Partial least-squares for imputation of white blood cell composition

Maarten van Iterson[1]

[1]Leiden University Medical Center, Department of Molecular Epidemiology, Leiden, The Netherlands

March 18, 2016

1 Introduction

Within the BIOS data white blood cell counts (WBC), i.e., neutrophils, lymphocytes, monocytes, eosinophils and basophils, were measures by the standard WBC differential as part of the CBC (Complete Blood Count). However, a considerable number of samples ($\approx 1/3$) lack CBC measurements. Since DNA methylation or gene expression levels are informative of the white blood cell composition[1] a linear predictor was build, based on either DNA methylation or gene expression, to infer the white blood cell composition of those samples lacking WBC measurements.

2 Method

The problem we face is to predict a multivariate response, counts or precentages for the five cell types using high-dimensional covariates; DNA methylation or gene expression measurements. Obviously, this model can not be fitted using ordinary least-squares, since, the number of covariates is much higher then the number of samples ($p >> n$); we need some kind of regularization.

We have chosen to use partial least-squares for fitting the linear model. The advantage of partial least-squares is that it both can handle multivariate responses and high-dimensional ($p >> n$) covariates. The R-package pls[2] was used to fit the model and to optimize the number of pls-components using five-fold cross-validation. Age and gender were included as covariates since WBCC seems to be dependent on both. Furthermore, we decided to use the fraction of white blood cell counts divided by the total amount of white blood cells, e.g., in contrast to the counts. We have tested both, as well as, included platelet counts and total white blood counts, also used log10-transformed percentages all gave more or less the same results. Therefore, we decided to use the simplest model, but the pls-approach can thus easily be used for the other models.

For validation the following approach was used: data with WBCC available was split in a train (2/3) and test set (1/3) (across cohorts). A pls-model was fitted on the train set and used to predict WBCC on the test set. See the supplemental sections for a description of the validation results.

3 Discussion

The multivariate responses, white blood counts on five cell types, represents compositional data, i.e., the data are percentages that sum up to 100%. The compositional nature of the data was not explicitly modelled, however the predicted WBC percentages sum up to 100% almost exactly (**ADD** mean +/- std).

One drawback of the pls approach is that all covariates a required for the predictor, although probably many will be noninformative. Therefore, we are currently looking in other approaches that give us a sparse predictor.

References

[1] E.A. Houseman et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 8(13), 2012.

[2] Wehrens R. Mevik, B.-H. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2), 2007.

3.1 Using DNA methylation as predictor

The DNA methylation data was preprocessed sample and probe filtering and normalized using Functional normalization. Subsequently, transformed to M-values NA's were imputed using k-nearest neighbour method.

In general abundant cell types are easier to predict e.g., neutrophils($\approx 62\%$) and lymphocytes$\approx 30\%$ compared to the low abundant basophils ($\approx 1\%$). See wikipedia white blood cell types.

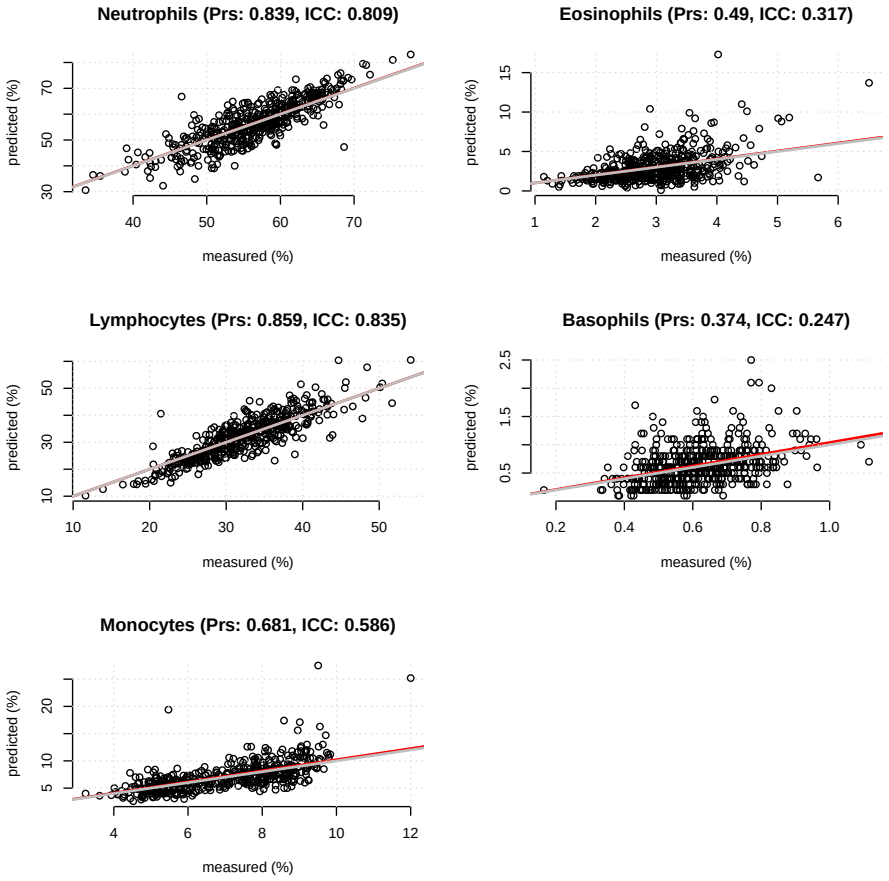

Figure 1: Pearson correlation of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

Both scatter plots and Bland-Altman plots show that there is good agreement between measured and predicted white blood cell percentages. The Bland-

