SUOMEN TILASTOSEURAN VUOSIKIRJA 2015–2016

ÅRSBOK FÖR STATISTISKA SAMFUNDET I FINLAND 2015–2016

THE YEARBOOK OF THE FINNISH STATISTICAL SOCIETY 2015–2016

2016

SISÄLLYSLUETTELO

ESIMIEHEN PALSTA

Jyrki Möttönen

Suomen Tilastoseuran sääntöjen toisessa pykälässä kerrotaan seuramme tarkoituksesta seuraavasti:

> 2 § Seuran tarkoituksena on edistää tilastotiedettä sekä toimia tilastotieteilijöiden ja muiden tilastoalasta kiinnostuneiden henkilöiden yhdyssiteenä. Tarkoituksensa toteuttamiseksi seura tukee jäsentensä tilastollista tutkimustyötä, toimeenpanee esitelmä- ja keskustelutilaisuuksia sekä edistää muullakin vastaavalla tavoin jäsentensä ammattitaidon kohottamispyrkimyksiä, harjoittaa julkaisutoimintaa ja on yhteistyössä muissa maissa toimivien vastaavien yhdistysten kanssa.

Tällä hetkellä yksi tärkeimmistä tilastotieteen edistämistehtävistä on tilastotieteen tunnettuuden lisääminen ja tärkeyden mainostaminen tiedeyhteisön ulkopuolella. Tilastotieteestä valmistuvien hyvistä työnäkymistä huolimatta pääaineena tilastotiedettä opiskelemaan hakeneiden määrä on ollut vuodesta toiseen pieni. Yhtenä syynä tilastotieteen huonoon suosioon on todennäköisesti se, että tiedemaailman ulkopuolella tilastotieteestä tiedetään melko vähän ja mielikuvat voivat olla myös vanhakantaisen värittyneitä. Tilastotiede saatetaan samaistaa pelkkään tilastojen tekemiseen, mihin on varmastikin osasyynä suomenkielinen harhaanjohtava nimi *tilasto*tiede. Tilastoseuran jäsenistöltä toivottaisiin tässä suhteessa aktiivisuutta. Mainostakaa omilla tahoillanne tilastotiedettä mielenkiintoisena ja monipuolisena alana sekä oikokaa virheellisiä käsityksiä! Tehdään yhdessä tilastotieteestä yhteiskunnallisesti arvostettu ja suosittu tieteenala!

Tilastoseuran toiminta pohjautuu lähes täysin vapaaehtoistyöhön oman palkkatyön ohella. Vapaaehtoistyön vuoksi jäsenmaksut on pystytty pitämään alhaisina mutta toisaalta täysipäiväisten työntekijöiden puuttuessa seuralla ei ole käytännön valmiuksia toiminnan laajentamiseen. Näistä selvistä rajoitteista huolimatta tavoitteenamme on edelleen toimia kaikkien tilastotieteilijöiden ja tilastotieteestä kiinnostuneiden henkilöiden yhdyssiteenä.

Tilastopäivät

Tilastopäivät järjestettiin Helsingin Meilahdessa 20.–21.8.2015 (Biomedicum Helsinki 1, luentosali 2, Haartmaninkatu 3). Teemaksi oli tällä kertaa valikoitunut *Big Data in Biological and Medical Research.* Pääpuhujana oli yksi alan merkittävimmistä nimistä, professori Mark Daly Harvardin yliopistosta. Dalyn lisäksi puhujina oli useita ulkomaisia ja kotimaisia tilastotieteen huippuja. Tilastopäivien konferenssipäivällinen järjestettiin Ravintola Lasipalatsin kabinetti Palmuhuoneessa.

Suomen Tilastoseura kiittää SAS Instituuttia heidän antamastaan merkittävästä taloudellisesta tuesta.

Palkinnot

Seura jakaa kahden vuoden välein Leo Törnqvist -palkinnon parhaalle suomalaisessa yliopistossa tai korkeakoulussa hyväksytylle tilastotieteen pro gradu -tutkielmalle. Palkintoon kuuluu stipendi, jonka suuruus on tällä hetkellä 1 000 euroa. Tilastoseura valitsi vuosina 2013-2014 hyväksytyistä pro gradu -tutkielmista parhaimmaksi Turun yliopistosta valmistuneen Joni Virran työn ”Some tools for linear dimension reduction”. Palkinnon saaja julkistettiin Tilastopäivien yhteydessä 21.8.2015.

Vuoden 2015 Eino H. Laurila -kansantulomitali myönnettiin valtiotieteen maisteri Helvi Kinnuselle, filosofian tohtori Ilmo Mäenpäälle ja valtiotieteen maisteri Olli Savelalle. Mitalit myönnettiin tunnustuksena heidän aktiivisesta työstään kansantalouden tilinpidon käytön ja tunnettuuden edistäjinä.

Suomen Tilastoseuran joka kolmas vuosi myöntämä Gunnar Modeen -mitali myönnettiin 2016 PhD Eva Elversille. Elvers on vaikuttanut merkittävästi virallisen tilastotoimen menetelmien kehittämiseen sekä kotimaassaan Ruotsissa että kansainvälisesti. Mitali luovutettiin Pohjoismaisen tilastokokouksen yhteydessä Tukholmassa 24.8.2016.

Iltapäiväseminaarit

Suomen Tilastoseura järjesti 11.6.2015 iltapäiväseminaarin aiheesta “Arkaluonteiset asiat tilastollisessa tutkimuksessa”. Paikkana oli Helsingin yliopiston päärakennuksen vanhan puolen auditorio XV. Puhujina olivat Jouni Kuha (London School of Economics), vanhempi tutkija Markku Heiskanen (Heuni) ja yliaktuaari Päivi Hokka (Tilastokeskus). Kommenttipuheenvuoron piti professori emeritus Seppo Laaksonen.

Kansainvälinen toiminta

Tilastoseuran varaesimiehen Ari Jaakolan terveiset ISI:n maailmankongressista:

Kansainvälisen tilastoinstituutin (ISI) maailmankongressin järjestettiin Brasilian Rio de Janeirossa 26.–31.7.2015. Joka toinen vuosi järjestettävä kongressin oli järjestyksessään kuudeskymmenes. Kongressin ohjelma oli perinteiseen tapaan runsas ja monipuolinen. Ohjelma koostui lähes kolmestasadasta sessiosta, joiden aiheet kattoivat käytännössä koko tilastotieteen kentän. Ajankohtaisia ja paljon keskustelua herättäneitä aiheita olivat esimerkiksi suurten tietoaineistojen käsittely ja analysointi, tiedonkeruumenetelmät sekä yleensä tilastotieteilijöiden asema tiedon tuottajina. Monissa puheenvuoroissa tuotiin esiin huoli siitä, että esimerkiksi matemaatikot ja tietojenkäsittelytieteilijät ovat nousseet monissa yhteyksissä esiin tilastotieteilijöitä näkyvämmin, vaikka juuri tilastotieteen eri osa-alueiden osaamisella tuntuisi olevan suuri kysyntä tiedon tuotannossa, analysoinnissa ja esittämisessä. Tieteellisten esitysten ohella ohjelma koostui lukuisista tieteellisten seurojen ja järjestöjen tapaamisista, varsinaisen kongressin rinnalla järjestetyistä satelliittikonferensseista, koulutustilaisuuksista. Myös vapaa-ajan ohjelmatarjonta oli runsas. Suomesta paikalla oli kaikkiaan noin viitisentoista osanottajaa.

Tulevaisuuden toimintaa

Keväällä 2017 Tilastoseura on järjestämässä iltapäiväseminaarin Smart City -aihepiiristä. Syksyllä 2017 on iltapäiväseminaari aiheesta ”Väestötutkimusten kato”.

Vuonna 2017 Tilastoseura on mukana ainakin kahden tärkeän tilastotieteen kokouksen järjestelyissä. Ensimmäinen kokous on toukokuussa Turussa järjestettävä Tilastopäivät. Aiheena on tällä kertaa *Missing Data* ja pääpuhujaksi on lupautunut yksi alan huipuista, professori Niels Keiding Kööpenhaminan yliopistosta. Toinen kokous on Helsingissä 24.–28.7.2017 järjestettävä 31st European Meeting of Statisticians (EMS2017). EMS on tärkein tilastotieteen ja todennäköisyyslaskennan konferenssi Euroopassa. Se järjestetään noin kahden vuoden välein ja sen sponsorina toimii *European Regional Committee of the Bernoulli Society.*

SIHTEERIN KOMMENTTI

Paula Bergman

Kaksi vuotta Suomen Tilastoseuran sihteerinä on kulunut kuin siivillä. Tehtävään ryhtyessäni en tiennyt tarkalleen, mihin soppaan olin lusikkaani työntämässä. Edellinen sihteeri, Kaisa Mäntysaari, perehdytti minut kuitenkin tehtävään hyvin, iso kiitos siitä! Seurassa vastuullani on ollut mm. tiedottamista, yhteydenpitoa jäsenistön kanssa, verkkosivujen päivitystä, jäsenrekisterin ylläpitoa sekä pöytäkirjojen laadintaa. Lisäksi kahteen vuoteen on mahtunut Tilastopäivien suunnittelua ja valmistelua, kuin myös tämän vuosikirjan kokoamista. Paljon uutta tietoa ja antoisia kokemuksia on tarttunut matkan varrella mukaan!

Tämän vuoden jäsenistön kannalta merkittäviä muutoksia ovat vaihtunut sähköpostiosoite, Tilastoseuran sosiaaliseen mediaan liittyminen, sekä uudistuvat verkkosivut. Olemme pyrkineet seuran rahastonhoitajan, Emma Kämäräisen, kanssa luomaan uudesta verkkosivusta kattavan ja paremmin jäsenistöämme palvelevan. Lisäksi seuran kuulumisia voi nykyään seurata Twitterissä (@tilastoseura1), sekä Facebookissa (Suomen Tilastoseura). Pyrimme julkaisemaan näiden kanavien kautta myös tietoa jäsenistöämme mahdollisesti kiinnostavista tapahtumista.

