Statistical methods for microarray data

Goeman, Jelle Jurjen

Samenvatting

In dit proefschrift worden statistische methoden ontwikkeld voor het analyseren van *microarray*-data. De microarray is een nieuwe technologie uit de moleculaire biologie, die onderzoekers in staat stelt metingen te doen aan genexpressie. Dit is het proces waarmee de informatie die in de genen ligt opgeslagen wordt gebruikt voor de productie van eiwitten. De activiteit van genexpressie kan worden gemeten via het RNA, de belangrijke tussenstap tussen gen en eiwit. Een microarray meet de concentratie van RNA behorend bij ieder specifiek gen en doet dit tegelijkertijd voor tienduizenden genen. Met een microarray is dus het patroon te zien van de gen-expressie van grote aantallen genen in een weefsel of een opgekweekte cellijn.

Door microarrays te vergelijken tussen verschillende typen weefsel, tussen weefsels van verschillende patienten of tussen cellijnen die verschillend behandeld zijn, kunnen allerlei wetenschappelijke vragen beantwoord worden. Interessante vragen zijn er bijvoorbeeld op het gebied van diagnose en prognose. Het vinden van gen-expressiepatronen die onderscheid maken tussen ernstige en minder ernstige vormen van ziekte kan de kwaliteit van diagnoses verbeteren, en daarmee de kwaliteit van de behandeling laten toenemen. Als de microarray bijvoorbeeld gebruikt kan worden om de overleving van borstkankerpatiënten nauwkeuriger te voorspellen zou een groot aantal patiënten een onnodige chemotherapie bespaard kunnen worden. Andere onderzoeksvragen die mogelijk worden gemaakt door microarrays gaan over de functie van genen: door uit te vinden van welke genen de gen-expressie verandert als cellijnen een bepaalde behandeling krijgen, kan iets worden afgeleid over de functie van die genen.

Een statistisch probleem bij het beantwoorden van deze vragen is de hoge dimensionaliteit van de microarray, gekoppeld aan de kleine steekproefgrootte. In een typisch klinisch onderzoek worden microarrays gemaakt van enkele tientallen tot hoogstens enkele honderden patiënten, terwijl voor iedere patiënt de gen-expressie gemeten is van tienduizenden genen. Deze hoge dimensionaliteit leidt tot problemen bij het toepassen van klassieke statistische methoden. Bij het zoeken naar genen die een verschillende gen-expressie hebben onder verschillende experimentele condities moet een zo groot aantal statistische toetsen worden uitgevoerd, dat bekende methoden om te corrigeren voor meervoudig toetsen niet meer goed functioneren. Bij het zoeken naar voorspelregels

om kenmerken van patienten te voorspellen, treedt het verschijnsel van *overfit* op: er zijn vele voorspelregels te vinden die de kenmerken van de onderzochte patiënten perfect voorspellen, maar waarvan de prestaties op nieuwe gevallen allerminst gegarandeerd zijn. De uitdaging die het oplossen van deze problemen biedt, heeft al geleid tot een groot aantal nieuwe statistische methoden.

De statistische methoden die in dit proefschrift ontwikkeld worden, maken zoveel mogelijk gebruik van inhoudelijke kennis uit de biologie, in het bijzonder annotatie van genen, om de kwaliteit en interpreteerbaarheid van de conclusies the verhogen. Annotatie koppelt genen aan de informatie die reeds over deze genen in de literatuur bekend is, bijvoorbeeld in welke celprocessen het gen betrokken is, met welke functies, organen of ziekten het gen is geassocieerd of op welk chromosoom het gen gelokaliseerd is. Een belangrijk concept hierbij is het begrip *pathway*: een *pathway* is een groep genen die met dezelfde functie geassocieerd wordt.

