See if new vruv2 works ok on real data

David Gerard

2016-07-16

Abstract

I try out a new variance inflation formulation for RUV2 by accounting for the variance inflation during the factor analysis. It doesn't work too well.

Inflation in RUV2

The model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{E} \tag{1}$$

we rotate to the three models

$$\boldsymbol{Y}_1 = \boldsymbol{R}_{11}\boldsymbol{\beta}_1 + \boldsymbol{R}_{12}\boldsymbol{\beta}_2 + \boldsymbol{Z}_1\boldsymbol{\alpha} + \boldsymbol{E}_1 \tag{2}$$
$$\boldsymbol{Y}_2 = \boldsymbol{R}_{22}\boldsymbol{\beta}_2 + \boldsymbol{Z}_2\boldsymbol{\alpha} + \boldsymbol{E}_2 \tag{3}$$
$$\boldsymbol{Y}_3 = \boldsymbol{Z}_3\boldsymbol{\alpha} + \boldsymbol{E}_3. \tag{4}$$

I earlier showed that RUV2 does the following steps:

1. Estimate $\boldsymbol{Z}_2$, $\boldsymbol{Z}_3$, and $\boldsymbol{\alpha}_{\mathcal{C}}$ by using factor analysis on $\binom{\boldsymbol{Y}_{2\mathcal{C}}}{\boldsymbol{Y}_{3\mathcal{C}}}$.
2. Estimate $\boldsymbol{\alpha}$ by regressing $\boldsymbol{Y}_3$ on $\boldsymbol{Z}_3$, i.e. $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{Z}}_3^T\hat{\boldsymbol{Z}}_3)^{-1}\hat{\boldsymbol{Z}}_3^T\boldsymbol{Y}_3$.
3. Estimate $\boldsymbol{\beta}$ with $\boldsymbol{R}_{22}^{-1}(\boldsymbol{Y}_2 - \hat{\boldsymbol{Z}}_2\hat{\boldsymbol{\alpha}})$.

My new idea for variance inflation in RUV2 is to account for variance inflation *during the factor analysis*. I am guessing that the reason why vruv4 works, is that for some reason, the variances in (3) are different from those in (4), so under that hypothesis it would make sense to account for these differences directly in the factor analysis. That is, fit the model

$$\boldsymbol{Y}_{2\mathcal{C}} = \boldsymbol{Z}_2\boldsymbol{\alpha}_{\mathcal{C}} + \boldsymbol{E}_{2\mathcal{C}} \tag{5}$$
$$\boldsymbol{Y}_{3\mathcal{C}} = \boldsymbol{Z}_3\boldsymbol{\alpha}_{\mathcal{C}} + \boldsymbol{E}_{3\mathcal{C}} \tag{6}$$
$$\boldsymbol{E}_{2\mathcal{C}} \sim N_{k_2\times m}(0, \lambda\boldsymbol{\Sigma}_{\mathcal{C}} \otimes \boldsymbol{I}_{k_2}) \tag{7}$$
$$\boldsymbol{E}_{3\mathcal{C}} \sim N_{n-k_2\times m}(0, \boldsymbol{\Sigma}_{\mathcal{C}} \otimes \boldsymbol{I}_{n-k_2}). \tag{8}$$

In words, we perform a factor analysis where the first k_2 rows have variances that differ by a multiplication factor from the variances of the last $n - k_2$ rows.

Factor Analysis

I fit the factor analysis by maximum likelihood where we assume $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ contain iid standard normals. This was fit using the EM described in Rubin and Thayer (1982) but modified to estimate the variance inflation parameter. This got estimates of $\boldsymbol{\alpha}_{\mathcal{C}}$, $\boldsymbol{\Sigma}_{\mathcal{C}}$ and λ. I obtained estimates of $\boldsymbol{Z}$ by GLS as described in section 6 of Bai and Li (2012).

Looking at vruv2

Load in data and estimate number of hidden confounders.

```
library(vicar)
source("../code/data_generators.R")
dout <- pois_thin(Nsamp = 20, nullpi = 1, path = "../../../data/gtex_tissue_gene_reads/",
                  ncontrol = 2000, Ngene = 10000, tissue = "muscle")
dout$num_sv
```

```
## [1] 7
```

```
vout <- vruv2(Y = dout$Y, X = dout$X, ctl = dout$control_genes, k = dout$num_sv,
              likelihood = "normal")
vout$multiplier
```

```
## [1] 0.8898
```

```
aout <- ashr::ash(betahat = vout$betahat, sebetahat = vout$sebetahat)
ashr::get_pi0(aout)
```

```
## [1] 0.6802
```

The variance inflation is super small, usually less than 1, and `ashr` estimates of π_0 aren't good.

Compare to `vruv4`.

```
vout4 <- vruv4(Y = dout$Y, X = dout$X, ctl = dout$control_genes, k = dout$num_sv,
               likelihood = "normal")
aout <- ashr::ash(betahat = vout4$betahat, sebetahat = vout4$sebetahat)
ashr::get_pi0(aout)
```

```
## [1] 0.9993
```

This worked fine.

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [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] vicar_0.1.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5          compiler_3.3.1       formatR_1.3
##  [4] iterators_1.0.8      tools_3.3.1          digest_0.6.9
##  [7] annotate_1.48.0      evaluate_0.9         RSQLite_1.0.0
## [10] nlme_3.1-128         cate_1.0.4           lattice_0.20-33
## [13] mgcv_1.8-12          foreach_1.4.3        Matrix_1.2-6
## [16] DBI_0.4              yaml_2.1.13          parallel_3.3.1
## [19] genefilter_1.52.0    stringr_1.0.0        knitr_1.12.28
## [22] REBayes_0.62         S4Vectors_0.8.6      IRanges_2.4.6
## [25] stats4_3.3.1         grid_3.3.1           Biobase_2.30.0
## [28] ruv_0.9.6            AnnotationDbi_1.32.3 XML_3.98-1.3
## [31] survival_2.39-4      rmarkdown_0.9.6      leapp_1.2
## [34] limma_3.26.3         sva_3.18.0           ashr_1.2.5
## [37] corpcor_1.6.8        magrittr_1.5         MASS_7.3-45
## [40] codetools_0.2-14     htmltools_0.3.5      BiocGenerics_0.16.1
## [43] splines_3.3.1        svd_0.4              assertthat_0.1
## [46] xtable_1.8-2         esaBcv_1.2.1         stringi_1.0-1
## [49] Rmosek_7.1.2         pscl_1.4.9           doParallel_1.0.10
## [52] truncnorm_1.0-7      SQUAREM_2014.8-1
```

References

Bai, Jushan, and Kunpeng Li. 2012. "Statistical Analysis of Factor Models of High Dimension." *The Annals of Statistics*. JSTOR, 436–65.

Rubin, Donald B, and Dorothy T Thayer. 1982. "EM Algorithms for ML Factor Analysis." *Psychometrika* 47 (1). Springer: 69–76.