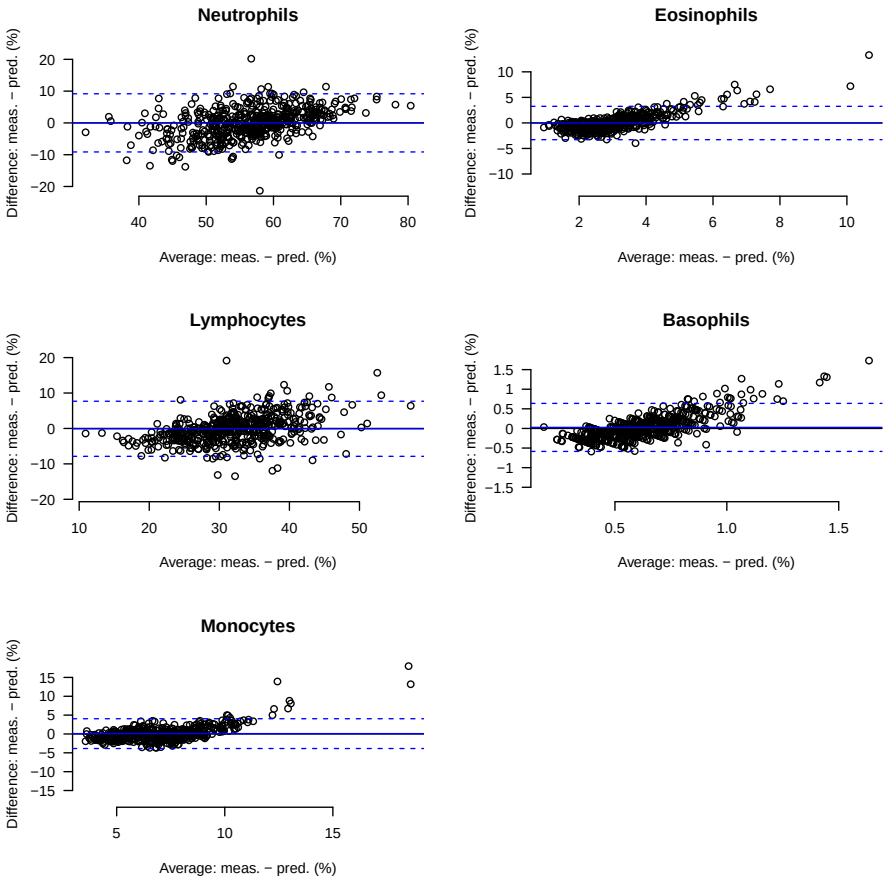

Figure 2: Bland-Altman plots of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

Altman plots show that there is a small systematic bias, as observed by the small positive trend between difference and average of measured vs predicted.

The observed negative correlation between neutrophils and lymphocytes is a consequences of being the most abundunt. If one is high the other necessarily need to be lower.

The heatmap show again the high abundant cell types have a distinct DNA methylation profile

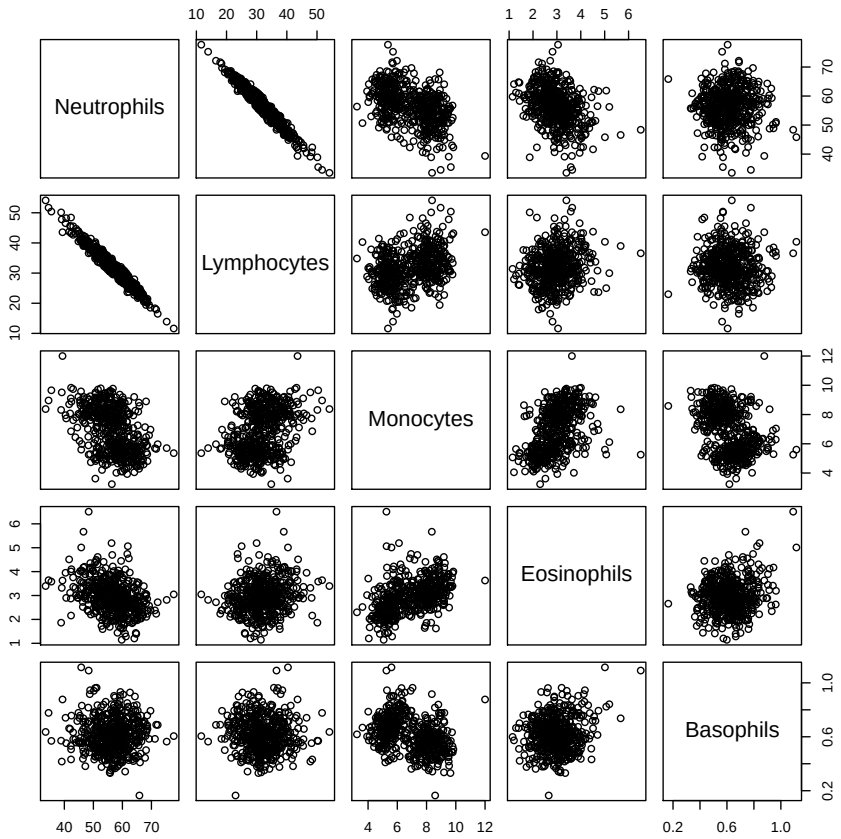

Figure 3: scatter plot (pairs) of the predicted white blood cell percentages of those samples lacking white blood cell counts.

