Universiteit
Leiden
The Netherlands

Statistical integration of diverse omics data

Bouhaddani, S. el

Citation
Bouhaddani, S. el. (2020, June 2). *Statistical integration of diverse omics data*. Retrieved from https://hdl.handle.net/1887/92366

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/92366

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/92366 holds various files of this Leiden University dissertation.

Author: Bouhaddani, S.
Title: Statistical integration of diverse omics data
Issue Date: 2020-06-02

Statistical integration of diverse omics data

Said el Bouhaddani

Cover design: Studio Antwan, Utrecht, the Netherlands
Printing: Ridderprint | www.ridderprint.nl

©2020 Said el Bouhaddani, Rotterdam, the Netherlands.
All rights reserved. No part of this publication may be reproduced without prior permission of the author.

ISBN: 978-94-6375-825-3

This thesis was funded by the EU under the FP7 project MIMOmics (Methods for Integrated Analysis of Multiple Omics Datasets), Grant/Award Number: FP7/Health/F5/2012/305280

Statistical integration of diverse omics data

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 2 juni 2020
klokke 15.00 uur

door

Said el Bouhaddani
geboren te Rotterdam in 1991

Promotor:	Prof. dr. J. J. Houwing-Duistermaat
Co-promotors:	Dr. H.-W. Uh · *Universitair Medisch Centrum Utrecht, Utrecht* Prof. dr. G. Jongbloed · *Technische Universiteit Delft, Delft*
Leden promotiecommissie:	Prof. dr. P. E. Slagboom Prof. dr. J. T. Kent · *University of Leeds, Leeds, Verenigd Koninkrijk* Prof. dr. T. Burzykowski · *Universiteit Hasselt, Hasselt, België*

Table of Contents

1 Introduction **1**
- 1.1 Background . . . 2
  - 1.1.1 Omics data characteristics . . . 3
- 1.2 Modeling the relationship . . . 4
  - 1.2.1 Unifying the data integration methods. . . . 5
  - 1.2.2 Estimating the data integration models . . . 7
- 1.3 Modeling the data-specific characteristics . . . 8
- 1.4 General outline of the thesis . . . 9

2 Evaluation of O2PLS in Omics data integration **15**
- 2.1 Abstract . . . 16
- 2.2 Introduction . . . 16
- 2.3 Methods . . . 18
  - 2.3.1 Previous methods . . . 18
  - 2.3.2 The O2PLS model . . . 18
  - 2.3.3 Proof of Optimality . . . 20
  - 2.3.4 Simulation Study . . . 22
  - 2.3.5 Availability of supporting data . . . 23
- 2.4 Results . . . 23
  - 2.4.1 Results of Simulation Study . . . 23
  - 2.4.2 Application to DILGOM data . . . 23
- 2.5 Discussion . . . 25
- 2.6 Tables . . . 28
- 2.7 Figures . . . 31

3 Integrating omics datasets with the OmicsPLS package **43**
- 3.1 Abstract . . . 44
- 3.2 Introduction . . . 44
- 3.3 Implementation . . . 45
  - 3.3.1 O2PLS Model . . . 45
  - 3.3.2 Implementation . . . 46
- 3.4 Results . . . 48
  - 3.4.1 Analysis of genetic and glycomic data . . . 48

3.4.2 Simulation study . . . 50
3.5 Discussion . . . 50
3.6 Conclusion . . . 52
3.7 Tables . . . 52
3.8 Figures . . . 54

4 Probabilistic Partial Least Squares: identifiability, estimation and application 59
4.1 Abstract . . . 60
4.2 Introduction . . . 60
4.3 Model and estimation . . . 61
4.3.1 The PPLS model . . . 61
4.3.2 Identifiability of probabilistic PLS . . . 62
4.3.3 Estimating the parameters . . . 63
4.4 Simulation study . . . 66
4.4.1 Results . . . 67
4.5 Data analysis . . . 69
4.6 Discussion . . . 73
4.7 Appendices for Chapter 4 . . . 76
4.8 Supplementary material for Chapter 4 . . . 81

5 Statistical integration of heterogeneous data with PO2PLS 117
5.1 Abstract . . . 118
5.2 Introduction . . . 118
5.3 PO2PLS: model and estimation . . . 120
5.3.1 The model . . . 120
5.3.2 Identifiability of PO2PLS . . . 121
5.3.3 Maximum Likelihood Estimation of the parameters . . . 122
5.4 Simulation study . . . 123
5.4.1 Results . . . 124
5.5 Analysis of heterogeneous omics data . . . 125
5.6 Discussion . . . 130
5.7 Appendices for Chapter 5 . . . 131
5.8 Supplementary material for Chapter 5 . . . 132

6 Statistical omics data integration in longitudinal family studies 149
6.1 Background . . . 150
6.2 Methods . . . 151
6.3 Data analysis results . . . 152
6.4 Discussion of the results . . . 153
6.5 Methodological future work . . . 156
6.5.1 A probabilistic model for supervised data integration in longitudinal family studies . . . 157
6.5.2 Further directions: extending the joint model . . . 158

Summary 163

Samenvatting 167

List of Publications **171**

Acknowledgments **173**

Curriculum Vitae **175**