Haluan kiittää molempien kausien hallituksia mukavasti sujuneesta yhteistyöstä, sekä erityisesti esimiestä Jyrki Möttöstä seuran kunniakkaasta luotsaamisesta vuodesta toiseen. Iso kiitos kuuluu myös Kati Tiirikaiselle, joka on auttanut todella paljon verkkosivujen kanssa. Kiitokset myös tämän vuosikirjan taittajalle Hilkka Lehtoselle.

Katsotaan, mitä mielenkiintoista tulevaisuus tuo tullessaan!

To make, or not to make any moment assumptions?

Klaus Nordhausen[a], Esa Ollila[b], and Sara Taskinen[c]
[a]Department of Mathematics and Statistics, University of Turku
[b]Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University
[c]Department of Mathematics and Statistics, University of Jyväskylä

Abstract

In August 2016 Professor Hannu Oja retired from his post as a Professor of Statistics at the University of Turku. Hannu can be regarded as one of the most influential statisticians in Finland. He has served as professor in four Finnish universities and as a Visiting Professor at three foreign universities. He has also held several appointments granted by the Academy of Finland including the highly respected Academy Professorship in 2008-2012. In this paper we will give a brief overview on Hannu's research achievements from the past four decades.

1 Introduction

In April 2016 Hannu Oja gave his last lecture as a Professor of Statistics at the Department of Mathematics and Statistics, University of Turku. In the past 45 years, he has lectured dozens of theoretical as well as applied statistics courses in five Finnish Universities. As he is an enjoyable and excellent teacher and supervisor, his retirement will leave an empty space to statistics teaching in Finland. On the other hand, Hannu's retirement will give him freedom to do what he enjoys the most, that is, research. From September 2016 on he will continue research work as a Professor Emeritus at the University of Turku. He also continues supervising doctoral students, having still several students at the time of writing.

Hannu Oja's academic career started in early 80's when he published his first paper "On location, scale, skewness and kurtosis of univariate distributions" in the *Scandinavian Journal of Statistics*. When browsing his latest publications, like the recently submitted paper "Joint use of third and fourth cumulants in independent component analysis" coauthored by M.Sc. Joni Virta and Dr. Klaus Nordhausen, one may get a feeling that not much has happened in the past 40 years. However, a lot has happened. Soon after publishing his first papers on descriptive statistics, Hannu started his journey in the world of nonparametric and robust methods and supervised eight PhD theses in this field. In 2006, after twenty years of developing methods that try to avoid moment assumptions among other things, Hannu Oja published his first paper on independent component analysis (ICA) based on fourth-order moments and since then his research work has focused on ICA and statistical dimension reduction methods. He has supervised two theses in this field.

In the following we will take a closer look at Hannu Oja's academic career. In Section 2 we review Hannu Oja's biography, and in Section 3, we summarize some achievements in three of his main research areas. Section 4 shortly describes how Hannu's 65th birthday was celebrated. The reference section contains then a list of all scientific works Hannu has published so far.

2 Brief biography of Hannu Oja

Hannu Frans Vilhelm Oja[1] was born in 1950 in Jämsä, Finland. He is married to Ritva Oja (maiden name Paavola) and has three children. He got his certificate of matriculation from the Tampere Classical Lyceum in 1969, the M.Sc. degree in statistics from the University of Tampere in 1973, and Ph.Lic. degree in statistics in 1978 from the University of Oulu. He received his Ph.D. degree in statistics from the University of Oulu in 1981 under the guidance of Prof. Elja Arjas. Hannu Oja is an Adjunct Professor of Statistics of University of Oulu, University of Tampere, University of Jyväskylä and Tampere University of Technology.

From 1975 to 1981 Hannu Oja served as a Research Assistant of Statistics, from 1981 to 1988 as a Lecturer of Statistics, and from 1988 to 1993 as a Senior Lecturer of Statistics, at the Department of Mathematical Sciences of the University of Oulu. From 1996 to 1997 he was appointed as an Associate Professor of Financial Mathematics and Statistics at the Turku School of Economics, from 1997 to 1998 as an Associate Professor of Statistics at the University of Oulu, from 1998 to 2004 as a Full Professor of Statistics at the University of Jyväskylä, from 2005 to 2012 as a Professor of Biometry at the University of Tampere and from 2013 to August 2016 as a Professor of Statistics at the Department of Mathematics and Statistics, University of Turku. He has also served from 1993 to 1998 as an acting Professor of Statistics at the University of Oulu and held numerous appointments granted by the Academy of Finland, including Research Assistant (1980-1981), Junior Researcher (1983-1986), Senior Scientist (1995-1996, 1999-2000 and 2004-2005) and the prestigious Academy Professorship (2008-2012). In 1991 he was a Visiting Professor of Statistics at the Department of Statistics, Pennsylvania State University, State College, PA, in 1993 he was as a Visiting Professor at the University of Bern, Switzerland, and in 2005 he was a visiting Kolmogorov-Professor at Moscow State University. Hannu

[1] The picture shows Hannu in Thailand 2012 and is reprinted with kind permission from Ritva Oja©.

Oja is a fellow of *IMS*, the *Institute of Mathematical Statistics*, a member of *Bernoulli Society* and *Finnish Academy of Science and Letters*.

Hannu Oja has served as an Associate Editor for various statistical journals, including *Scandinavian Journal of Statistics*, *Journal of Statistical Planning and Inference* and *Statistics and Probability Letters*, and is a member of the steering committee of *ICORS*, *International Conferences on Robust Statistics*. His research interests in statistics include nonparametric and robust statistics, multivariate statistical analysis, biostatistics and statistical signal processing and image analysis. He has written nearly two hundred scholarly articles in these fields, including two recent text-books *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks* published by Springer in 2010 and *Robust Correlation: Theory and Applications*, which is published by Wiley in 2016 and co-authored with Prof. Georgy Shevlyakov (St. Petersburg State Polytechnical University).

3 Summary of research activities of Hannu Oja

To summarize and visualize Hannu's research interest in statistics we used the titles and abstracts of 111 of Hannu's methodological research articles and obtained the word cloud shown in Figure 1.

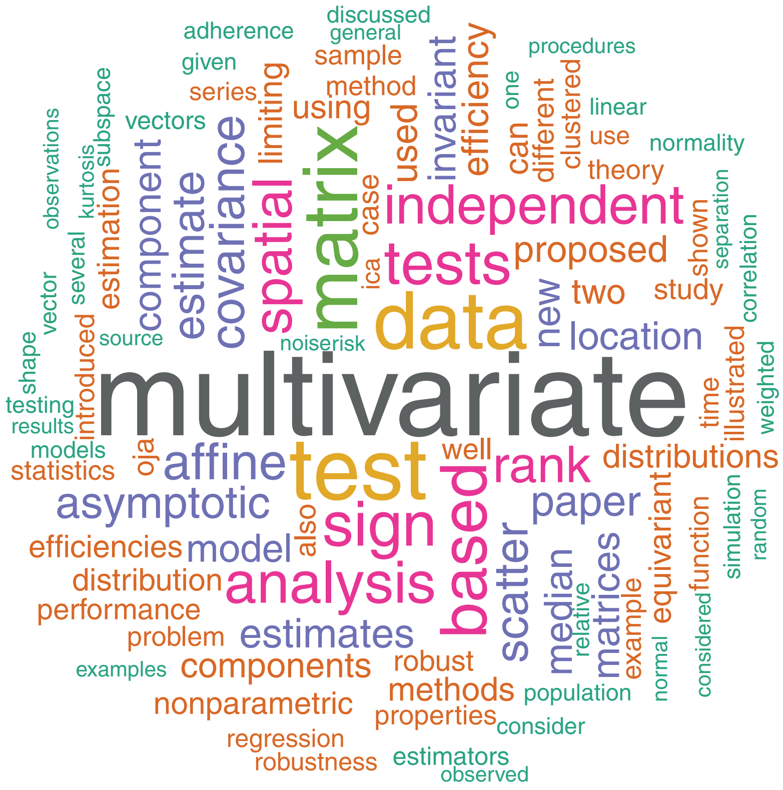

Figure 1: Word cloud based on 111 titles and abstracts of Hannu Oja's publications.

The figure clearly illustrates that Hannu's research focused mainly on diverse multivariate and nonparametric methods. His main research areas are now addressed in a bit more detail.

3.1 Nonparametric and descriptive statistics

Hannu Oja has been a pioneer in the field of multivariate nonparametric methods. He has written dozens of scholarly articles on a wide variety of topics including descriptive statistics, multivariate signs and ranks and related nonparametric inference and multivariate analysis as well as extensions of these concepts to cluster correlated data. Many of his contributions have been influential which is attested in the large number of citations of his papers, many of which are published in the leading journals of statistics including *Annals of Statistics*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society Ser. A-C*, *Statistical Science* and *Biometrika*. His main collaborators in this field of research have been Prof. emeritus Thomas P. Hettmansperger (Pennsylvania State University, USA), Prof. emeritus Jukka Nyblom (University of Jyväskylä), Prof. emeritus Marc Hallin and Prof. Davy Paindaveine (Universite Libre de Bruxelles, Belgium), Prof. Christophe Croux (KU Leuven, Belgium), Prof. Probal Chaudhuri (Indian Statistical Institute, India), Dr. Gleb A. Koshevoy (Russian Academy of Sciences, Russia), Prof. Yuri Tyurin (Moscow State University, Russia), Prof. Denis Larocque (HEC Montréal, Canada), Prof. David Tyler (Rutgers University, USA), Prof. Emeritus Ronald Randles and Prof. Somnath Datta (University of Gainesville, USA), Prof. Robert Serfling (University of Texas, USA) and his former students, Dr. Ahti Niinimaa, Dr. Jyrki Möttönen, Dr. Samuli Visuri, Assoc. Prof. Esa Ollila, Dr. Sara Taskinen, Prof. Jaakko Nevalainen, Dr. Seija Sirkiä and Dr. Klaus Nordhausen.