3.2 Using gene expression as predictor

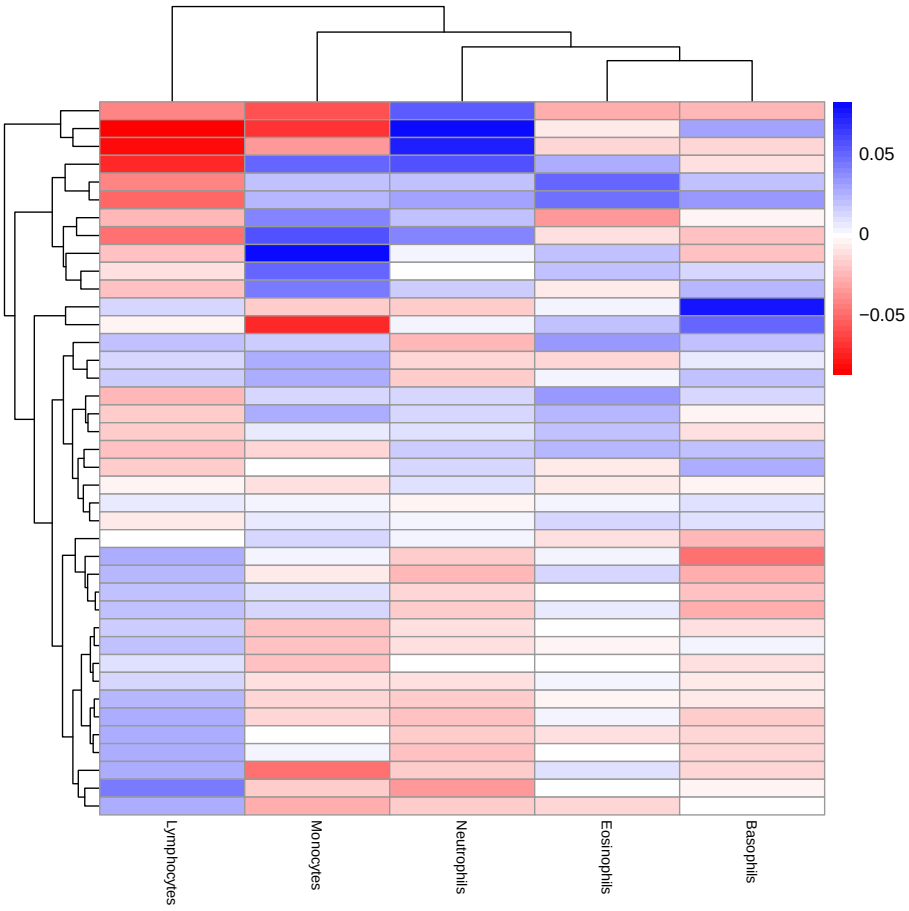

Figure 4: Heatmap of correlation of the CpGs with highest contribution to the prediction with the cell percentages.

4 Supplemental Figures

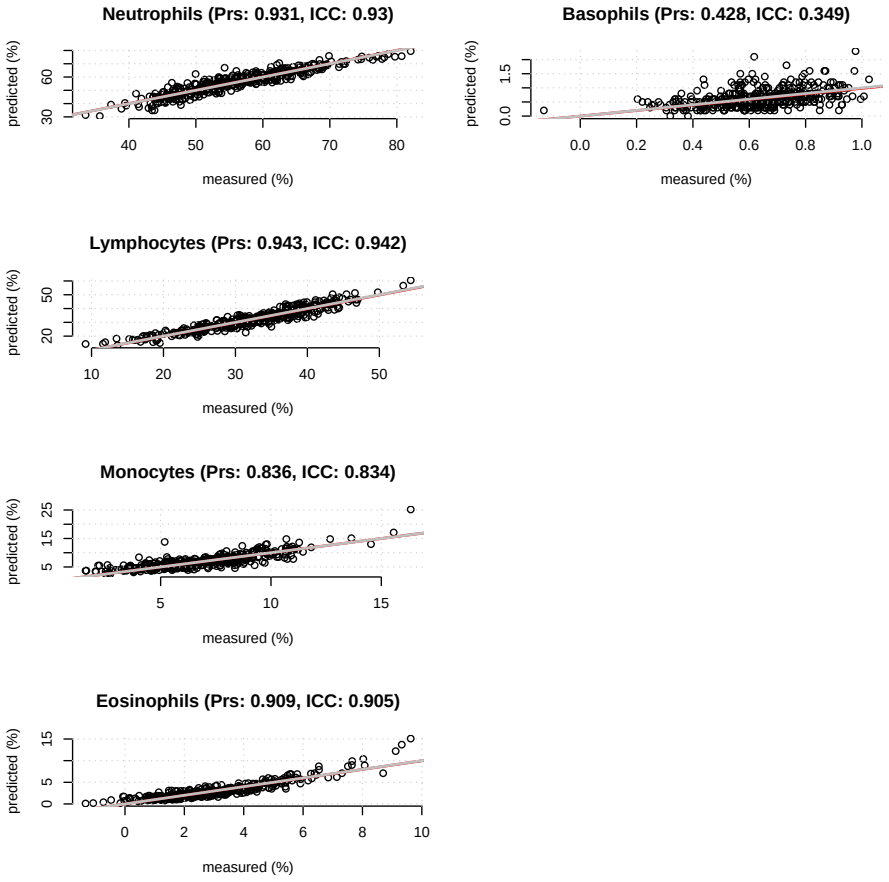

Figure 5: Pearson correlation of measured with predicted cell percentages based on a pls-predictor using gene expression levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

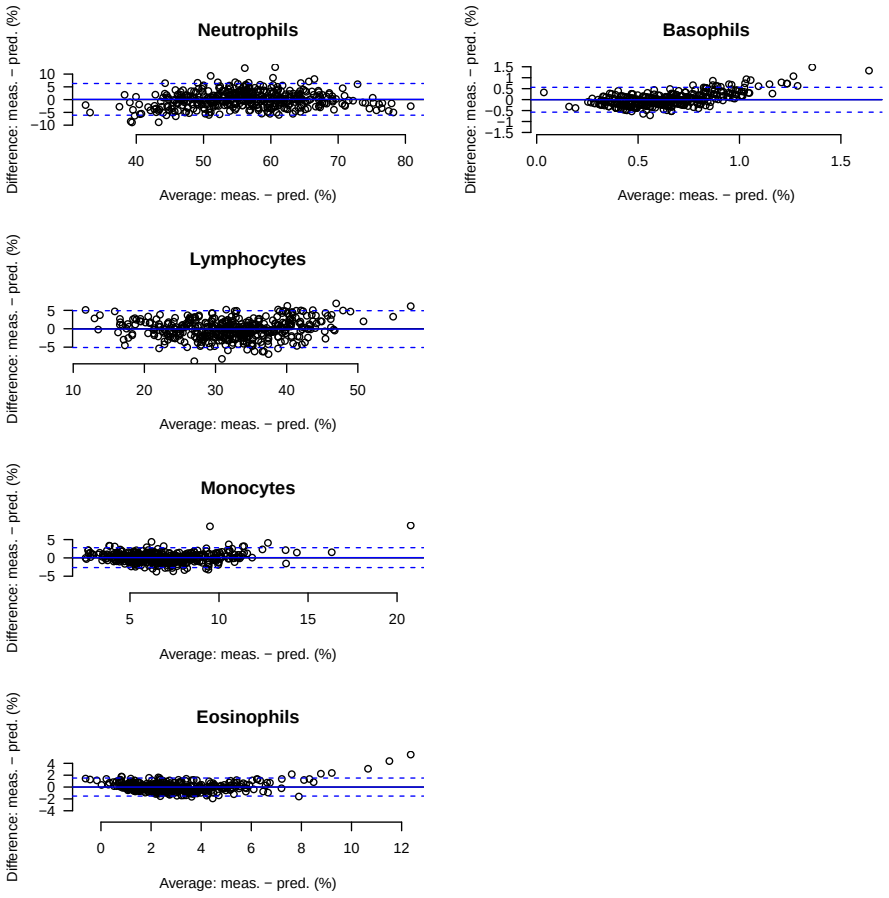

Figure 6: Bland-Altman plots of measured with predicted cell percentages based on a pls-predictor using gene expression levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