De belangrijkste nieuwe methode in dit proefschrift is de *GlobalTest*-methodologie. Deze wordt uiteengezet in de hoofdstukken 2 tot en met 5. Deze methode biedt een statistische toets die onderzoekers in staat stelt om microarray data te analyseren op het niveau van pathways, in plaats van op het niveau van individuele genen. De onderzoeker gaat dan niet op zoek naar genen waarvan de expressie geassocieerd is met bepaalde kenmerken van patiënten, maar naar pathways waarvan de expressie met deze kenmerken geassocieerd is. Dit is een andere manier van werken, die vaak een tegengestelde onderzoeksvraag heeft. Methoden die zoeken naar individuele genen hebben meestal tot doel de functie van het gen af te leiden uit het kenmerk waarmee de expressie van dat gen associatie vertoont. Als bijvoorbeeld de expressie van genen in gekweekte cellen sterk verandert na kortdurend verhitten van deze cellen, zullen die genen waarschijnlijk een functie hebben bij het herstellen van celschade na hitte. Omgekeerd probeert een methode die zoekt naar pathways juist iets te leren over biologie achter een geobserveerd kenmerk, vanuit de bekende functies van de pathways. Als bijvoorbeeld blijkt dat de expressie van de apoptose-pathway (die de geprogrammeerde celdood regelt) in tumorweefsel dat zich heeft uitgezaaid duidelijk anders is dan in tumorweefsel dat zich niet heeft uitgezaaid, kan geconcludeerd worden dat een storing in de apoptose een stap is in het proces van uitzaaien van tumoren.

Hoofdstuk 2 introduceert de GlobalTest-methodologie die kan toetsen of het gen-expressiepatroon van een bepaalde pathway geassocieerd is met een bepaalde responsvariabele. De details van de methode worden uitgewerkt voor het geval de respons ofwel twee mogelijke waarden aanneemt, ofwel een normaal verdeelde grootheid is.

Hoofdstuk 3 geeft een uitbreiding van dezelfde methodologie naar de si-

tuatie waarin gezocht wordt naar pathways die geassocieerd zijn met overlevingsduur. Het introduceert bovendien de mogelijkheid te corrigeren voor de effecten van verstorende variabelen, wat van groot belang is bij observationeel onderzoek.

Hoofdstuk 4 werkt de wiskunde uit die nodig is om de GlobalTest-methode toe te passen op een responsvariabele die meer dan twee ongeordende waarden aanneemt. De toets die dit artikel presenteert, wordt niet beschreven als een toets voor microarray data, maar in de vorm van een *goodness-of-fit* toets voor het mutinomiale logistische regressiemodel. Dit is een toets waarmee kan worden onderzocht of een dergelijk multinomiaal logistisch model een dataset adequaat beschrijft. Wiskundig gezien is deze toets dezelfde als de toets die nodig is om de GlobalTest-methode te generaliseren naar responsvariabelen met meerdere uitkomstcategorieën.

Hoofdstuk 5 plaatst de toetsen van de vorige drie hoofdstukken in een algemener kader door te laten zien dat ze deel uitmaken van een brede klasse van toetsen die een eenvoudige nulhypothese toetsen tegen een hoogdimensionaal alternatief. Het laat bovendien zien dat dit soort toetsen gemiddeld in een omgeving van de nulhypothese een optimaal onderscheidend vermogen heeft.

Hoofdstuk 6 staat buiten de GlobalTest-methodologie. Het behandelt het probleem hoe een klinische variabele van een patiënt te voorspellen uit de microarray data van die patiënt. Ook hier wordt zoveel mogelijk gebruik gemaakt van kennis over de microarray-data om een goede voorspelregel te construeren. Hiertoe wordt een model van de simultane verdeling van de gen-expressiemetingen en de te voorspellen uitkomstvariabele geconstrueerd. Dit model is gebouwd op de aanname dat er een klein aantal onobserveerbare onderliggende variabelen bestaat, dat zowel de gen-expressiemetingen beinvloedt als de uitkomstvariabele, en dat alle gemeten waarden gepaard gaan met ruis. Op basis van deze eenvoudige aannamen wordt een voorspelregel geconstrueerd die goede eigenschappen heeft in dit model.

Hoofdstuk 7, tenslotte, gaat in op het belangrijke onderwerp van visualisatie van microarray data. Een veelgebruite visualisatiemethode als de puntenwolk geeft al snel een vertekend beeld als er duizenden punten in één diagram weergegeven moeten worden. Het is dan beter om in plaats van een puntenwolk een kleurenweergave van de dichtheid te presenteren, omdat een dergelijke weergave veel duidelijker aangeeft waar de massa van de punten zich bevindt. In het hoofdstuk wordt een snel algoritme gegeven om een dergelijke visualisatie te genereren.

Bibliography

Abraham, B. and G. Merola (2005). Dimensionality reduction approach to multivariate prediction. *Computational Statistics & Data Analysis 48*(1), 5–16.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics 20*(4), 578–580.