Hannu's early papers [1, 2, 4, 3], developed during his Ph.D. studies at the University of Oulu under the supervision of Prof. Emeritus Elja Arjas, considered descriptive statistics (e.g., location, scale, skewness, kurtosis), goodness-of-fit tests and tests of normality. It is notable that these papers were single-authored and the promise of a successful career can be foreseen in papers [1, 3], which have been influential in the field, collecting more than 600 citations up to date (source: Google scholar, June 1st, 2016). In [1], he developed a theoretical framework based on partial ordering for better understanding what the commonly used descriptive statistics such as skewness and kurtosis really measure. This paper was followed by a seminal paper [3], where multivariate extensions of these concepts were considered, including his now celebrated affine invariant multivariate median, commonly known as *Oja median*. The Oja median is based on an intuitive geometric description of the most central location of a point cloud. In the bivariate case, it can be defined as a point $\hat{\boldsymbol{\theta}} \in \mathbb{R}^2$ which minimizes the scatter of the data defined as the sum of areas of triangles formed by $\boldsymbol{\theta}$ and pairs of observations $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbb{R}^2$, $i < j$,

$$\sum_{i<j} \mathrm{abs}\Big\{ \begin{vmatrix} 1 & 1 & 1 \\ \mathbf{x}_i & \mathbf{x}_j & \boldsymbol{\theta} \end{vmatrix} \Big\}. \tag{1}$$

Note that an extension to a higher dimensional setting ($\mathbb{R}^d$, $d \geq 3$) is straightforward, now defining the scatter of the data as the sum of volumes of simplices formed by a point $\boldsymbol{\theta} \in \mathbb{R}^d$ and a set of d observations. This seminal paper can be seen to be influential in increasing interest on the concept of statistical depth of Tukey and which has now emerged as a new advanced area in this field consisting of novel concepts such as regression depth or multivariate quantiles. Hannu Oja and his co-workers later derived the properties (robustness, asymptotic normality, computation) of the Oja median in [6, 9, 16, 31, 72, 76, 155] and considered its applications in image restoration in [36]. In [14, 15] based on the criterion (1), the concept of affine invariant multivariate signs and ranks, now bearing the name *Oja sign and ranks*, were introduced as natural extensions of the univariate sign and rank. These works were followed by a series of papers [21, 22, 40, 48, 51, 53, 82, 111] in which nonparametric multivariate one-sample and several sample sign, rank and signed-rank tests were developed. These developments, launched from his 1983 seminal paper [3] then culminated in an invited survey article [56], “Affine invariant multivariate sign and rank tests and corresponding estimates: a review”, published in *Scandinavian Journal of Statistics* in 1999 and to an invited survey article [87], ”Multivariate nonparametric tests”, co-authored with Prof. Ronald Randles and published in *Statistical Science* in 2004.

Also the spatial median and the related multivariate spatial sign and rank concepts have been fundamental components in Hannu Oja’s research. The spatial median is defined as point $\boldsymbol{\theta}$ minimizing

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\theta}\|.$$

In other words, a point for which the sum of Euclidean distances to the data points is at its minimum. Statistical properties of the spatial median were investigated in [31, 128] while [47] introduced an affine equivariant version of the spatial median. The spatial sign function, defined as a gradient vector of the underlying loss function $\rho(\mathbf{x}) = \|\mathbf{x}\|$, is simply $\mathbf{S}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ and the spatial rank function is defined as $\mathbf{R}(\mathbf{x}) = \mathrm{ave}_i\{\mathbf{S}(\mathbf{x} - \mathbf{x}_i)\}$. In a series of papers [14, 15, 29, 40, 43, 86], Hannu Oja and his co-workers developed sign and rank tests based on these concepts and derived their properties such as Pitmann efficiencies. In [57, 63, 61] covariance (scatter) matrices based on spatial and Oja signs and ranks were introduced. Their statistical properties were studied in [78, 88, 70] whereas affine equivariant symmetrised versions of the spatial sign and rank covariance matrices were studied in [107, 120, 132].

Using the developed toolbox of multivariate sign and ranks and related covariance matrices, Hannu Oja and his co-workers developed several multivariate analysis procedures [61, 70, 83, 95, 96, 98, 100, 108, 112, 114, 116, 123, 152], multivariate regression estimators [71, 79, 173], tests of independence [81, 80, 89, 92, 172], tests of normality [84, 101], estimates of direction-of-arrivals of plane waves in sensor array signal processing applications [62, 63, 64, 68, 69] as well as advanced analysis methods of cluster-correlated data [104, 102, 103, 109, 117, 127, 129, 144, 145, 147, 159]. These developments then culminated in a text-book [175], “Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks”, published by Springer in 2010.

3.2 Scatter matrices and dimension reduction

Another major research interest of Hannu Oja has been the investigation of different scatter matrices and their roles in multivariate statistics, especially in the context of dimension reduction. His main collaborators in this research area include Prof. David Tyler (Rutgers University, USA), Prof. Frank Critchley (Open University, UK), Professor Fabian Theis (Helmholtz Zentrum München, Germany), Professor Christophe Croux (KU Leuven, Belgium), Prof. Bing Li (Pennsylvania State University, USA), Prof. Robert Serfling (University of Texas, USA), Prof. Visa Koivunen (Aalto University), Dr. Jari Miettinen (University of Jyväskylä) and his former and current students, Dr. Samuli Visuri, A.Prof. Esa Ollila, Dr. Sara Taskinen, Dr. Seija Sirkiä, Dr. Klaus Nordhausen, A.Prof. Pauliina Ilmonen, M.Sc. Markus Matilainen and M.Sc. Joni Virta.

A central concept in this research line has been a *scatter matrix* functional defined as follows. Let $\mathbf{x}$ be p-variate random vector with c.d.f. F, Then a matrix-valued functional $\mathbf{S}(F)$, also denoted as $\mathbf{S}(\mathbf{x})$, is a scatter matrix if it is affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{x}+\mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^T,$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all p-variate vectors $\mathbf{b}$. An example of a scatter matrix functional is the covariance matrix, the corresponding estimator (final sample counterpart) being the sample covariance matrix. Although the sample covariance matrix is an optimal estimator for an i.i.d. sample from a multivariate normal model, it is known to be very inefficient under heavy-tailed non-Gaussian distributions and highly sensitive to outliers (i.e., non-robust). Robust scatter and shape matrices based on sign and rank concepts were therefore developed in [61, 78, 88, 94, 132], and in [74] scatter matrices based on zonotopes were studied.

An important property of a scatter matrix is that it is proportional to the covariance matrix when F belongs to the class of elliptically symmetric distributions. This means that many multivariate analysis procedures, which require an estimate of the covariance matrix up to a scale, can also be based on a scatter matrix estimate. For example, robust principal component analysis (PCA) can be based on an robust scatter matrix estimate by first computing its eigen-value decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ (where $\mathbf{\Lambda}$ denotes the diagonal matrix of ordered eigenvalues and $\mathbf{U}$ the matrix of eigenvectors) and then obtaining the principal component (PC) variables in a standard way as $\mathbf{z} = \mathbf{U}^T\mathbf{x}$. Robust PCA approaches were developed in [61, 70, 152], for example.

Another important contribution of Hannu Oja is developing a general framework of using two (or several) scatter matrix functionals for independent component analysis (ICA) or linear dimension reduction. The seminal paper [122], a discussion paper read at the Royal Statistical Society, introduces the so-called *invariant coordinate selection (ICS)* which finds a transformation matrix $\mathbf{B}$ that jointly diagonalizes two scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$, i.e.,

$$\mathbf{B}\mathbf{S}_1(\mathbf{x})\mathbf{B}^T = \mathbf{I}_p \quad \text{and} \quad \mathbf{B}\mathbf{S}_2(\mathbf{x})\mathbf{B}^T = \mathbf{D},$$

where $\mathbf{D}$ denotes a diagonal matrix with decreasing diagonal elements. Then the invariant coordinates $\mathbf{z} = \mathbf{B}\mathbf{x}$ of the data $\mathbf{x}$ can differ only by their signs no matter what is the used co-ordinate system (basis) for representing $\mathbf{x}$. ICS has been used in transforming non-affine equivariant methods to be affine equivariant [97, 113], as a method for model selection [141] or finding clusters in the data [98], and as a Fisher's linear discrimination rule [122]. Theoretical properties of ICS were studied in [123, 146]. In supervised dimension reduction the dimension of $\mathbf{x}$ should be reduced in such a way that no information is lost about the dependence of $\mathbf{x}$ with a response $\mathbf{y}$. In [157], supervised scatter matrix functional is introduced as a matrix-valued functional verifying the following equivariance principle

$$\mathbf{S}(\mathbf{A}\mathbf{x}+\mathbf{b};\mathbf{y}) = \mathbf{A}\mathbf{S}(\mathbf{x};\mathbf{y})\mathbf{A}^T$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all p-variate vectors $\mathbf{b}$. The method of using a regular scatter functional with a supervised scatter functional in ICS is then called as supervised ICS (SICS) [157]. SICS includes many well-known supervised dimension methods like SIR, SAVE or pHd for appropriate choices of $\mathbf{S}_1$ and $\mathbf{S}_2$.

Another application of ICS is independent component analysis (ICA) where the observed p-variate random vector $\mathbf{x}$ is modelled as an unknown linear mixture of unobserved (latent) statistically independent source random variables $z_1, \ldots, z_p$ (also called as independent components), that is,

$$\mathbf{x} = \mathbf{A}\mathbf{z}.$$

The goal is then to recover $\mathbf{z}$ or equivalently to estimate the unknown full rank mixing matrix $\mathbf{A}$ given a data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. In [98, 108, 113, 114, 122], Hannu Oja and his co-workers showed that the ICS transformation $\mathbf{B}$ based on two scatter matrices with independence property can also recover the independent components. Most scatter matrices do not possess independence property and therefore [107] proposed symmetrized M-estimators of scatter which possess this desirable property. The above idea was later extended to independent subspace analysis (ISA) model in [138, 140]. In [148, 158, 163, 166, 165, 171] and [168], the statistical properties of ICA estimators were studied in time series and spatial data contexts, respectively. For other studies of Hannu Oja in the field of ICA, see [124, 142, 151, 154, 174]. Hannu's most recent interests in this area are to develop ICA models and methods for financial times series and for functional data as well as for tensor-valued observations.