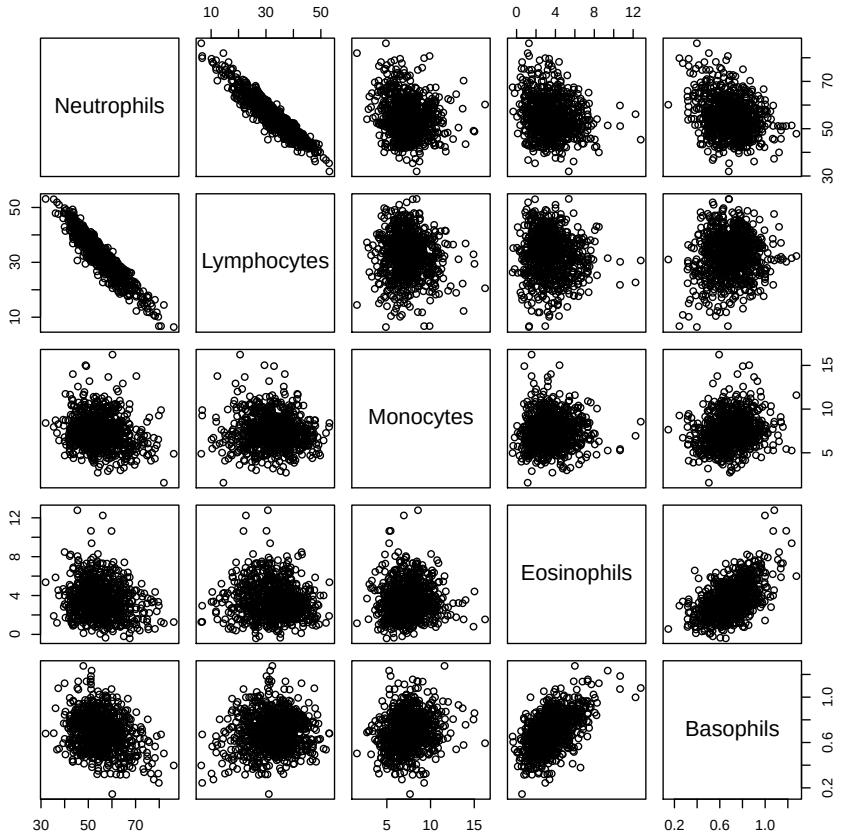

Figure 7: scatter plot (pairs) of the predicted white blood cell percentages of those samples lacking white blood cell counts.

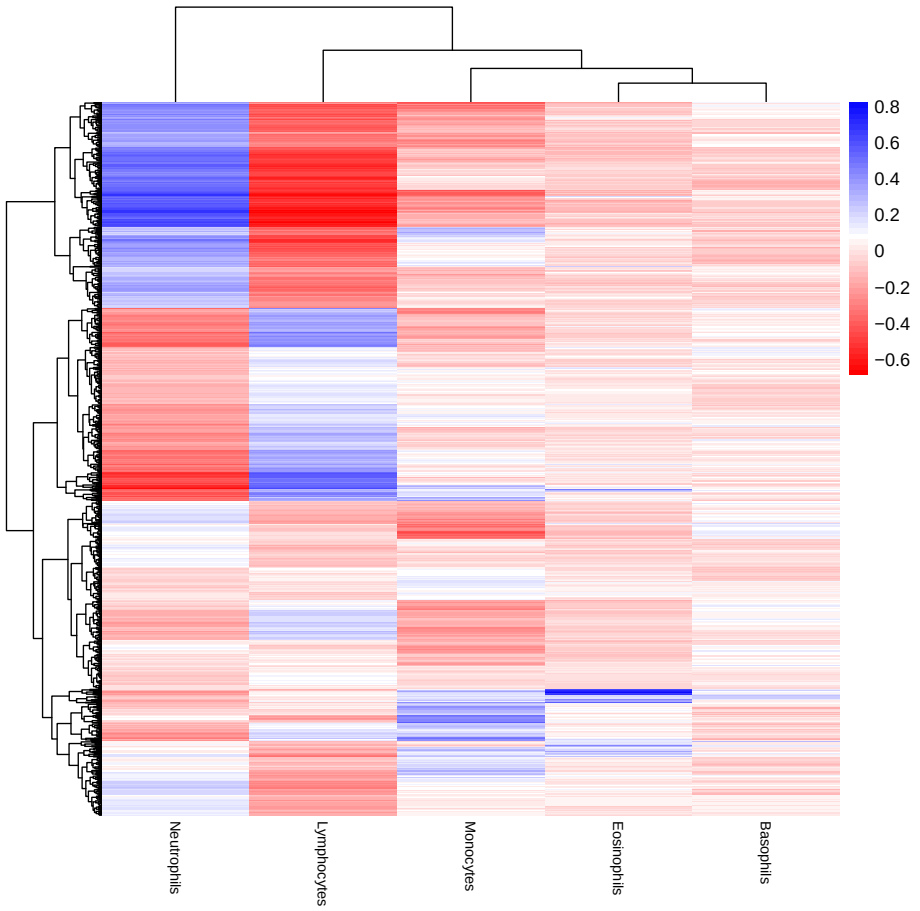

Figure 8: Heatmap of correlation of the genes with highest contribution to the prediction with the cell percentages.