Albert, A. and E. Harris (1987). *Multivariate interpretation of clinical laboratory data*. New York: Marcel Dekker.

Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. New York: Springer.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 25*(1), 25–29.

Azzalini, A. and A. Bowman (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society Series B-Methodological 55*(2), 549–557.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2004). Prediction by supervised principal components. Technical report, Dept. of Statistics, Stanford University.

Bartholomew, D. J. and M. Knott (1999). *Latent variable models and factor analysis* (2nd ed.). Arnold.

Beer, D. G., S. L. R. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. A. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine 8*(8), 816–824.

Beissbarth, T. and T. P. Speed (2004). GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics 20*(9), 1464–1465.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*(4), 1165–1188.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian theory*. Chichester: Wiley.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*(2), 185–193.

Boyle, E. I., S. A. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock (2004). GO-TermFinder: open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics 20*(18), 3710–3715.

Brown, P. J. (1993). *Measurement, regression, and calibration*. Oxford: Oxford University Press.

Burnham, A. J., J. F. MacGregor, and R. Viveros (1999a). Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems 48*(2), 167–180.

Burnham, A. J., J. F. MacGregor, and R. Viveros (1999b). A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of Chemometrics 13*(1), 49–65.

Burnham, A. J., J. F. MacGregor, and R. Viveros (2001). Interpretation of regression coefficients under a latent variable regression model. *Journal of Chemometrics 15*(4), 265–284.

Burnham, A. J., R. Viveros, and J. F. MacGregor (1996). Frameworks for latent variable multivariate regression. *Journal of Chemometrics 10*(1), 31–45.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical statistics*. Boca Raton: Chapman & Hall.

De Menezes, R. X., J. M. Boer, and J. C. van Houwelingen (2004). Microarray data analysis: a hierarchical t-test to handle heteroscedasticity. *Applied Bioinformatics 3*, 229–235.

Díaz-Uriarte, R. (2005). Supervised methods with genomic data: a review and cautionary review. In F. Azuaje and J. Dopazo (Eds.), *Data Analysis and Visualization in Genomics and Proteomics*, pp. 193–214. Chichester: Wiley.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association 97*(457), 77–87.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science 18*(1), 71–103.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–451.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*(456), 1151–1160.

Eilers, P. (1994). Smoothing and interpolation with finite differences. In P. Heckbert (Ed.), *Graphics Gems*, Volume IV, pp. 241–250. London: Academic Pres.

Eilers, P. (2003). A perfect smoother. *Analytical Chemistry 75*(14), 3631–3636.

Eilers, P., J. Boer, G. van Ommen, and J. C. van Houwelingen (2001). Classification of microarray data with penalized logistic regression. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (Eds.), *Proceedings of SPIE*, Volume 4266, pp. 187–198.

Eilers, P. H. C. and J. J. Goeman (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics 20*(5), 623–628.

Ein-Dor, L., I. Kela, G. Getz, D. Givol, and E. Domany (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics 21*(2), 171–178.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America 95*(25), 14863–14868.

Ewis, A. A., Z. Zhelev, R. Bakalova, S. Fukuoka, Y. Shinohara, M. Ishikawa, and Y. Baba (2005). A history of microarrays in biomedicine. *Expert Review of Molecular Diagnostics 5*(3), 315–328.

Fleming, T. R. and D. P. Harrington (1991). *Counting processes and survival analysis*. New York: Wiley.

Ge, Y. C., S. Dudoit, and T. P. Speed (2003). Resampling-based multiple testing for microarray data analysis. *Test 12*(1), 1–77.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. H. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 5*(10), R80.

Goeman, J. J. and J. Oosting (2005). *Globaltest: testing association of a pathway with a clinical variable*. R package version 3.2.0.

Goeman, J. J., J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and J. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21*(9), 1950–1957.

Goeman, J. J., S. A. van de Geer, F. de Kort, and J. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*(1), 93–99.

Goeman, J. J., S. A. van de Geer, and J. C. van Houwelingen (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society Series B-Statistical Methodology 68*, to appear.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning*. Springer.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hosmer, D. W. and S. Lemeshow (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Houwing-Duistermaat, J. J., B. H. F. Derkx, F. R. Rosendaal, and J. C. van Houwelingen (1995). Testing familial aggregation. *Biometrics 51*(4), 1292–1301.