3.3 Biometry

In addition to nonparametric statistics and multivariate methods Hannu Oja has contributed broadly in the field of biometry. He has coauthored nearly 50 research articles with more than 100 researchers from Finnish and foreign universities. His main collaborators in this field include deceased Prof. Emerita Paula Rantakallio, Prof. Olli-Pekka Alho, Prof. Esa Läärä, Prof. Matti Isohanni, Prof. Martti Sorri (University of Oulu) and Professor Seija Sandberg (University College London). In the following we will review few of Hannu Oja's achievements in biostatistics.

The Northern Finland Birth Cohort 1966 (NFBC1966) study, started by deceased Professor Emerita Paula Rantakallio, is world famous. It covers medical records of more than 12000 mothers and their children and has allowed researchers to study how perinatal events affect subsequent morbidity and mortality of children. Working as a senior lecturer and an acting professor at the Department of Mathematical Sciences, University of Oulu, Hannu Oja participated in several research projects concerning the NFBC1966 study. In [12, 13] birth weight data and intrauterine weight data was modeled using mixture models. The risk factors involved in pre-term birth were studied in [11, 17, 24], the risk factors for teenage smoking and alcohol use are reported in [19, 23] and for adult schizophrenia and other psychoses in [66].

Another large cohort study conducted at the University of Oulu by Professor Olli-Pekka Alho aimed at finding risk factors for acute otitis media. In this study, infection data as well as background information of more than 2500 children up to age 2 years was gathered. The challenge in the modeling was in the dependency of the response variables, therefore a new dynamic logistic regression model was constructed in which the child's observation time was analyzed in short periods by recording the values of the determinants for each period separately. Such dynamic modeling allows one to control both the confounding effects and time dependent responses, and has proved to outperform conventional approaches when analyzing longitudinal data. The results of studies concerning acute otitis media data are reported in [20, 26, 32, 34, 33, 38] among others.

A similar dynamic modelling approach as in [32] was used in Professor Oja's most recognised biometrics paper that was written in co-operation with Professor Seija Sandberg and her research group. The group had followed 90 Scottish child patients suffering from chronic asthma up to 18 months. Children's asthma was monitored using diaries and daily peak-flow measurements. In addition to this, repeated interview assessments of life events and long-term psychosocial experiences were carried. The dynamic logistic regression model was constructed as follows: For child i, denote the observed follow-up time by $t(i)$ (in 2-week periods). Let then $d(i,j) = 0$ indicate that the peak flow data are missing for child i in period j. Otherwise $d(i,j) = 1$. The response variable $\mathbf{y}(i) = (y(i,1),\ldots,y(i,t(i)))'$ is a vector of $t(i)$ dependent binary outcomes, each $y(i,j)$ indicating whether or not an episode occurred for child i in period j. Let $\mathbf{x}(i,1),\ldots,\mathbf{x}(i,t(i))$ be p-vectors, each $\mathbf{x}(i,j) = (x(i,j,1),\ldots,x(i,j,p))'$ containing the values of p covariates (fixed and time-dependent determinants: age, gender, social class, baseline asthma severity, parental smoking, chronic stress, life events, season) for child i in period j as explained above. As the analyses involved repeated events, $\mathbf{x}(i,j)$ also includes relevant indicators describing the episode history $y(i,1),\ldots,y(i,j-1)$. According to the dynamic logistic model the conditional probability $r(i,j)$ for child i of encountering an episode in period j conditional on his or her history $\mathbf{x}(i,j)$ is

$$r(i,j) = \frac{1}{1+\exp\{-\beta'\mathbf{x}(i,j)\}},$$

and the total log-likelihood (for constructing estimates, confidence intervals and statistical tests) using these conditional probabilities is then

$$l(b) = \sum_i \sum_j d(i,j)(y(i,j)\,\log(r(i,j)) + (1-y(i,j))\,\log(1-r(i,j))).$$

Notice that the sum is over the cases $d(i,j) = 1$ only (periods with no missing peak flow data). As can be seen from the formula, the model can then be fitted using standard logistic regression algorithms for independent observations, if the design matrix is constructed as explained above, $\mathbf{x}(i,j)$ explicitly depending on the episode history $y(i,1),\ldots,y(i,j-1)$. The main founding of this study was that severe life events, both on their own and in conjunction with high chronic stress, significantly increased the risk of new asthma attacks. The effect of severe events without accompanying chronic stress involved a small delay. The results of this study were published in [60], “The role of acute and chronic stress in asthma attacks in children”, in the highly-ranked journal *Lancet* having nearly 400 citations up to date (source: Google scholar, June 1st, 2016). The role of positive life events on asthma attacks is studied in a similar fashion in [67].

In 2005-12, while acting as a Professor of Biometry and as a Academy Professor at the University of Tampere, Hannu Oja participated in several biometrical research projects. See [99, 110, 115, 119, 121] for example. In addition to these projects, Professor Oja has participated in supervising three doctoral theses in biometry. Dr. Maarit Laaksonen’s thesis concerned new estimation methods for population attributable fraction (PAF) in longitudinal studies. The result of applying new estimating estimation methods for Mini-Finland Health Survey are reported in [126, 134]. In Dr. Silvia Kiwuwa-Muyingo’s thesis, new tools to analyze adherence data were developed. Methods were applied to antiretroviral therapy data collected in Uganda and Zimbabwe and revealed new evidence between adherence and mortality in HIV-infected adults [133, 153]. Dipl.-Stat. Daniel Fischer (LUKE) develops in his thesis work novel statistical methods for the analysis of gene expression data [156, 161, 162].

4 Book and conference

On December 2015 several Hannu Oja’s closest colleagues and friends gathered together at the ERCIM2015 conference in London to celebrate Hannu’s 65th birthday and looming retirement. In the conference a session *“Special session on modern multivariate and robust methods in honor of H. Oja’s 65th birthday”* was organized, and invited talks were given by Prof. Robert Serfling (University of Texas), Prof. Emeritus Marc Hallin (Universite Libre de Bruxelles) and Prof. Bing Li (Pennsylvania State University). On the same occasion, a Festchrift *“Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja”* was published by Springer with Klaus Nordhausen and Sara Taskinen as editors. The book includes 25 original research papers written by Hannu’s colleagues and coauthors from the past 45 years. The topics vary from univariate nonparametric and robust methods to signals processing applications. The research papers are joined with recollections of Hannu Oja’s early years of academic life written by Prof. Emeritus Elja Arjas and an overview of Hannu Oja’s publication and coauthorship networks written by Dipl.-Stat. Daniel Fischer and the editors of the book.

5 Conclusions

Hannu Oja’s research activities are covering an exceptionally broad spectrum of fundamental theoretical and applied statistical topics, mainly in the areas of non-parametric and robust multivariate methods, biometry and signal processing. This is seen in his publication list. The full list of papers published so far is given in the end of the article. We hope that we have been able to outline his outstanding contributions to statistical sciences in this essay. Besides his ground-breaking research work that still continues after his retirement, his contributions to supervising Ph.D. students and education of future generation of statisticians has been unique in Finland.

In total Hannu has supervised in several disciplines the following eleven PhD students:

1. Ahti Niinimaa, Statistics, University of Oulu, 1993.
2. Jyrki Möttönen, Statistics, University of Oulu, 1997.
3. Samuli Visuri, Statistical Signal Processing, Helsinki University of Technology, 2001 (co-supervisor Visa Koivunen).
4. Esa Ollila, Statistics, University of Jyväskylä, 2002.
5. Sara Taskinen, Statistics, University of Jyväskylä, 2003 (co-supervisor Annaliisa Kankainen).
6. Jaakko Nevalainen, Statistics, University of Tampere, 2007.
7. Seija Sirkiä, Statistics, University of Jyväskylä, 2007, (co-supervisor Sara Taskinen).
8. Klaus Nordhausen, Biometry, University of Tampere, 2008 (co-supervisor Tapio Nummi).
9. Maarit Laaksonen, Biometry, University of Tampere, 2010 (co-supervisors Paul Knekt and Tommi Härkänen).
10. Pauliina Ilmonen, Biometry, University of Tampere, 2011 (co-supervisor Jaakko Nevalainen).
11. Sylvia Kiwuwa-Muyingo, Biometry, University of Tampere, 2012 (co-supervisor Arto Palmu).

And at least three more students are supposed to obtain their PhD within the next two years.

References

[1] H. Oja. On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 8:154–168, 1981.

[2] H. Oja. Two location and scale-free goodness-of-fit tests. *Biometrika*, 68:637–640, 1981.

[3] H. Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1:327–332, 1983.

[4] H. Oja. New tests for normality. *Biometrika*, 70:297–299, 1983.

[5] H. Oja. Partial ordering of distributions. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Science*, pages 490–494. Wiley, New York, 1985.

[6] H. Oja and A. Niinimaa. Asymptotical properties of the generalized median in the case of multivariate normality. *Journal of the Royal Statistical Society, Series B*, 47:372–377, 1985.

[7] H. Oja. On permutation tests in multiple regression and analysis of covariance problems. *Australian Journal of Statistics*, 29:81–100, 1987.

[8] H. Oja and J. Nyblom. On bivariate sign tests. *Journal of the American Statistical Association*, 84:249–259, 1989.

[9] A. Niinimaa, H. Oja, and M. Tableman. The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. *Statistics & Probability Letters*, 10:325–328, 1990.

[10] P. Rantakallio, A.-L. Hartikainen-Sorri, P. Sipilä, H. Oja, U. ElSaid, and M. Koiranen. Computer-based perinatal risk prediction in a geographically defined parturient population - an intervention study. *Pediatric Research*, 28:305, 1990.

[11] P. Rantakallio and H. Oja. Perinatal risk for infants of unmarried mothers during 20 years. *Early Human Development*, 22:157–169, 1990.

[12] H. Oja, M. Koiranen, and P. Rantakallio. Fitting mixture models to birth weight data: a case study. *Biometrics*, 47:883–897, 1991.

[13] P. Rantakallio, H. Oja, and M. Koiranen. Has intrauterine weight gain curve changed in shape? *Pediatric & Perinatal Epidemiology*, 5:201–220, 1991.