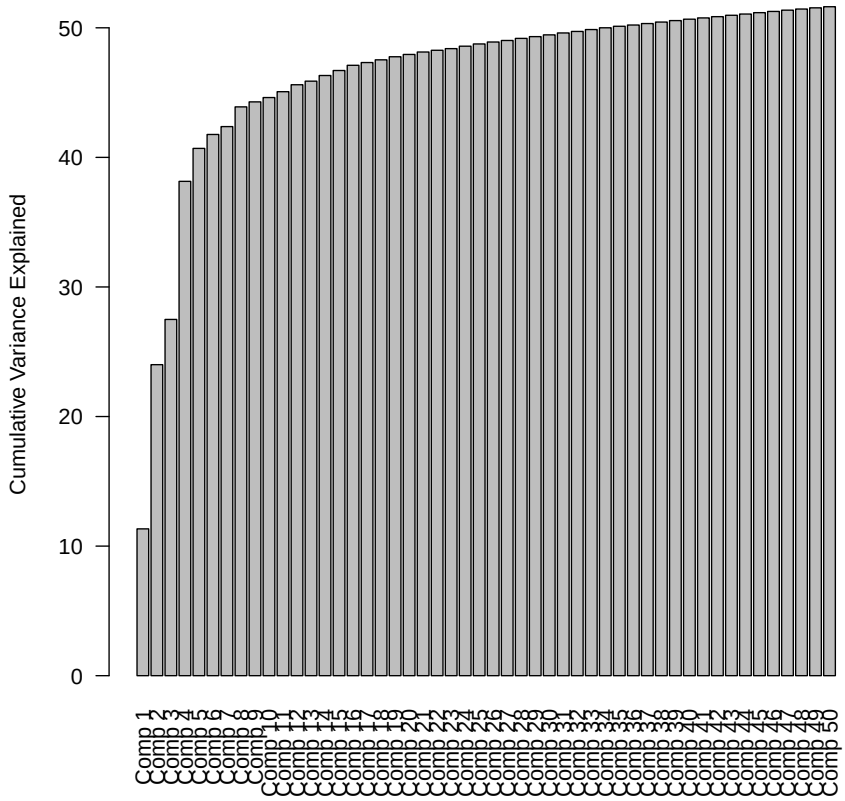

Figure 9: DNA methylation: Cumulative variance explained vs the number of pls-components, truncated at 50 components.

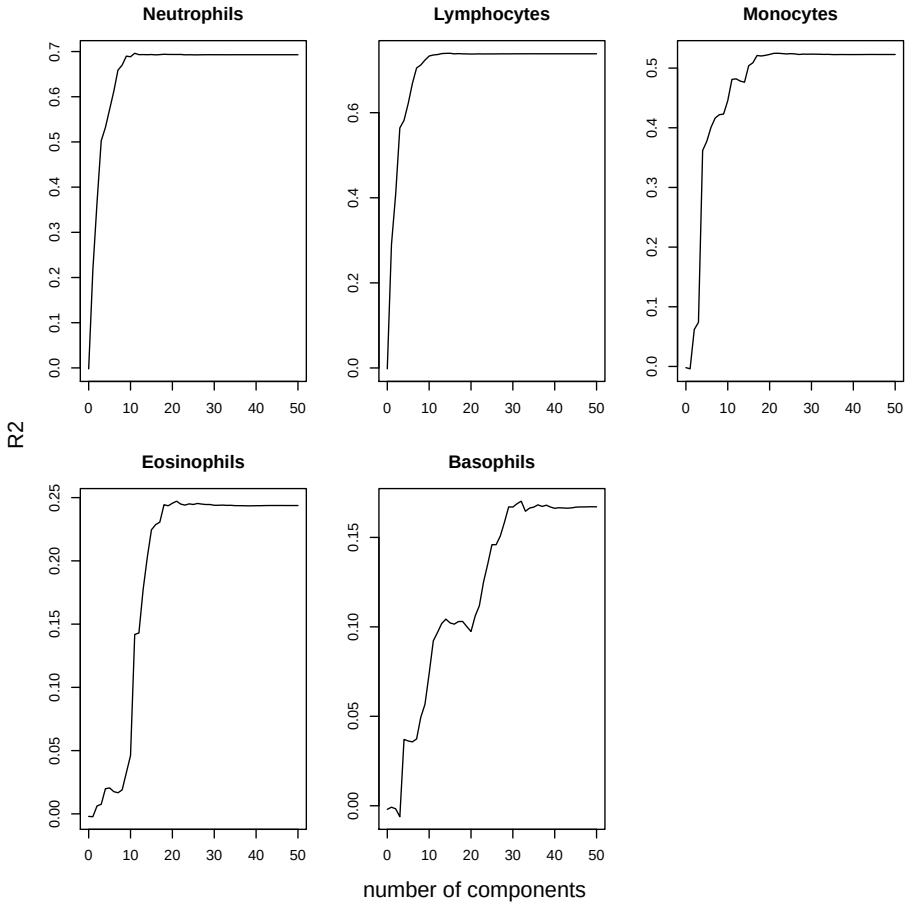

Figure 10: DNA methylation: : R^2 per cell type as function of the number of pls-components.

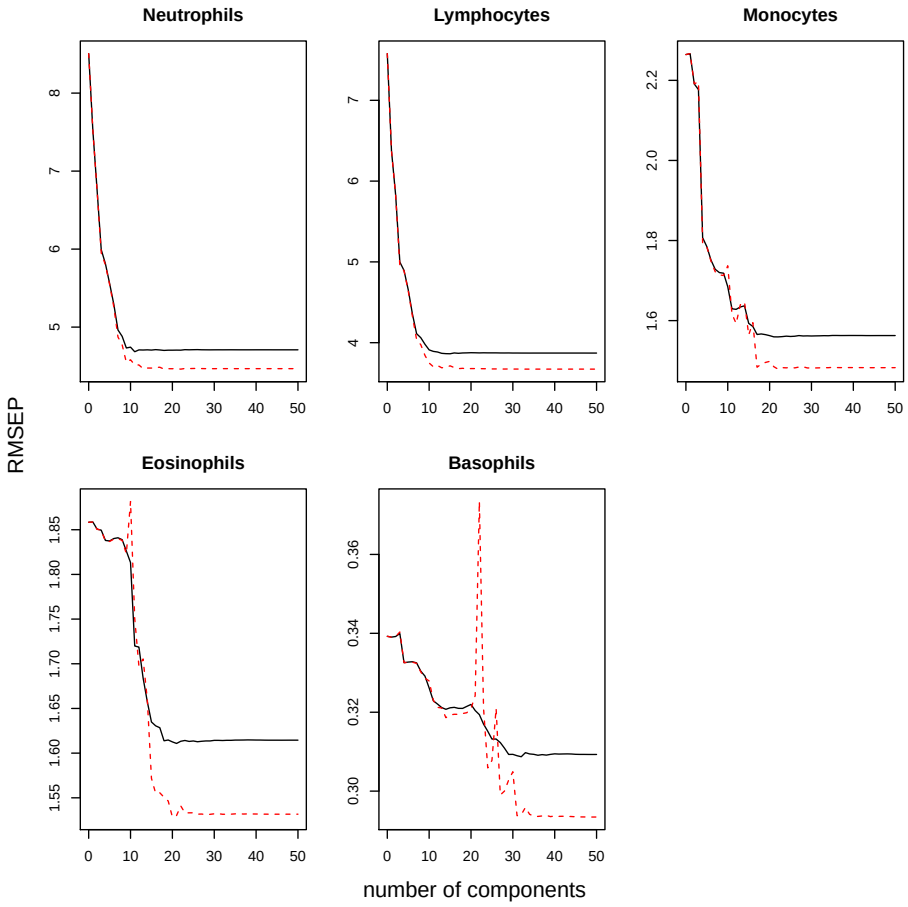

Figure 11: DNA methylation: Root mean squared error of the prediction as function of the number of pls-components (there are two methods to calculate this see ?pls).

