "adrenal.1.1"     "kidney.1.1"      "lung.1.1"        "spleen.1.1"
[5] "cerebrum.1.1"    "hippocampus.1.1" "muscle.1.1"      "spinal.1.1"
[9] "thymus.1.1"
```

and then to perform the test using the ordered dataset

```
> meanmat.ts(VEGForder,40,group.sizes=c(4,1,1,1,1,1),voi="columns")

MEAN MATRIX TEST
Sample Size       = 40
Row Variables     = 46
Column Variables  = 9

Hypothesis Test
H_0: 6 prespecified group(s) of columns with the same mean vector
vs.
H_1: not H_0

The number of columns in the 6 predefined groups are 4, 1, 1, 1, 1 and 1  respectively.

Test Statistic = 218.5071 , p-value = <0.0001
```

Note that we included 5 additional column groups of size one in the `group.sizes` argument to reflect the fact that the mean gene expression levels vector in each of the remaining 5 tissues remained unspecified. The null hypothesis is rejected, and hence we may conclude that the mean gene expression levels vector is not constant in the adrenal glands, the spleen, the kidney and the liver.
Apart from hypothesis testing, it is also important to estimate the mean relationship between the genes and the tissues. In this example the mean matrix seems to be unstructured and thus the mean gene expression levels in the 9 tissues can be estimated via the sample mean matrix

```
> sample.mean <- meanmat.hat(VEGFmouse,40)
> sample.mean

ESTIMATION OF THE MEAN MATRIX
Sample Size           =  40
Row Variables         =  46
Column Variables      =  9

Estimated Mean Matrix [1:5,1:5] =
      adrenal.1 cerebrum.1 hippocampus.1 kidney.1  lung.1
Akt1     0.8399     1.2157        1.0597   1.1469  1.2673
Akt2    -0.2333    -0.6201       -0.3881  -0.5524 -0.5359
Akt3    -1.0856    -0.4351       -0.5490  -0.2534 -0.6091
Arnt     0.1089     0.1898        0.0968   0.2551 -0.1171
Casp9    0.0877     0.2600        0.4812   0.3203  0.7416
```

Note that the output preserves the order of the genes and the columns. For example, 0.8399 is the average $\log_2$ intensity for gene "Akt1" in the adrenal gland based on 40 mice. The mean matrix for the first 10 genes across the 9 tissues is

```
> head(round(sample.mean$estmeanmat,4),n=10)

         adrenal.1 cerebrum.1 hippocampus.1 kidney.1  lung.1 muscle.1 spinal.1
Akt1        0.8399     1.2157        1.0597   1.1469  1.2673   0.8459   1.2201
Akt2       -0.2333    -0.6201       -0.3881  -0.5524 -0.5359  -0.2082  -0.3877
Akt3       -1.0856    -0.4351       -0.5490  -0.2534 -0.6091  -1.1794  -0.7467
```

```
Arnt        0.1089     0.1898        0.0968   0.2551 -0.1171  -0.2919  -0.0861
Casp9       0.0877     0.2600        0.4812   0.3203  0.7416  -0.2492   0.3691
Cdc42      -0.0538     0.1657       -0.1516  -0.0548  0.0254  -0.3577  -0.0816
Grb2       -0.2765    -0.5322        0.0948   0.0162 -0.1499   0.3015   0.2385
Hif1a      -0.5760     1.3233       -3.5652   1.5485  1.0256  -0.1264   1.1011
Hras1      -0.8040    -0.5952       -0.4063  -0.4964 -0.4997  -0.4900  -0.4475
Hsp90aa1   -0.3007     0.1292        0.7474   0.5589  0.2163   0.2634   0.7414
         spleen.1 thymus.1
Akt1       1.2142   1.2025
Akt2      -0.5154  -0.6836
Akt3      -0.8073  -0.4024
Arnt      -0.5188  -0.2219
Casp9      0.5682   0.0726
Cdc42     -0.0888   0.1986
Grb2      -0.4664  -0.3850
Hif1a     -2.5046   0.3866
Hras1     -0.8511  -0.7226
Hsp90aa1  -0.2552  -0.2784
```

4 Dependence Structure of the Genes and of the Tissues

The matrix-variate normal distribution is utilized to estimate two covariance matrices, one for the genes (rows) and the other for the multiple tissues (columns). We developed shrinkage estimators for both covariance matrices but we let the user decide if shrinkage is required to both, one or neither of these matrices. In principle, we recommend shrinking both covariance matrices in order to obtain well-defined and invertible covariance matrix estimators. The `covmat.hat` function provides the corresponding covariance estimators:

```
> estcovmat <- covmat.hat(datamat = VEGFmouse, N = 40, shrink = "both", centered = FALSE)
> estcovmat

ESTIMATION OF THE ROW AND/OR THE COLUMN COVARIANCE MATRIX
Sample Size           =  40
Row Variables         =  46
Column Variables      =  9
Shrinking             =  Both Sets of Variables
Centered Data         =  FALSE