[14] B. M. Brown, T. P. Hettmansperger, J. Nyblom, and H. Oja. On certain bivariate sign tests and medians. *Journal of the American Statistical Association*, 87:127–135, 1992.

[15] T.P. Hettmansperger, J. Nyblom, and H. Oja. On multivariate notions of sign and rank. In Y. Dodge, editor, *L_1-Statistical Analysis and Related Methods*, pages 267–278. Elsevier, Amsterdam: North Holland, 1992.

[16] A. Niinimaa, H. Oja, and J. Nyblom. Algorithm AS 277: The Oja bivariate median. *Applied Statistics*, 41:611–617, 1992.

[17] P. Sipilä, A.-L. Hartikainen-Sorri, H. Oja, and L. von Wendt. Perinatal outcome of pregnancies complicated by vaginal bleeding. *British Journal of Obstetrics and Gynaecology*, 99:959–963, 1992.

[18] O. P. Alho, O. Kilkku, H. Oja, M. Koivu, and M. Sorri. Control of the temporal aspect when considering risk factors for acute otitis media. *Archives of Otolaryngology - Head & Neck Surgery*, 119:444–449, 1993.

[19] M. Isohanni, H. Oja, P. Rantakallio, I. Moilanen, and M. Koiranen. The relation of the teenage smoking and alcohol use, with special reference to the deficient family background. *Scandinavian Journal of Social Medicine*, 21:24–30, 1993.

[20] O. P. Alho, M. Koivu, M. Sorri, H. Oja, and O. Kilkku. Which children are being operated on for recurrent acute otitis media? *Archives of Otolaryngology - Head & Neck Surgery*, 120:807–811, 1994.

[21] T.P. Hettmansperger, J. Nyblom, and H. Oja. Affine invariant multivariate one-sample sign test. *Journal of the Royal Statistical Society, Series B*, 56:221–234, 1994.

[22] T.P. Hettmansperger and H. Oja. Affine invariant multivariate multisample sign tests. *Journal of the Royal Statistical Society, Series B*, 56:235–249, 1994.

[23] M. Isohanni, H. Oja, I. Moilanen, and M. Koiranen. Teenage alcohol drinking and nonstandard family background. *Social Science & Medicine*, 38:1565–1574, 1994.

[24] P. Sipilä, A.-L. Hartikainen, L. von Wendt, and H. Oja. Changes in risk factors for unfavourable pregnancy outcome among singletons during twenty years. *Acta Obstetricia et Gynecologica Scandinavica*, 73:612–618, 1994.

[25] O. Alho, K. Jokinen, T. Pirilä, A. Ilo, and H. Oja. Acute epiglottitis and infant conjugate haemophilus influenzae type b vaccination in northern Finland. *Archives of Otolaryngology - Head & Neck Surgery*, 121:898–902, 1995.

[26] O.P. Alho, H. Oja, M. Koivu, and M. Sorri. Chronic otitis media with effusion in infancy - how frequent is it and how does it develop? *Archives of Otolaryngology-Head & Neck Surgery*, 121:432–436, 1995.

[27] O.P. Alho, H. Oja, M. Koivu, and M. Sorri. Risk factors for chronic otitis media with effusion in infancy. each acute otitis media episode induces a high but transient risk. *Archives of Otolaryngology - Head & Neck Surgery*, 121(8):839–843, 1995.

[28] K. Isohanni, H. Oja, I. Moilanen, M. Koiranen, and P. Rantakallio. Smoking or quitting during pregnancy: associations with background and future social factors. *Scandinavian Journal of Social Medicine*, 23:32–38, 1995.

[29] J. Möttönen and H. Oja. Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5:201–213, 1995.

[30] J. Möttönen, H. Oja, U. Krause, and P. Rantakallio. Application of random coefficient regression model to myopia data: a case study. *Biometrical Journal*, 37:657–672, 1995.

[31] A. Niinimaa and H. Oja. On the influence functions of certain bivariate medians. *Journal of the Royal Statistical Society, Series B*, 57:565–574, 1995.

[32] O. P. Alho, E. Läärä, and H. Oja. How should relative risk estimates for acute otitis media in children aged less than 2 years be perceived? *Journal of Clinical Epidemiology*, 49:9–14, 1996.

[33] O.P. Alho, E. Läärä, and H. Oja. Public health impact of various risk factors for acute otitis media in northern Finland. *American Journal of Epidemiology*, 143:1149–1156, 1996.

[34] O.P. Alho, E. Läärä, and H. Oja. What is the natural history of recurrent acute otitis media in infancy? *The Journal of Family Practice*, 43:258–264, 1996.

[35] J. Kiuttu, P. Larivaara, E. Väisänen, S. Keinänen-Kiukaanniemi, and H. Oja. The effects of family systems medicine training on the practice orientations of general practitioners. *Families, Systems & Health*, 14:453–462, 1996.

[36] S. Luukkonen, V. Koivunen, and H. Oja. Multivariate median filter for correlated noise. In *Proceedings of IEEE NORSIG'96*, pages 411–414, 1996.

[37] H. Oja, O.P. Alho, and E. Läärä. Model-based estimation of the excess fraction (attributable fraction): day care and middle ear infection. *Statistics in Medicine*, 15:1519–1534, 1996.

[38] M. Sorri, O.P. Alho, and H. Oja. Dynamic multivariate modelling: day care and consultation rate for acute otitis media. *Acta Otolaryngologica*, 116:299–301, 1996.

[39] B.M. Brown, T.P. Hettmansperger, J. Möttönen, and H. Oja. Rank plots in the affine invariant case. In Yadolah Dodge, editor, *L_1-statistical procedures and related topics: Papers from the 3rd International Conference on L_1-Norm and Related Methods held in Neuchtel, August 11–15, 1997*, pages 351–362, 1997.

[40] T.P. Hettmansperger, J. Möttönen, and H. Oja. Affine invariant multivariate one-sample signed-rank tests. *Journal of the American Statistical Association*, 92:1591–1600, 1997.

[41] J. Kiultu, E. Väisänen, P. Larivaara, S. Keinänen-Kiukaanniemi, and H. Oja. Family systems medicine training: Helping to meet psychiatric and psychosomatic problems. *Nordic Journal of Psychiatry*, 51:259–265, 1997.

[42] T. Mäkikyrö, M. Isohanni, J. Moring, H. Oja, H. Hakko, P. Jones, and P. Rantakallio. Is a child's risk of early onset schizophrenia increased in the highest social class? *Schizophrenia Research*, 23:245–252, 1997.

[43] J. Möttönen, H. Oja, and J. Tienari. On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics*, 25:542–552, 1997.

[44] T. Pirilä, A. Talvisara, O.P. Alho, and H. Oja. Physiological fluctuations in nasal resistance may interfere with nasal monitoring in the nasal provocation test. *Acta Oto-Laryngologica*, 117:596–600, 1997.

[45] M. Sorri, E. Mäki-Torkko, M.-R. Järvelin, and H. Oja. Prevalence figures for mild hearing impairments cannot be based on clinical data. *Acta Oto-laryngologica*, 117:179–181, 1997.

[46] K.-E. Wahlberg, L.C. Wynne, H. Oja, P. Keskitalo, L. Pykäläinen, I. Lahti, J. Moring, M. Naarala, A. Sorri, M. Seitamaa, K. Läksy, J. Kolassa, and P. Tienari. Gene-environment interaction in vulnerability to schizophrenia: Findings from the Finnish adoptive family study of schizophrenia. *American Journal of Psychiatry*, 154:355–362, 1997.

[47] B. Chakraborty, P. Chaudhuri, and H. Oja. Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8:767–784, 1998.

[48] T.P. Hettmansperger, J. Möttönen, and H. Oja. Affine invariant multivariate rank tests for several samples. *Statistica Sinica*, 8:785–800, 1998.

[49] V. Koivunen, S. Luukkonen, and H. Oja. Affine equivariance in multichannel OS-filtering. In *IEEE ICASSP'98*, volume V, pages 2881–2884, 1998.

[50] E.M. Mäki-Torkko, M.-R. Järvelin, M.J. Sorri, A.A. Muhli, and H. Oja. Aetiology and risk indicators of hearing impairments in a one-year birth cohort for 1985-86 in northern Finland. *Scandinavian Audiology*, 27:237–247, 1998.

[51] J. Möttönen, T.P. Hettmansperger, H. Oja, and J. Tienari. On the efficiency of affine invariant multivariate rank tests. *Journal of Multivariate Analysis*, 66:118–132, 1998.

[52] T. P. Hettmansperger, J. Möttönen, and H. Oja. The geometry of the affine invariant multivariate sign and rank methods. *Journal of Nonparametric Statistics*, 11:271–285, 1999.

[53] T.P. Hettmansperger, H. Oja, and S. Visuri. Discussion of "Multivariate analysis by data depth: Descriptive statistics, graphics and inference" by Liu, Parelius and Singh. *Annals of Statistics*, 3:845–853, 1999.

[54] J. Möttönen, V. Koivunen, and H. Oja. Robust autocovariance estimation based on sign and rank correlation coefficients. In *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 187–190, 1999.

[55] A. Niinimaa and H. Oja. Multivariate median. In S. Kotz, C. B. Read, and D. Banks, editors, *Encyclopedia of Statistical Sciences*, volume 3, pages 497–505, New York, USA, 1999. John Wiley & Sons.

[56] H. Oja. Affine invariant multivariate sign and rank tests and corresponding estimates: a review. *Scandinavian Journal of Statistics*, 26:319–343, 1999.

[57] S. Visuri, V. Koivunen, J. Möttönen, and H. Oja. Matrix perturbations in covariance and autocovariance matrix estimators. In *Proceedings of the 1999 Finnish Signal Processing Symposium (FINSIG'99)*, pages 142–146, 1999.

[58] S. Visuri, H. Oja, and V. Koivunen. Multichannel signal processing using spatial rank covariance matrices. In *Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99)*, pages 75–79, 1999.

[59] S. Sandberg, J. Y. Paton, S. Ahola, D. C. McCann, D. McGuinness, C. R. Hillary, and H. Oja. Stressi lisää lasten astmakohtausten riskiä. *Duodemic*, 116:2305–2306, 2000.