Huber, W., A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*(Suppl. 1), S96–S104.

Hwang, J. T. G. and D. Nettleton (2003). Principal components regression with data-chosen components and related methods. *Technometrics 45*(1), 70–79.

Imhof, J. P. (1961). Computing distribution of quadratic forms in normal variables. *Biometrika 48*(3-4), 419–426.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Jenssen, T. K., W. P. Kuo, T. Stokke, and E. Hovig (2002). Associations between gene expressions in breast cancer and patient survival. *Human Genetics 111*(4-5), 411–420.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Kerr, M. K., M. Martin, and G. A. Churchill (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7*(6), 819–837.

Klein, J. P. and M. L. Moeschberger (1997). *Survival analysis: techniques for truncated data*. New York: Springer.

Le Cessie, S. and J. C. Van Houwelingen (1991). A goodness-of-fit test for binary regression models based on smoothing methods. *Biometrics 47*, 1267–1282.

Le Cessie, S. and J. C. van Houwelingen (1992). Ridge estimators in logistic regression. *Applied Statistics 41*(1), 191–201.

Le Cessie, S. and J. C. van Houwelingen (1995). Testing the fit of a regression-model via score tests in random effects models. *Biometrics 51*(2), 600–614.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Lesaffre, E. and A. Albert (1989). Multiple-group logistic regression diagnostics. *Applied Statistics 38*, 425–440.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Magnus, J. R. and H. Neudecker (1999). *Matrix differential calculus with applications in statistics and econometrics* (revised ed.). Wiley.

Mansmann, U. and R. Meister (2005). Testing differential gene expression in functional

groups: Goeman's global test versus an ANCOVA approach. *Methods of Information in Medicine 44*(3), 449–453.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall.

McLachlan, G. J., R. W. Bean, and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics 18*(3), 413–422.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop (2003). PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*(3), 267–273.

Nguyen, D. V. and D. M. Rocke (2002a). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics 18*(12), 1625–1632.

Nguyen, D. V. and D. M. Rocke (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics 18*(1), 39–50.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research 27*(1), 29–34.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford: Clarendon.

Pawitan, Y., J. Bjohle, S. Wedren, K. Humphreys, L. Skoog, F. Huang, L. Amler, P. Shaw, P. Hall, and J. Bergh (2004). Gene expression profiling for prognosis using Cox regression. *Statistics in Medicine 23*(11), 1767–1780.

Pigeon, J. G. and J. F. Heyse (1999). An improved goodness-of-fit statistic for probability prediction models. *Biometrical Journal 41*, 71–82.

R Development Core Team (2005). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reiner, A., D. Yekutieli, and Y. Benjamini (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics 19*(3), 368–375.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science 270*(5235), 467–470.

Shevade, S. K. and S. S. Keerthi (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19*(17), 2246–2253.

Simon, R., E. Korn, L. McShane, M. Radmacher, G. Wright, and Y. Zhao (2003). *Design*

and analysis of DNA microarray investigations. New York: Springer.

Simonoff, J. (1996). *Smoothing methods in statistics*. New York: Springer.

Smid, M. and L. C. J. Dorssers (2004). GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics 20*(16), 2618–2625.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology 3*(1), article 3.

Sohler, F., D. Hanisch, and R. Zimmer (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics 20*(10), 1517–1521.

Solomon, H. and M. A. Stephens (1978). Approximations to density functions using Pearson curves. *Journal of the American Statistical Association 73*(361), 153–160.

Speed, T. E. (2003). *Statistical analysis of gene expression microarray data*. Boca Raton: Chapman & Hall.

Stone, M. and R. J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least-squares, partial least-squares and principal components regression. *Journal of the Royal Statistical Society Series B-Methodological 52*(2), 237–269.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Methodological 64*, 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America 100*(16), 9440–9445.

Tan, Y. X., L. M. Shi, W. D. Tong, and C. Wang (2005). Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Research 33*(1), 56–65.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological 58*(1), 267–288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine 16*(4), 385–395.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America 99*(10), 6567–6572.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America 98*(9), 5116–5121.

Van de Vijver, M. J., Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen,

A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine 347*(25), 1999–2009.

