Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/33717 holds various files of this Leiden University dissertation.

Author: Meijer, Rosa Janna
Title: Efficient multiple testing for large structured problems
Issue Date: 2015-06-30

Efficient Multiple Testing for Large Structured Problems

Rosa Janna Meijer

Cover design: Ridderprint BV, the Netherlands
Printing: Ridderprint BV, the Netherlands

©2015 Rosa Meijer, Rotterdam, the Netherlands.
All rights reserved. No part of this publication may be reproduced without prior permission of the author.

ISBN: 978-946-299-080-7

Efficient Multiple Testing for Large Structured Problems

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 30 juni 2015
klokke 13.45 uur

door

Rosa Janna Meijer
geboren te Rotterdam in 1985

PROMOTIECOMMISSIE

Promotores:	Prof. dr. J. J. Goeman · *Radboud Universitair Medisch Centrum, Nijmegen* Prof. dr. T. Stijnen
Overige leden:	Prof. dr. M. A. van de Wiel · *Vrije Universiteit Medisch Centrum en Vrije Universiteit, Amsterdam* Dr. A. Solari · *University of Milano-Bicocca , Milano* Dr. E. W. van Zwet

Table of Contents

1 Introduction **1**
- 1.1 Introduction . . . 1
- 1.2 Hypothesis Testing . . . 2
  - 1.2.1 Type of test . . . 2
  - 1.2.2 Multiple Testing . . . 4
- 1.3 Prediction modeling . . . 5
  - 1.3.1 Type of model . . . 6
  - 1.3.2 Predictive performance . . . 7
- 1.4 Computational feasibility . . . 8
- 1.5 Outline of the thesis . . . 9

2 Extending Hommel's procedure **13**
- 2.1 Introduction . . . 14
- 2.2 Hommel's procedure . . . 15
- 2.3 More efficient implementation of Hommel's procedure . . . 18
- 2.4 Confidence Sets . . . 20
  - 2.4.1 Determining whether an hypothesis can be rejected by the closed testing procedure in an efficient way . . . 22
  - 2.4.2 Algorithm . . . 26
- 2.5 Application . . . 28
- 2.6 Discussion . . . 33

3 A region-based multiple testing method **35**
- 3.1 Introduction . . . 36
- 3.2 Region hypotheses . . . 37
  - 3.2.1 A FWER controlling multiple testing procedure, based on the sequential rejection principle . . . 39
  - 3.2.2 Confidence sets for the number of true rejections in regions chosen post-hoc . . . 44
  - 3.2.3 A weighted-version of the region-method . . . 45
- 3.3 Algorithms . . . 46
  - 3.3.1 Dynamic programming algorithm to calculate ratios . . . 47
  - 3.3.2 Complexity . . . 48

3.3.3 Confidence statements . . . 51
3.4 Applications . . . 51
3.5 Discussion . . . 57
3.6 Appendix . . . 58

4 A multiple testing method for DAGs 61
4.1 Introduction . . . 62
4.2 The method . . . 63
4.2.1 Sequential rejection principle . . . 65
4.2.2 Testing procedure for DAGs with one-way logical relationships . 67
4.2.3 Using two-way logical relationships: a more powerful procedure . 70
4.2.4 Confidence sets for the number of true findings in sets chosen post-hoc . . . 73
4.3 Algorithms . . . 74
4.3.1 Maximizing the denominator . . . 75
4.3.2 Constructing confidence sets . . . 77
4.4 Applications . . . 77
4.4.1 Comparison of the approaches that can be used when two-way logical relationships are present . . . 78
4.4.2 Comparing the any-parent and all-parents rule to existing FWER controlling methods . . . 80
4.4.3 A simulation study to further explore the method's behavior . . . 82
4.5 Discussion . . . 85

5 Multiple testing of gene sets from GO 89
5.1 Introduction . . . 90
5.2 Structure of gene sets . . . 91
5.3 Appropriate multiple testing procedures . . . 94
5.4 More powerful FWER controlling methods . . . 98
5.5 Discussion . . . 103

6 Finding important variables in high-dimensional data 107
6.1 Introduction . . . 107
6.2 Selecting variables by fitting a prediction model . . . 108
6.2.1 Screening by penalized methods . . . 109
6.2.2 Screening by univariate selection . . . 113
6.2.3 Practical usefulness of methods possessing screening properties . 114
6.2.4 The Van de Vijver data set . . . 116
6.3 Selecting variables by testing individual covariates . . . 118
6.3.1 Methods for FWER control . . . 121
6.3.2 Methods for FDR control . . . 124
6.3.3 Confidence intervals for the number of true discoveries . . . 126

6.4 Reducing the number of variables beforehand and incorporating background knowledge . . . 128
6.5 Discussion . . . 129

7 Efficient approximate cross-validation 131
7.1 Introduction . . . 132
7.2 Approximation procedure . . . 133
7.2.1 Generalized linear models . . . 134
7.2.2 Cox' proportional hazards model . . . 138
7.3 Results . . . 140
7.3.1 Measure of predictive performance: the cross-validated log-likelihood . . . 141
7.3.2 Data sets . . . 141
7.3.3 Ridge regression . . . 141
7.3.4 Influence of n, k and λ on the accuracy of the approximations . . 144
7.3.5 Computation Time . . . 145
7.4 Discussion . . . 147

Bibliography 149

List of Publications 157

Samenvatting 159

Dankwoord 165

Curriculum Vitae 167