ROW VARIABLES
Estimated Shrinkage Intensity = 0.0115
Estimated Covariance Matrix [1:5,1:5] =
         Akt1    Akt2    Akt3    Arnt   Casp9
Akt1   0.4139 -0.0248  0.0420 -0.0010  0.1084
Akt2  -0.0248  0.3341 -0.0240 -0.0029 -0.0151
Akt3   0.0420 -0.0240  0.6954  0.1733 -0.0168
Arnt  -0.0010 -0.0029  0.1733  0.4746  0.0850
Casp9  0.1084 -0.0151 -0.0168  0.0850  0.5337

COLUMN VARIABLES
Estimated Shrinkage Intensity = 0.3341
Estimated Covariance Matrix [1:5,1:5] =
              adrenal.1 cerebrum.1 hippocampus.1 kidney.1  lung.1
adrenal.1        0.0368    -0.0006        0.0001  -0.0006  0.0010
cerebrum.1      -0.0006     0.0432       -0.0002   0.0000 -0.0034
hippocampus.1    0.0001    -0.0002        0.0266   0.0019  0.0000
```

```
kidney.1        -0.0006     0.0000        0.0019   0.0317  0.0012
lung.1           0.0010    -0.0034        0.0000   0.0012  0.0809
```

The output summarizes the results but the user can recover the full covariance matrix estimators. For example, the covariance matrix of the tissues is

```
> round(estcovmat$cols.covmat,4)
```

```
              adrenal.1 cerebrum.1 hippocampus.1 kidney.1  lung.1 muscle.1
adrenal.1        0.0368    -0.0006        0.0001  -0.0006  0.0010   0.0006
cerebrum.1      -0.0006     0.0432       -0.0002   0.0000 -0.0034   0.0012
hippocampus.1    0.0001    -0.0002        0.0266   0.0019  0.0000  -0.0005
kidney.1        -0.0006     0.0000        0.0019   0.0317  0.0012  -0.0006
lung.1           0.0010    -0.0034        0.0000   0.0012  0.0809   0.0009
muscle.1         0.0006     0.0012       -0.0005  -0.0006  0.0009   0.0300
spinal.1        -0.0003    -0.0012        0.0004   0.0001  0.0155   0.0007
spleen.1        -0.0010    -0.0038       -0.0005  -0.0009  0.0014  -0.0004
thymus.1        -0.0006     0.0003        0.0012   0.0026  0.0008  -0.0001
              spinal.1 spleen.1 thymus.1
adrenal.1      -0.0003  -0.0010  -0.0006
cerebrum.1     -0.0012  -0.0038   0.0003
hippocampus.1   0.0004  -0.0005   0.0012
kidney.1        0.0001  -0.0009   0.0026
lung.1          0.0155   0.0014   0.0008
muscle.1        0.0007  -0.0004  -0.0001
spinal.1        0.0393  -0.0012   0.0012
spleen.1       -0.0012   0.0389  -0.0012
thymus.1        0.0012  -0.0012   0.0410
```

Moreover, the user can study the gene-wise or tissue-wise correlation by using the `covmat.ts` function. For example, the identity and sphericity test

```
> covmat.ts(datamat = VEGFmouse, N = 40, voi = "columns", centered = FALSE)
```

```
SPHERICITY AND IDENTITY TESTS FOR THE ROW OR COLUMN VARIABLES
Sample Size           =  40
Row Variables         =  46
Column Variables      =  9
Variables Tested      =  Columns
Centered Data         =  FALSE

Sphericity test for the covariance matrix of the Columns
Test Statistic = 10.474 , p-value = <0.0001

Identity Test for the covariance matrix of the Columns
Test Statistic = 39.0576 , p-value = <0.0001
```

suggest that the tissues might not be uncorrelated since the p-values of the sphericity and the identity test are both < 0.001.

5 Citation

```
> citation("HDTD")
```

```
Please use the following guidelines for citing `HDTD' in publication:
```

```
To cite the mean matrix hypothesis testing methodology, please use

  Touloumis, A., Tavare, S. and Marioni, J.C. (2014). Testing the Mean
  Matrix in High-Dimensional Transposable Data, To appear in Biometrics

To cite the covariance matrix hypothesis testing methodology, please
use

  Touloumis, A., Marioni, J.C. and Tavare, S. (2013). Hypothesis
  Testing for the Covariance Matrix in High-Dimensional Transposable
  Data with Kronecker Product Dependence Structure,
  http://arxiv.org/abs/1404.7684

To cite the software, please use

  Touloumis, A., Marioni, J.C. and Tavare, S. (2014). HDTD: Statistical
  Inference about the Mean Matrix and the Covariance Matrices in
  High-Dimensional Transposable,
  URL=http://www.bioconductor.org/packages/3.0/bioc/html/HDTD.html
```

6 Session Info

```
> sessionInfo()

R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] HDTD_1.2.0

loaded via a namespace (and not attached):
[1] tools_3.2.0
```