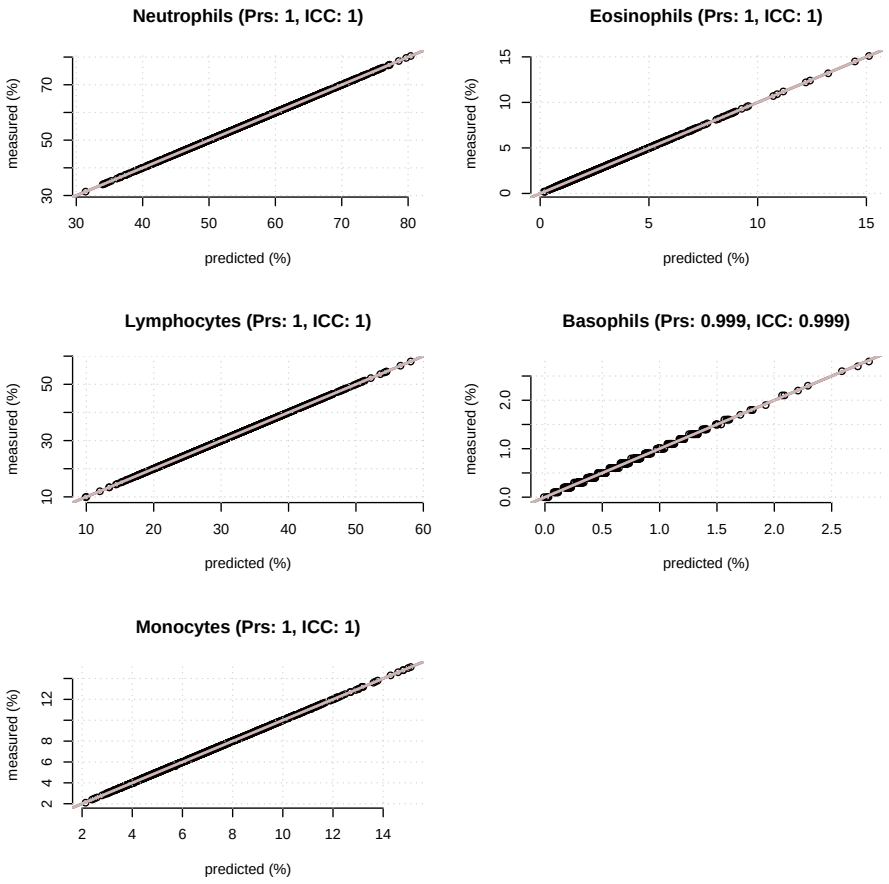

Figure 12: DNA methylation: Train set: Pearson correlation of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

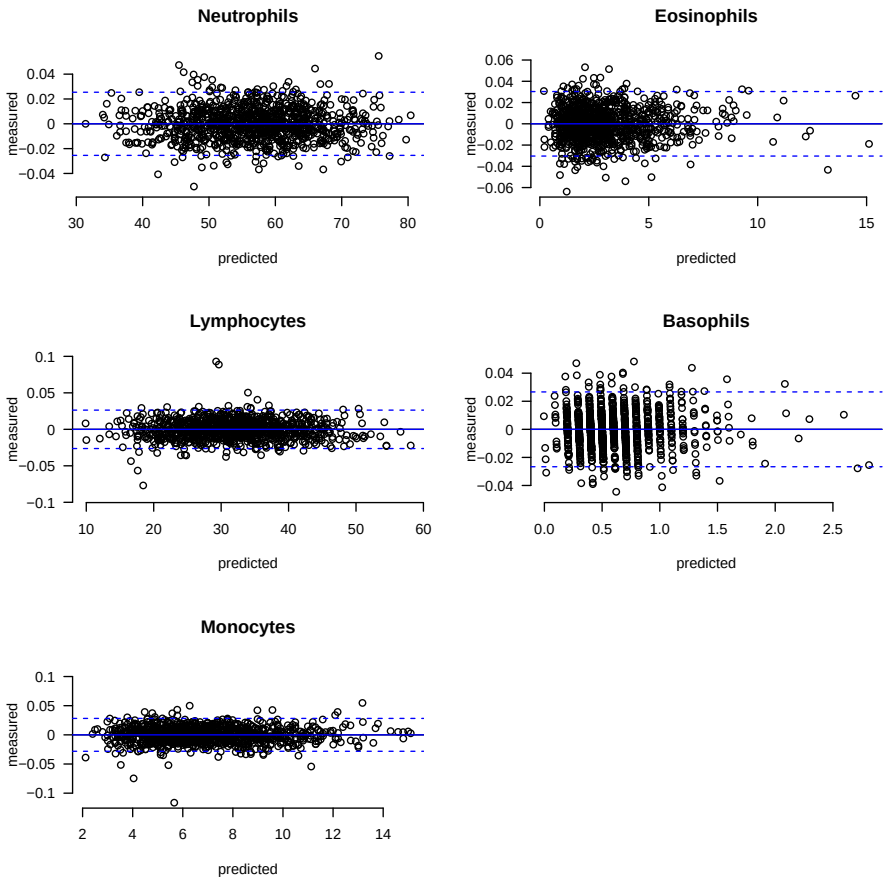

Figure 13: DNA methylation: Tain set: Bland-Altman plots of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