Van Houwelingen, J. C. (1984). Principal components of large matrices with missing elements. In P. Mandl and M. Hukov (Eds.), *Asymptotic statistics 2: Proceedings of the 3rd Prague symposium on asymptotic statistics 29 August-2 September 1983*, pp. 295–302. North Holland.

Van Houwelingen, J. C. (2001). Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica 55*(1), 17–34.

Van Houwelingen, J. C., T. Bruinsma, A. A. M. Hart, L. J. van t Veer, and L. F. A. Wessels (2005). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine 24*, to appear.

Van Houwelingen, J. C. and R. M. Schipper (1981). The efficiency of a test based on the asymptotic distribution of the MLE for a linear functional relationship. *Mathematische Operationsforschung und Statistik, Series Statistics 12*(1), 21–30.

Van 't Veer, L. J., H. Y. Dai, M. J. van de Vijver, Y. D. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*(6871), 530–536.

Verweij, P. J. M., J. C. van Houwelingen, and T. Stijnen (1998). A goodness-of-fit test for Cox's proportional hazards model based on martingale residuals. *Biometrics 54*(4), 1517–1526.

Vingron, M. (2001). Bioinformatics needs to adopt statistical thinking. *Bioinformatics 17*(5), 389–390.

Wall, M. M. and R. F. Li (2003). Comparison of multiple regression to two latent variable techniques for estimation and prediction. *Statistics in Medicine 22*(23), 3671–3685.

Wentzell, P. D., D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics 11*(4), 339–366.

Westfall, P. H. and S. S. Young (1989). P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association 84*(407), 780–786.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society 41*, 63–75.

Wigle, D. A., I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B. J. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd, and M. S. Tsao (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research 62*(11), 3005–3008.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn (1984). The collinearity problem in linear regression: the partial least-squares (PLS) approach to generalized inverses. *SIAM*

Journal on Scientific and Statistical Computing 5(3), 735–743.

Wright, G. W. and R. M. Simon (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics 19*(18), 2448–2455.

Wu, Z. J., R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association 99*(468), 909–917.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*(4), e15.

Zeeberg, B. R., W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology 4*(4), R28.

Zhang, B., D. Schmoyer, S. Kirov, and J. Snoddy (2004). GO Tree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics 5*, 16.

Zhang, J. (2004). *GO: a data package containing annotation data for GO*. R package version 1.6.5.

Curriculum Vitae

De auteur van dit proefschrift werd geboren op 24 juni 1976 in Leiderdorp. Hij bezocht vanaf 1988 het Stedelijk Gymnasium te Leiden, waar hij in 1994 zijn V.W.O. diploma behaalde. In datzelfde jaar begon hij aan de Universiteit Leiden met de studie wiskunde, die hij in 2001 afrondde. De doctoraalcriptie die hieruit voortkwam, getiteld *Using survival to predict survival*, werd in 2002 bekroond met de scriptieprijs van de Vereniging voor Statistiek, en is in verkorte vorm gepubliceerd in *Statistica Neerlandica*. In dezelfde periode studeerde de auteur ook geschiedenis aan de Universiteit Leiden. Deze studie werd in 2001 *cum laude* afgerond met een doctoraalscriptie op het gebied van de historische methodologie, getiteld *Grondslagen van de vergelijkende methode*.

Het onderzoek dat leidde tot dit proefschrift werd uitgevoerd tussen 2001 en 2005, toen de auteur als promovendus verbonden was aan de afdeling Medische Statistiek en Bioinformatica van het LUMC en aan het Mathematisch Instituut van de Universiteit Leiden. De resultaten van het onderzoek werden eerder gepresenteerd op verschillende conferenties, workshops en colloquia, onder andere in Freiburg, Aarhus, Kaunas, Diepenbeek, Wye, Heidelberg, Boston, Marseille en Leicester. De presentatie op de laatste conferentie werd met een prijs bekroond. In deze periode heeft de auteur ook meegewerkt aan de organisatie van de workshop *On high-dimensional data*, die gehouden werd aan het Lorentz Center in Leiden in september 2002. Daarnaast was hij van 2003 tot 2005 secretaris van het Leids Promovendi Overleg.

Op dit moment werkt de auteur als wetenschappelijk onderzoeker aan de afdeling Medische Statistiek en Bioinformatica van het LUMC.