

**ADDITIONAL TEXT “CONTROLLING BIAS AND INFLATION  
IN EPIGENOME- AND TRANSCRIPTOME-WIDE ASSOCIATION  
STUDIES USING THE EMPIRICAL NULL DISTRIBUTION”**

ABSTRACT. Here we proof that genomic control is equivalent to the use of an empirical null distribution. Furthermore, a sketch of a proof is given to show that the omission of a variable introduces bias and we give a brief description how we impute white blood cell counts.

CONTENTS

|                                                                         |   |
|-------------------------------------------------------------------------|---|
| 1. Genomic control is similar to an inflated or overdispersed null:     | 1 |
| 2. Unobserved covariates introduce bias and inflation                   | 2 |
| 2.1. Omission of a variable introduces bias                             | 2 |
| 3. Partial least squares for imputation of white blood cell composition | 3 |
| References                                                              | 3 |

1. GENOMIC CONTROL IS SIMILAR TO AN INFLATED OR OVERDISPERSED NULL:

To test the hypothesis  $H_{0i} : \beta_{1i} = 0$  we use the regular test statistic:

$$(1) \quad T_i = \frac{\hat{\beta}_{1i}}{\text{SE}(\hat{\beta}_{1i})} \sim t_{n-2} \approx \mathcal{N}(0, 1),$$

and assume that we can approximate the Student’s  $t$ -distribution by the standard normal.

Genomic control divides the test-statistic by the inflation factor before the calculation of  $P$  values. The following derivation shows that applying genomic control is identical to the use of an overdispersed or inflated null. Consider, the interest is in two-sided  $P$ -values  $U_i$  then:

$$(2) \quad \begin{aligned} U_i &= 2 \left[ 1 - \Phi \left( \left| \frac{T_i}{\lambda} \right| \right) \right] \\ &= 2 \left[ 1 - \Pr \left\{ Z \leq \left| \frac{T_i}{\lambda} \right| \right\} \right], \\ &= 2 [1 - \Pr \{ \lambda Z \leq |T_i| \}] \\ &= 2 [1 - \Pr \{ X \leq |T_i| \}] \end{aligned}$$

here,  $Z \sim \mathcal{N}(0, 1)$  with CDF given by  $\Phi(\cdot)$ . Furthermore,  $X \sim \mathcal{N}(0, \lambda^2)$ , represents the overdispersed or inflation normal distribution. Here  $\lambda$  is  $\sqrt{\lambda_{\chi^2_1}}$  as defined in the main text.

---

*Date:* January 4, 2017.

And in general if both inflation and bias are present.

$$\begin{aligned}
 (3) \quad U_i &= 2 \left[ 1 - \Phi \left( \left| \frac{T_i - \mu}{\sigma} \right| \right) \right] \\
 &= 2 [1 - \Pr \{ \sigma Z + \mu \leq |T_i| \}] \\
 &= 2 [1 - \Pr \{ X \leq |T_i| \}]
 \end{aligned}$$

here,  $\mu$  and  $\sigma$  represent the bias and inflation. Furthermore, now  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

## 2. UNOBSERVED COVARIATES INTRODUCE BIAS AND INFLATION

In a note from P. Rao [1] it was shown that the omission of a variable introduces bias and decreases the variance of all least squares estimates, i.e. introduces both bias and inflation.

**2.1. Omission of a variable introduces bias.** Here is a sketch of the proof for the introduction of bias, within the framework of the main text.

Considered the omission of the unobserved technical or biological covariates,  $\mathbf{W}$ . For the sake of simplicity, we assume there are no known covariates,  $\mathbf{Z}$ .

$$\begin{aligned}
 (4) \quad \mathbf{y}_j &= \mathbf{x}\tilde{\beta}_j + \tilde{\epsilon}_{.j} \\
 \mathbf{y}_j &= \mathbf{x}\beta_j + \mathbf{W}\gamma_j + \epsilon_{.j}
 \end{aligned}$$

The latter model is true but we are unaware of this and continue estimating the regression coefficient of interest,  $\tilde{\beta}_j$  of the former, misspecified, model.

$$\begin{aligned}
 (5) \quad \tilde{b}_j &= \frac{\mathbf{x}^T \mathbf{y}_j}{\mathbf{x}^T \mathbf{x}} \\
 &= \frac{\mathbf{x}^T (\mathbf{x}\beta_j + \mathbf{W}\gamma_j + \epsilon_{.j})}{\mathbf{x}^T \mathbf{x}} \\
 &= \beta_j + \frac{\mathbf{x}^T \mathbf{W}\gamma_j + \mathbf{x}^T \epsilon_{.j}}{\mathbf{x}^T \mathbf{x}}
 \end{aligned}$$

where, we substituted the true model for  $\mathbf{y}_j$ . Now, since  $E[\epsilon_{.j}] = 0$ , the expected regression coefficient is given by:

$$(6) \quad E[\tilde{b}_j] = \beta_j + \frac{\mathbf{x}^T \mathbf{W}}{\mathbf{x}^T \mathbf{x}} \gamma_j$$

and the bias is given by the last fraction, which can be interpreted as a weighted sum of correlations between the covariate of interest,  $\mathbf{x}$  and  $q$  omitted variables.

$$(7) \quad \sum_k^q \gamma_{jk} \text{cor}(\mathbf{x}, \mathbf{w}_k)$$

If all weights are zero or all correlations are zero than the bias will be zero too. The bias is not equal to zero if an omitted variable is confounded with the outcome, i.e., both  $\gamma_{jk}$  and  $\mathbf{x}^T \mathbf{W}$  are not zero.

### 3. PARTIAL LEAST SQUARES FOR IMPUTATION OF WHITE BLOOD CELL COMPOSITION

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 minority of samples lack CBC measurements. Since DNA methylation levels are informative of the white blood cell composition [2] we build a linear predictor to infer the white blood cell composition of those samples lacking WBC measurements.

Obviously, this model can not be fitted using ordinary least-squares, since  $p \gg n$ , we need some kind of regularization. Furthermore, the multivariate responses, white blood counts on five cell types, represents compositional data, i.e., the data are percentages that sum up to 100%.

We have chosen to use partial least-squares for fitting a model with cell counts as a multivariate response and the  $> 400,000$  CpGs age and sex as covariates. It is known that the WBCC is dependent on age and gender. The advantage of partial least-squares is that it both can handle multivariate responses and high-dimensional ( $p \gg n$ ) covariates. We used the R-package pls [3] to fit the model and optimize the number of pls-components using five-fold cross-validation. The fitted model was used to predict WBCC using the 450K data, age and sex of those samples lacking WBCC.

The pls-approach has been validated by splitting the data with WBCC available in a train and test set. Fit the pls-model on the train set and predict WBCC on the test set. Correlation (Pearson) between predicted and measured WBCC range from 0.86 – 0.37 for lymphocytes and basophils respectively, the intraclass correlation was 0.84 – 0.25.

This approach has been implemented in a R package `wbccPredictor`.

### REFERENCES

- [1] Rao, P.: Some notes on misspecification in multiple regression. *The American Statistician* **25**(5) (1971)
- [2] Houseman, E.A., *et al.*: DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012)
- [3] Mevik, B.H., Wehrens, R.: The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software* **18**(2) (2007)