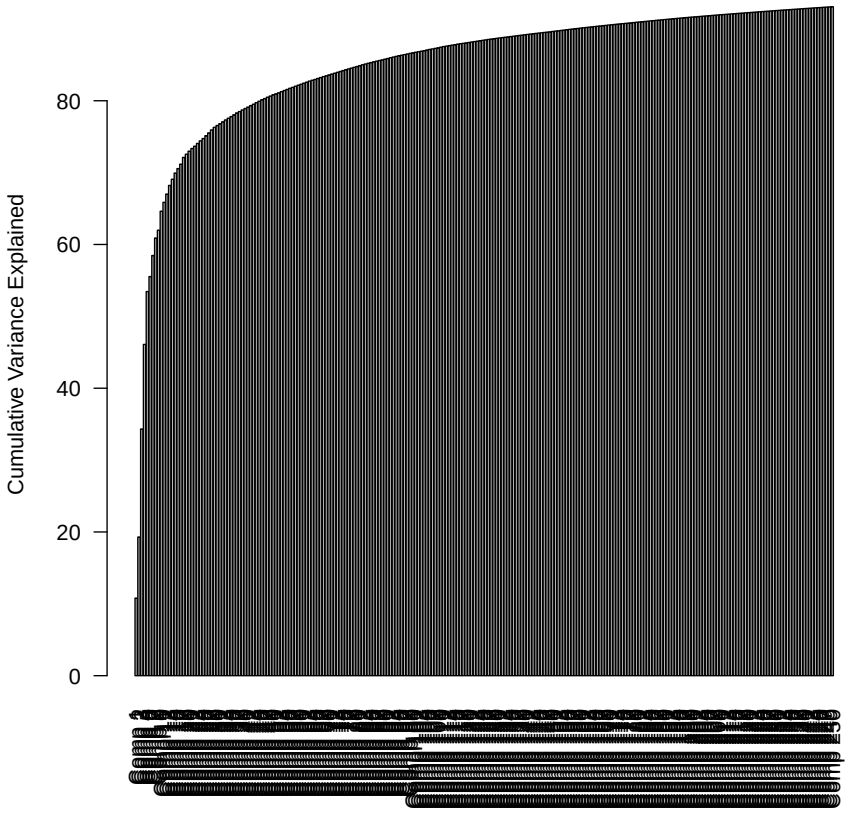

Figure 14: Gene expression: Cumulative variance explained vs the number of pls-components, truncated at 50 components.

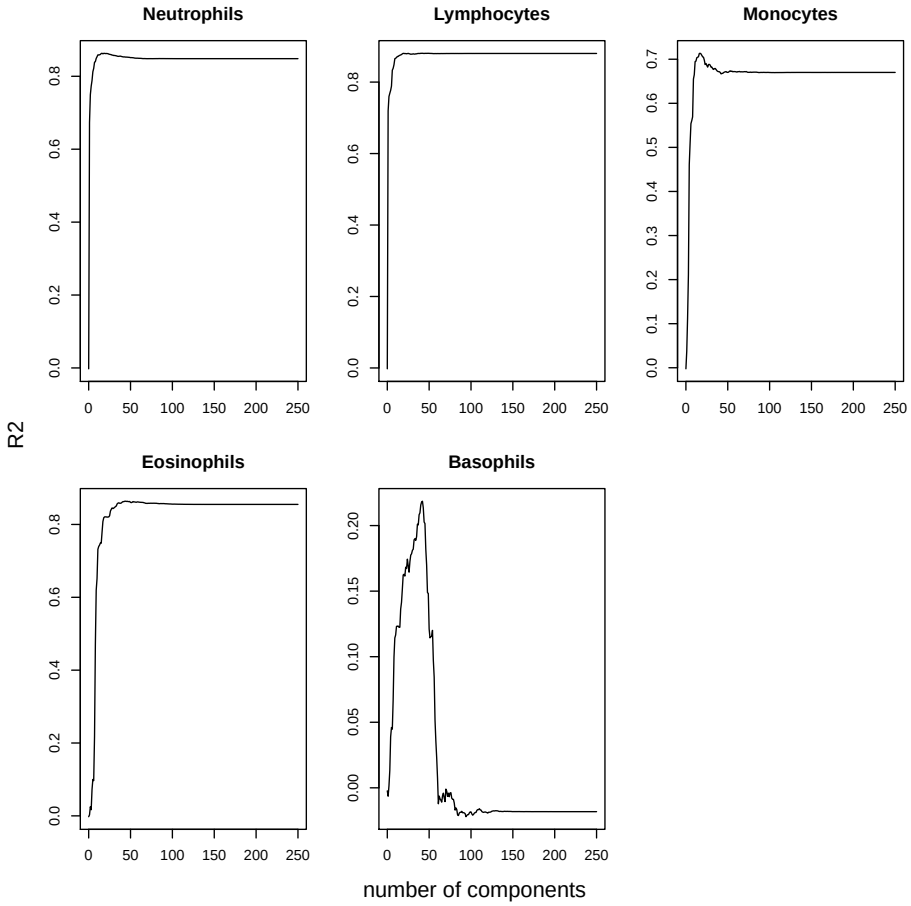

Figure 15: Gene expression: R^2 per cell type as function of the number of pls-components.