[60] S. Sandberg, J.Y. Paton, S. Ahola, D. C. McCann, D. McGuinness, C. R. Hillary, and H. Oja. The role of acute and chronic stress in asthma attacks in children. *Lancet*, 356:982–988, 2000.

[61] S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91:557–575, 2000.

[62] S. Visuri, H. Oja, and V. Koivunen. Nonparametric method for subspace based frequency estimation. In *EUSIPCO-2000*, pages 1261–1264, 2000.

[63] S. Visuri, H. Oja, and V. Koivunen. Nonparametric statistics for DOA estimation in the presence of multipath. In *Proceedings of the First IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2000)*, pages 356–360, 2000.

[64] S. Visuri, H. Oja, and V. Koivunen. Robust subspace DOA estimation for wireless communications. In *The IEEE Annual Vehicular Technology Conference VTC2000*, pages 2551–2555, 2000.

[65] K.E. Wahlberg, L.C. Wynne, H. Oja, P. Keskitalo, H. Anais-Tanner, P. Koistinen, T. Tarvainen, H. Hakko, I. Lahti, J. Moring, M. Naarala, A. Sorri, and P. Tienari. Thought disorder index of Finnish adoptees and communication deviance of their adoptive parents. *Psychological Medicine*, 30:127–136, 2000.

[66] M. Isohanni, P.B. Jones, K. Moilanen, P. Rantakallio, J. Veijola, H. Oja, M. Koiranen, J. Jokelainen, T. Croudace, and M. Järvelin. Early developmental milestones in adult schizophrenia and other psychoses. a 31-year follow-up of the northern Finland 1966 birth cohort. *Schizophrenia Research*, 52:1–19, 2001.

[67] S. Sandberg, D. C. McCann, S. Ahola, H. Oja, J. Y. Paton, and D. McGuinness. Positive experiences and the relationship between stress and asthma in children. *Acta Paediatrica*, 91:152–158, 2001.

[68] S. Visuri, H. Oja, and V. Koivunen. Blind channel identification using robust subspace estimation. In *Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 281–284, 2001.

[69] S. Visuri, H. Oja, and V. Koivunen. Subspace-based direction of arrival estimation using nonparametric statistics. *IEEE Transactions on Signal Processing*, 49:2060–2073, 2001.

[70] C. Croux, E. Ollila, and H. Oja. Sign and rank covariance matrices: statistical properties and application to principal component analysis. In Yadolah Dodge, editor, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 257–269. Birkäuser Verlag, Basel, Switzerland, 2002.

[71] E. Ollila, T.P. Hettmansperger, and H. Oja. Estimates of regression coefficients based on sign covariance matrix. *Journal of the Royal Statistical Society, Series B*, 64:447–466, 2002.

[72] T. Ronkainen, H. Oja, and P. Orponen. Computation of the multivariate Oja median. In R. Dutter, P. Filzmoser, U. Gather, and P. J. Rousseeuw, editors, *Developments in Robust Statistics*, pages 344–359, Heidelberg, 2002. Springer.

[73] S. Taskinen, A. Kankainen, and H. Oja. Tests of independence based on sign and rank covariances. In R. Dutter, P. Filzmoser, U. Gather, and P.J. Rousseeuw, editors, *Developments in Robust Statistics.*, pages 387–403, Heidelberg, 2002. Springer.

[74] G.A. Koshevoy, J. Möttönen, and H. Oja. A scatter matrix estimate based on the zonotope. *Annals of Statistics*, 31:1439–1459, 2003.

[75] J. Möttönen, J. Hüsler, and H. Oja. Multivariate nonparametric tests in a randomized complete block design. *Journal of Multivariate Analysis*, 85:106–129, 2003.

[76] M. Nadar, T.P. Hettmansperger, and H. Oja. The asymptotic covariance matrix of the Oja median. *Statistics & Probability Letters*, 64:431–442, 2003.

[77] H. Oja. Multivariate M-estimates of location and shape. In R. Höglund, M. Jäntti, and G. Rosenqvist, editors, *Statistics, Econometrics and Society. Essays in Honor of Leif Nordberg*. Statistics Finland, 2003.

[78] E. Ollila, H. Oja, and C. Croux. The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. *Journal of Multivariate Analysis*, 87:328–355, 2003.

[79] E. Ollila, H. Oja, and V. Koivunen. Estimates of regression coefficients based on lift rank covariance matrix. *Journal of the Americal Statistical Association*, 99:90–98, 2003.

[80] S. Taskinen, A. Kankainen, and H. Oja. Sign test of independence between two random vectors. *Statistics and Probability Letters*, 62:9–21, 2003.

[81] S. Taskinen, A. Kankainen, and H. Oja. Tests of independence based on sign and rank covariances. In Rudolf Dutter, Peter Filzmoser, Ursula Gather, and Peter J. Rousseeuw, editors, *Developments in Robust Statistics*, pages 387–403. Physica-Verlag HD, 2003.

[82] A. Topchii, Y. Tyurin, and H. Oja. Inference based on the affine invariant multivariate Mann-Whitney-Wilcoxon statistic. *Journal of Nonparametric Statistics*, 15:403–419, 2003.

[83] S. Visuri, E. Ollila, V. Koivunen, J. Möttönen, and H. Oja. Affine equivariant multivariate rank methods. *Journal of Statistical Planning and Inference*, 114:161–185, 2003.

[84] A. Kankainen, S. Taskinen, and H. Oja. On Mardia's tests of multinormality. In M. Hubert, G. Pison, A. Stryuf, and S. Van Aelst, editors, *Statistics for Industry and Technology*, pages 152–164. Birkhäuser, Basel, 2004.

[85] G.A. Koshevoy, J. Möttönen, and H. Oja. On the geometry of multivariate L_1 objective functions. *Allgemeines Statistisches Archiv*, 88:137–154, 2004.

[86] J. Möttönen, H. Oja, and R.J. Serfling. Multivariate generalized spatial signed-rank methods. *Journal of Statistical Research*, 39:19–42, 2004.

[87] H. Oja and R.H. Randles. Multivariate nonparametric tests. *Statistical Science*, 19:598–605, 2004.

[88] E. Ollila, C. Croux, and H. Oja. Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. *Statistica Sinica*, 14:297–316, 2004.

[89] S. Taskinen, A. Kankainen, and H. Oja. Rank scores tests of multivariate independence. In M. Hubert, G. Pison, A. Stryuf, and S. Van Aelst, editors, *Statistics for Industry and Technology*, pages 329–341. Birkhäuser, Basel, 2004.

[90] H. Oja. Discussion of "breakdown and groups" by P. L. Davies and U. Gather. *Annals of Statistics*, 33:1000–1004, 2005.

[91] H. Oja and D. Paindaveine. Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323, 2005.

[92] S. Taskinen, H. Oja, and R. Randles. Multivariate nonparametric tests of independence. *Journal of American Statistical Association*, 100:916–925, 2005.

[93] D. Busarova, Y. Tyurin, J. Möttönen, and H. Oja. Multivariate Theil estimator with the corresponding test. *Mathematical Methods of Statistics*, 15:1–19, 2006.

[94] M. Hallin, H. Oja, and D. Paindaveine. Semiparametrically efficient rank-based inference for shape. II. Optimal R-estimation of shape. *Annals of Statistics*, 34:2757–2789, 2006.

[95] A. Hartikainen and H. Oja. On nonparametric discrimination rules. In R. Liu, editor, *Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications.*, AMS DIMACS Book Series, pages 61–70, 2006.

[96] J. Nevalainen and H. Oja. SAS/IML macros for a multivariate analysis of variance based on spatial signs. *Journal of Statistical Software*, 16:1–17, 2006.

[97] K. Nordhausen, H. Oja, and D.E. Tyler. On the efficiency of invariant multivariate sign and rank test. In Erkki P. Liski, Jarkko Isotalo, Jarmo Niemelä, Simo Puntanen, and George P. H. Styan, editors, *Festschrift for Tarmo Pukkila on his 60th Birthday*, pages 217–231. University of Tampere, Tampere, FINLAND, 2006.

[98] H. Oja, S. Sirkiä, and J. Eriksson. Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35:175–189, 2006.

[99] A. Saastamoinen, H. Oja, E. Huupponen, A. Värri, J. Hasan, and S.L. Himanen. Topographic differences in mean computational sleep depth between healthy controls and obstructive sleep apnoea patients. *Journal of Neurosicence Methods*, 157:178–184, 2006.

[100] S. Taskinen, C. Croux, A. Kankainen, E. Ollila, and H. Oja. Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*, 97:359–384, 2006.

[101] A. Kankainen, S. Taskinen, and H. Oja. Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications*, 16:357–379, 2007.

[102] D. Larocque, J. Nevalainen, and H. Oja. A weighted multivariate sign test for cluster correlated data. *Biometrika*, 94:267–283, 2007.

[103] J. Nevalainen, D. Larocque, and H. Oja. On the multivariate spatial median for clustered data. *The Canadian Journal of Statistics*, 35:215–283, 2007.

[104] J. Nevalainen, D. Larocque, and H. Oja. A weighted spatial median for clustered data. *Statistical Methods & Applications*, 15:355–379, 2007.

[105] K. Nordhausen, H. Oja, and E. Ollila. Robust ICA based on two scatter matrices. In S. Aivazian, P. Filzmoser, and Y. Kharin, editors, *Proceedings of the 8th International Conference on Computer Data Analysis and Modeling*, pages 84–91, Minsk, 2007. Minsk Publishing Center BSU.

[106] H. Oja and F. Critchley. Discussion of the paper "a survey on robust statistics" by S. Morgenthaler. *Statistical Methods & Applications*, 15:271–293, 2007.

[107] S. Sirkiä, S. Taskinen, and H. Oja. Symmetrised M-estimators of scatter. *Journal of Multivariate Analysis*, 98:1611–1629, 2007.

[108] S. Taskinen, S. Sirkiä, and H. Oja. Independent component analysis based on symmetrised scatter estimators. *Computational Statistics and Data Analysis*, 51:5103–5111, 2007.