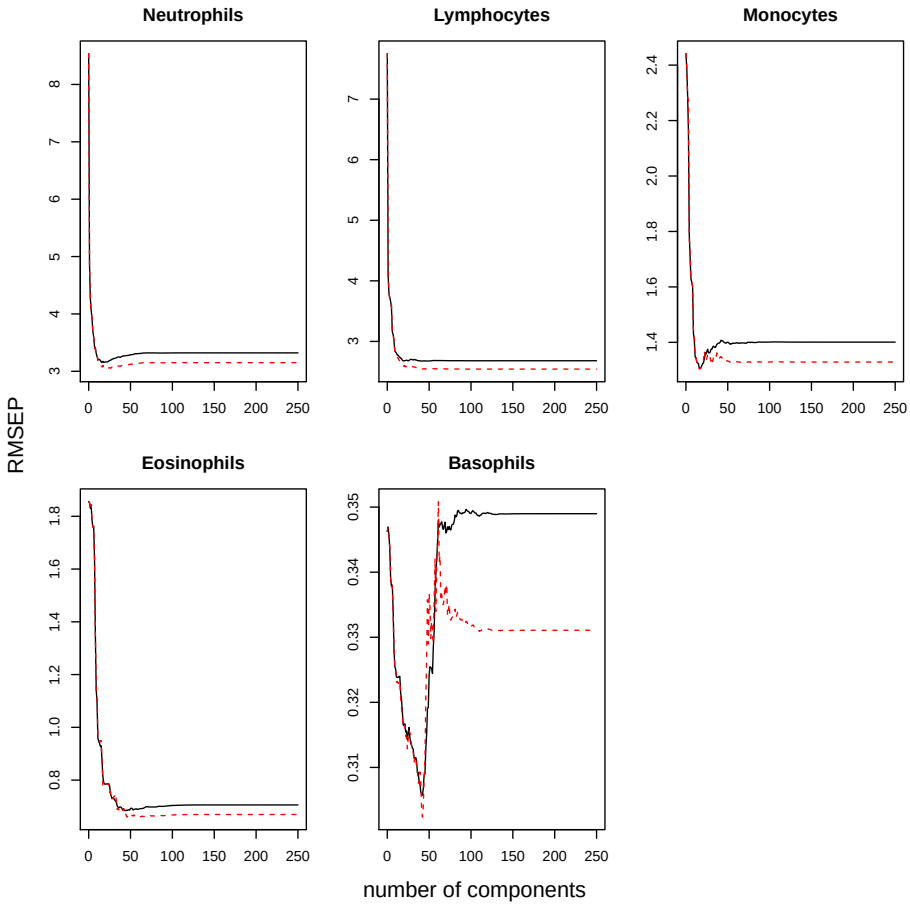

Figure 16: Gene expression: Root mean squared error of the prediction as function of the number of pls-components (there are two methods to calculate this see ?pls).

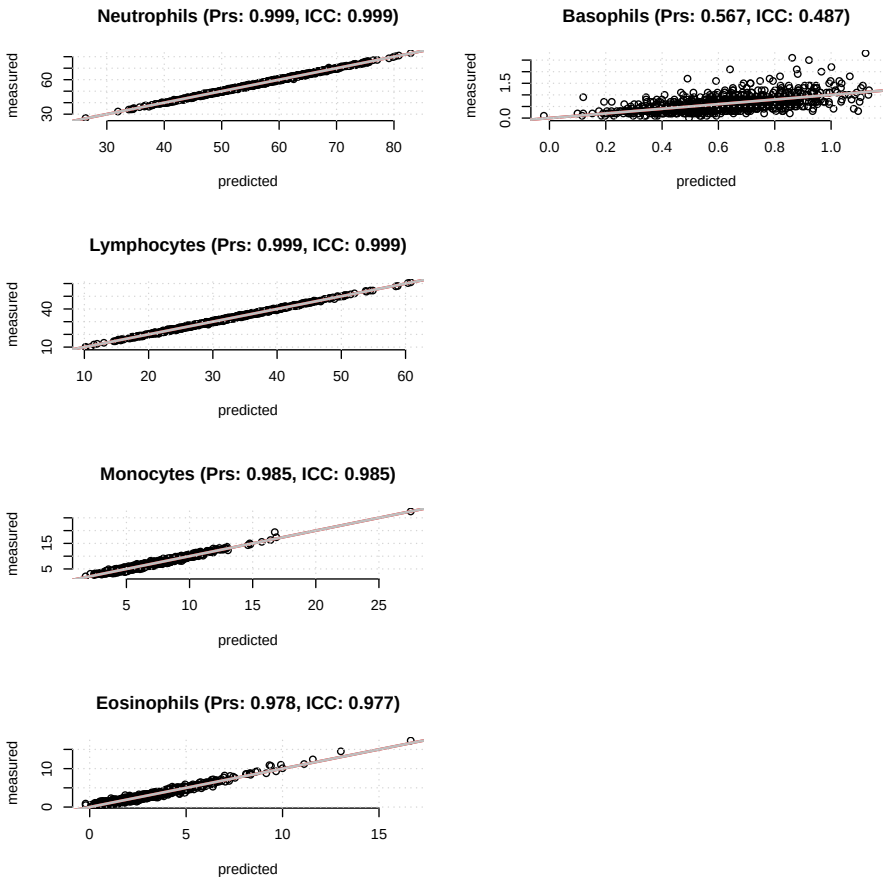

Figure 17: Gene expression: Train set: Pearson correlation of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).

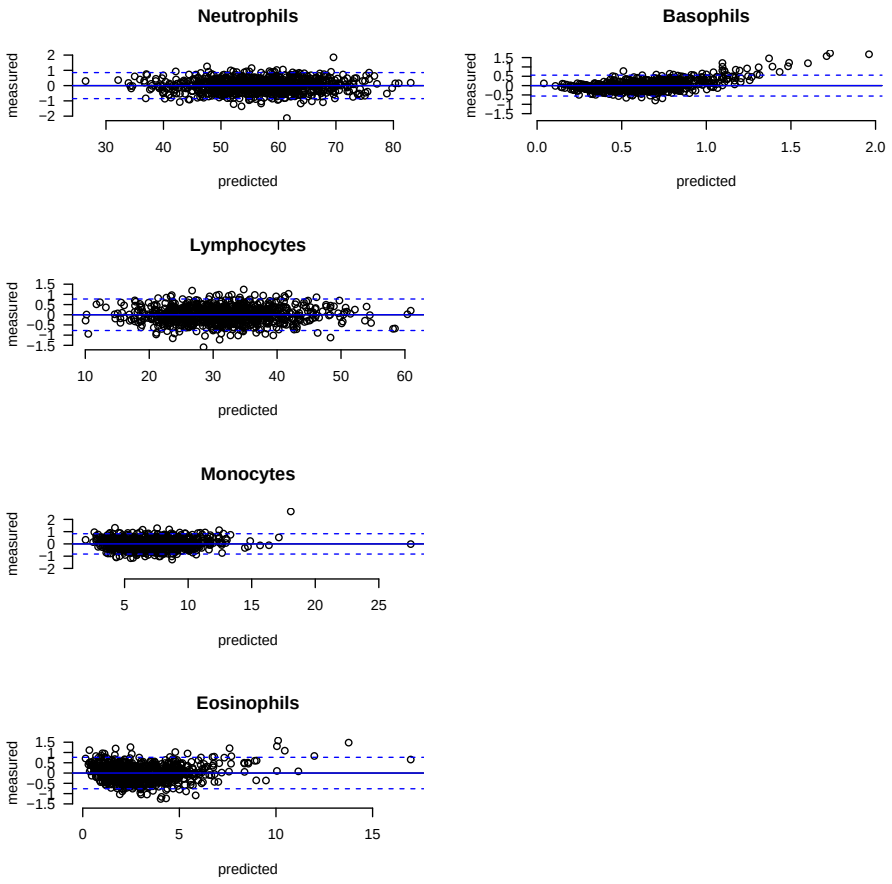

Figure 18: Gene expression: Tain set: Bland-Altman plots of measured with predicted cell percentages based on a pls-predictor using DNA methylation levels as covariates (Prs = Pearson correlation, ICC = intra class correlation).