[109] D. Larocque, J. Nevalainen, and H. Oja. One-sample location tests for multilevel data. *Journal of Statistical Planning and Inference*, 138:2469–2482, 2008.

[110] B. Lindroos, K. Mäenpää, T. Ylikomi, H. Oja, R. Suuronen, and S. Miettinen. Characterisation of human dental stem cells and buccal mucosa fibroblasts. *Biochemical and Biophysical Research Commununications*, 368:329–335, 2008.

[111] J. Nevalainen, J. Möttönen, and H. Oja. A spatial rank test and corresponding estimators for several samples. *Statistics & Probability Letters*, 78:661–668, 2008.

[112] K. Nordhausen, H. Oja, and E. Ollila. Robust independent component analysis based on two scatter matrices. *Austrian Journal of Statistics*, 37:91–100, 2008.

[113] K. Nordhausen, H. Oja, and D.E. Tyler. Tools for exploring multivariate data: The package ICS. *Journal of Statistical Software*, 28:1–31, 2008.

[114] E. Ollila, H. Oja, and V. Koivunen. Complex-valued ICA based on a pair of generalized covariance matrices. *Computational Statistics & Data Analysis*, 52:3789–3805, 2008.

[115] L. Uusitalo, J. Nevalainen, S. Niinistö, G. Alfthan, J. Sundvall, T. Korhonen, M. G. Kenward, H. Oja, R. Veijola, O. Simell, J. Ilonen, M. Knip, and S. M. Virtanen. Serum alpha- and gamma-tocopherol concentrations and risk of advanced beta cell autoimmunity in children with HLA-conferred susceptibility to type 1 diabetes mellitus. *Diabetologia*, 51:773–780, 2008.

[116] P. Chaudhuri, A. K. Ghosh, and H. Oja. Classification based on hybridization of parametric and nonparametric classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1153–1164, 2009.

[117] R. Haataja, D. Larocque, J. Nevalainen, and H. Oja. A weighted multivariate signed-rank test for cluster-correlated data. *Journal of Multivariate Analysis*, 100:1107–1119, 2009.

[118] K. Nordhausen, H. Oja, and D. Paindaveine. Signed-rank tests for location in the symmetric independent component model. *Journal of Multivariate Analysis*, 100:821–834, 2009.

[119] H. Rantanen, A.M. Koivisto, R.K. Salokangas, M. Helminen, H. Oja, S. Pirkola, K. Wahlbeck, and M. Joukamaa. Five-year mortality of Finnish schizophrenia patients in the era of deinstitutionalization. *Social Psychiatry and Psychiatric Epidemiology*, 44:135–142, 2009.

[120] S. Sirkiä, S. Taskinen, H. Oja, and D. Tyler. Tests and estimates for shape based on spatial signs and ranks. *Journal of Nonparametric Statistics*, 21:155–176, 2009.

[121] A. Tahvanainen, M. Leskinen, J. Koskela, E. Ilveskoski, K. Nordhausen, H. Oja, M. Kähönen, T. Kööbi, J. Mustonen, and I. Pörsti. Ageing and cardiovascular responses to head-up tilt in healthy subjects. *Atherosclerosis*, 2007:445–451, 2009.

[122] D.E. Tyler, F. Critchley, L. Dümbgen, and H. Oja. Invariant co-ordinate selection. *Journal of the Royal Statistical Society, Series B*, 71:549–592, 2009.

[123] P. Ilmonen, J. Nevalainen, and H. Oja. Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, 80:1844–1853, 2010.

[124] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila. A new performance index for ICA: Properties, computation and asymptotic analysis. In Vincent Vigneron, Vicente Zarzoso, Eric Moreau, Rémi Gribonval, and Emmanuel Vincent, editors, *Latent Variable Analysis and Signal Separation - 9th International Conference, LVA/ICA 2010, St. Malo, France, September 27-30, 2010. Proceedings*, volume 6365 of *Lecture Notes in Computer Science*, pages 229–236. Springer, 2010.

[125] M. Laaksonen, P. Knekt, T. Härkänen, E. Virtala, and H. Oja. Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. *American Journal of Epidemiology*, 171:837–847, 2010.

[126] M.A. Laaksonen, T. Härkänen, P. Knekt, E. Virtala, and H. Oja. Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. *Statistics in Medicine, Special Issue: Papers from the 29th Annual Conference of the International Society for Clinical Biostatistics*, 29:860–874, 2010.

[127] D. Larocque, R. Haataja, J. Nevalainen, and H. Oja. Two sample tests for the nonparametric Behrens-Fisher problem with clustered data. *Journal of Nonparametric Statistics*, 22:755–771, 2010.

[128] J. Möttönen, K. Nordhausen, and H. Oja. Asymptotic theory of the spatial median. In J. Antoch, M. Huskova, and P.K. Sen, editors, *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckova*, pages 182–193, Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2010.

[129] J. Nevalainen, D. Larocque, H. Oja, and I. Pörsti. Nonparametric analysis of multivariate clustered data. *Journal of the Americal Statistical Association*, 105:864–872, 2010.

[130] K. Nordhausen and H. Oja. Three scatter matrices and independent subspace analysis. In S. Aivazian, P. Filzmoser, and Y. Kharin, editors, *Proceedings of the 9th International Conference on Computer Data Analysis and Modeling*, pages 93–100, Mink, 2010. Minsk Publishing Center BSU.

[131] H.E. Rauhala, S.E. Jalava, J. Isotalo, H. Bracken, S. Lehmusvaara, T.L. Tammela, H. Oja, and T. Visakorpi. mir-193b is an epigenetically regulated putative tumor suppressor in prostate cancer. *International Journal of Cancer*, 127:1363–1372, 2010.

[132] S. Taskinen, S. Sirkiä, and H. Oja. k-step shape estimators based on spatial signs and ranks. *Journal of Statistical Planning and Inference*, 140:3376–3388, 2010.

[133] S. Kiwuwa-Muyingo, H. Oja, A.S. Walker, P. Ilmonen, J. Levin, and J. Todd. Clustering based on adherence data. *Epidemiologic Perspectives & Innovations (methodology)*, 8, 2011.

[134] M. Laaksonen, E. Virtala, P. Knekt, H. Oja, and T. Härkänen. SAS macros for calculation of population attributable fraction in a cohort study design. *Journal of Statistical Software*, 43:1–25, 2011.

[135] H. Mattila, M. Schindler, J. Isotalo, T. Ikonen, M. Vihinen, H. Oja, T.L.J Tammela, T. Wahlfors, and J. Schleutker. Nmd and microrna expression profiling of the hpcx1 locus reveal magec1 as a candidate prostate cancer predisposition gene. *BMC Cancer 2011*, 11, 2011.

[136] K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja, and E. Ollila. Deflation-based FastICA reloaded. In *Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011)*, pages 1854–1858, 2011.

[137] K. Nordhausen and H. Oja. Discussion on the paper "the asymptotic efficiency of the spatial median for elliptically symmetric distributions" by Andrew Magyar and David E. Tyler. *Sankhya, Series B*, 73:188–191, 2011.

[138] K. Nordhausen and H. Oja. Independent subspace analysis using three scatter matrices. *Austrian Journal of Statistics*, 40:93–101, 2011.

[139] K. Nordhausen and H. Oja. Multivariate L_1 methods: The package MNM. *Journal of Statistical Software*, 43:1–28, 2011.

[140] K. Nordhausen and H. Oja. Scatter matrices with independent block property and ISA. In *Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011)*, pages 1738–1742, 2011.

[141] K. Nordhausen, H. Oja, and E. Ollila. Multivariate models and the first four moments. In D. R. Hunter, D. St. P. Richards, and J. L. Rosenberger, editors, *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger*, pages 267–287. World Scientific, Singapore, 2011.

[142] K. Nordhausen, E. Ollila, and H. Oja. On the performance indices of ICA and blind source separation. In *Proceedings of IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2011)*, pages 486–490, 2011.

[143] R.K. Salokangas, M. Helminen, A.M. Koivisto, H. Rantanen, H. Oja, S. Pirkola, K. Wahlbeck, and M. Joukamaa. Incidence of hospitalised schizophrenia in Finland since 1980: decreasing and increasing again. *Social Psychiatry and Psychiatric Epidemiology*, 46:343–250, 2011.

[144] K. Tokola, D. Larocque, J. Nevalainen, and H. Oja. Power, sample size and sampling costs for clustered data. *Statistics & Probability Letters*, 81:852–860, 2011.

[145] S. Datta, J. Nevalainen, and H. Oja. A general class of signed-rank tests for clustered data when the cluster size is potentially informative. *Journal of Nonparametric Statistics*, 24:797–808, 2012.

[146] P. Ilmonen, H. Oja, and R. Serfling. On invariant coordinate system (ICS) functionals. *International Statistical Review*, 80:93–110, 2012.

[147] R. Lemponen, D. Larocque, J. Nevalainen, and H. Oja. Weighted rank tests and Hodges-Lehmann estimates for the multivariate two-sample location problem with clustered data. *Journal of Nonparametric Statistics*, 24:977–991, 2012.

[148] J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Statistical properties of a blind source separation estimator for stationary time series. *Statistics & Probability Letters*, 82:1865–1873, 2012.

[149] K. Nordhausen, H.W. Gutch, H. Oja, and F.J. Theis. Joint diagonalization of several scatter matrices for ICA. In Fabian J. Theis, Andrzej Cichocki, Arie Yeredor, and Michael Zibulevsky, editors, *Latent Variable Analysis and Signal Separation - 10th International Conference, LVA/ICA 2012, Tel Aviv, Israel, March 12-15, 2012. Proceedings*, volume 7191 of *Lecture Notes in Computer Science*, pages 172–179. Springer, 2012.

[150] H. Oja. Descriptive statistics for nonparametric models. The impact of some Erich Lehmann's papers. In J. Rojo, editor, *Selected Works of E. L. Lehmann*, pages 451–457. Springer, New York, 2012.

[151] H. Oja and K Nordhausen. Independent component analysis. In A.-H. El-Shaarawi and W. Piegorsch, editors, *Encyclopedia of Environmetrics*, pages 1352–1360. Wiley, Chichester, 2nd edition, 2012.

[152] S. Taskinen, I. Koch, and H. Oja. Robustifying principal component analysis with spatial sign vectors. *Statistics and Probability Letters*, 82:765–774, 2012.

[153] S. Kiwuwa-Muyingo, H. Oja, A.S. Walker, P. Ilmonen, J. Levin, I. Mambule, P. Mugyenyi, J. Todd, and DART Trial team. Dynamic logistic regression model and population attributable fraction to investigate the association between adherence, missed visits and mortality: a study of HIV-infected adults surviving the first year of ART. *BMC Infectious Diseases*, 13, 2013.

[154] J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Fast equivariant JADE. In *Proceedings of 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pages 6153–6157, 2013.

[155] H. Oja. Multivariate median. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures*, pages 3–15. Springer, Berlin, 2013.

[156] D. Fischer, H. Oja, J. Schleutker, P.K. Sen, and T. Wahlfors. Generalized Mann-Whitney type tests for microarray experiments. *Scandinavian Journal of Statistics*, 41:672–692, 2014.

[157] E. Liski, K. Nordhausen, and H. Oja. Supervised invariant coordinate selection. *Statistics: A Journal of Theoretical and Applied Statistics*, 48:711–731, 2014.

[158] J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Deflation-based separation of uncorrelated stationary time series. *Journal of Multivariate Analysis*, 123:214–227, 2014.

[159] J. Nevalainen, S. Datta, and H. Oja. Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers*, 55:71–92, 2014.

[160] K. Tokola, A. Lundell, J. Nevalainen, and H. Oja. Design and cost optimization for hierarchical data. *Statistica Neerlandica*, 68:130–148, 2014.

[161] D. Fischer and H. Oja. Mann-Whitney type tests for microarray experiments: The R package gMWT. *Journal of Statistical Software*, 65:1–19, 2015.

[162] D. Fischer, T. Wahlfors, H. Mattila, H. Oja, T. Tammela, and J. Schleutker. MiRNA profiles in lymphoblastoid cell lines of Finnish prostate cancer families. *Plos ONE*, 10:1–17, 2015.

[163] K. Illner, J. Miettinen, C. Fuchs, S. Taskinen, K. Nordhausen, H. Oja, and F.J. Theis. Model selection using limiting distributions of second-order blind source separation algorithms. *Signal Processing*, 113:95–103, 2015.

[164] P. Ilmonen, K. Nordhausen, H. Oja, and F.J. Theis. An affine equivariant robust second order blind source separation method. In E. Vincent, A. Yeredor, Z. Koldovsky, and P. Tichavsky, editors, *Latent Variable Analysis and Signal Separation*, LNCS 9237, pages 328–335. Springer, 2015.

[165] M. Matilainen, K. Nordhausen, and H. Oja. New independent component analysis tools for time series. *Statistics and Probability Letters*, 105:80–87, 2015.

[166] J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Statistical Science*, 30:372–390, 2015.

[167] K. Nordhausen, H. Oja, P. Filzmoser, and C. Reimann. Blind source separation for spatial compositional data. *Mathematical Geosciences*, 47:753–770, 2015.

[168] K. Nordhausen, H. Oja, and O. Pärssinen. Mixed effects regression splines to model myopia data. *Journal of Biometrics and Biostatistics*, 6:1–9, 2015.

[169] E. Liski, H. Oja K. Nordhausen, and A. Ruiz-Gazen. Combining linear dimension reduction estimates. In Claudio Agostinelli, Ayanendranath Basu, Peter Filzmoser, and Diganta Mukherjee, editors, *Recent Advances in Robust Statistics: Theory and Applications*, pages 131–149. Springer India, Delhi, 2016.

[170] M. Matilainen, J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. ICA and stochastic volatility models. In S. Aivazian, P. Filzmoser, and Y. Kharin, editors, *Proceedings of the XI International Conference on Computer Data Analysis and Modeling*, pages 30–37, Minsk, 2016. Publishing center of BSU.

[171] J. Miettinen, K. Illner, K. Nordhausen, H. Oja, S. Taskinen, and F. Theis. Separation of uncorrelated stationary time series using autocovariance matrices. *Journal of Time Series Analysis*, 37:337–354, 2016.

[172] H. Oja, D. Paindaveine, and S. Taskinen. Parametric and nonparametric tests for multivariate independence in IC models. *Electronic Journal of Statistics*, 10:2372–2419, 2016.

[173] S. Taskinen and H. Oja. Influence functions and efficiencies of k-step Hettmansperger-Randles estimators. In R. Liu and J.W. McKean, editors, *Robust Rank-Based and Nonparametric Methods*, pages 189–207. Springer, Heidelberg, 2016.

[174] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, and J. Virta. The squared symmetric FastICA estimator. *Signal Processing*, 131:402–411, 2017.

[175] H. Oja. *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer, 2010.

[176] G.L. Shevlyakov and H. Oja. *Robust Correlation: Theory and Applications*. Wiley, 2016.

SUOMEN TILASTOSEURA

Tilastopäivät 2015 – Statistical Days 2015

Big Data in Biological and Medical Research

University of Helsinki, August 20-21, 2015

Day 1 – Thursday, August 22, 2015

Biomedicum Helsinki 1, Lecture Room 2, Haartmaninkatu 8

11:00-12:15 Registration

12:15-12:30 ***Opening***
Jyrki Möttönen, President of the Finnish Statistical Society

Invited Session I: Human Genomics

Chair: *Matti Pirinen*, FIMM

12:30-13:00 ***Research infrastructures and Big Data***
Professor Juni Palmgren, University of Helsinki

13:00-13:15 ***Out of disc space and CPU hours: adventures in SISu-projects and current statistical genetics***
Samuli Ripatti, FIMM

13:15-14:00 ***UK Biobank, A Large-scale, Extensively Phenotyped, Prospective Resource: How to Use the Genotype Data for 500,000 Individuals***
Doctor Desislava Petkova, Oxford University

Contributed Session 1

Chair: *Samuli Ripatti*, FIMM

14:00-14:15 ***Efficient fine mapping of thousands of correlated genetic variants using summary data from genome-wide association studies***
Christian Benner, FIMM

14.15-14.30 ***metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis***
Anna Cichonska, FIMM & Aalto University

14:30-15:00 Coffee Break

Invited Session II: Bioinformatics and Molecular Evolution

Chair: *Ida Surakka*, FIMM

15:00-15:45 ***Voodoo or real inference? ABC meets machine learning with applications to evolutionary epidemiology***
Professor Jukka Corander, University of Helsinki

15:45-16:30 ***Modeling the Impact of Recombination on the Genomic Distribution of Streptococcus Pneumoniae***
Doctor Pekka Marttinen, Aalto University

Sponsor's Address and Conference Dinner

16:30-17:00 ***Patient Safety in Focus – Using Text Analytics to Detect Adverse Event Related Triggers***
Pertti Viitamäki, SAS Institute Oy

19:00-22:00 Conference dinner at Ravintola Lasipalatsi

Day 2 – Friday, August 21, 2015

Biomedicum Helsinki 1, Lecture Room 2, Haartmaninkatu 8

09:00-9.30 Morning coffee with posters

Invited Session III: Population registry data

Chair: *Ari Jaakola*, Vice President of the Finnish Statistical Society

09:30-10.00 ***Finnish Cancer Registry – Population-based Big Data Covering Six Decades***
Doctor Janne Pitkäniemi, The Finnish Cancer Registry

10:00-10:30 ***Register data – To be or not to be Big Data?***
Doctor Reijo Sund, University of Helsinki

Invited Session IV:

Chair: *Jyrki Möttönen*, President of the Finnish Statistical Society

10:30-11:15 ***Progress in Human Genetics – with Big Data***
Professor Mark Daly, Harvard University

Final Session

Chair: *Jyrki Möttönen*, President of the Finnish Statistical Society

11:15-11:45 ***Leo Törnqvist Prize***
11:45-12:00 ***Closing***

Efficient fine mapping of thousands of correlated genetic variants using summary data from genome-wide association studies

Christian Benner[1,2], Chris Spencer[3], Samuli Ripatti[1,2,4] and Matti Pirinen[1]

1 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
2 Department of Public Health, University of Helsinki, Helsinki, Finland
3 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
4 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Abstract

To date, Genome-Wide Association Studies (GWAS) have linked thousands of genomic regions to complex diseases and traits. For any associated region, the next challenge is fine mapping, that is, pinpointing individual variants and genes that have a direct effect on the disease or trait. However, large sample sizes are needed for fine mapping and existing methods are not scalable to terabytes of data currently emerging from dense genotype imputation or next-generation sequencing.

We introduce FINEMAP, a tool for Bayesian fine mapping genomic regions scalable to big data. By collapsing individual-level data to summary statistics, our method using Bayesian linear regression enables fine mapping of associated regions from studies with hundreds of thousands of individuals, for instance from large-scale biobank genotyping projects. Importantly, our method is computationally efficient permitting fine mapping of large genomic regions with thousands of variants. We attain high computational efficiency by 1) eliminating nuisance parameters through analytic marginalization of the likelihood function, 2) reducing the complexity of the likelihood evaluation by rewriting the determinant and quadratic form in the Normal likelihood and 3) using a search algorithm to define a set of associated variants that is efficient to run and strongly scalable in parallel computing clusters. Moreover, FINEMAP is able to integrate prior information about external functional annotation of variants to improve its context-specific accuracy.

We compare FINEMAP with existing fine mapping methods including CAVIARBF and PAINTOR and achieve similar or better fine mapping accuracy using orders of magnitude less processing time. For example, FINEMAP was able to identify the correct causal variants in a test case with 1250 SNPs in 1.3 hours while CAVIARBF needed 3.6 hours and PAINTOR did not finish in 2 days. We illustrate FINEMAP on data from the UK Parkinson's Disease Consortium and the Wellcome Trust Case Control Consortium 2 by fine mapping 4q22/SNCA region that contains a complex association pattern with Parkinson's disease.

FINEMAP is a user-friendly, freely available, fine-mapping tool for big data and we believe that it will prove useful in near future in prioritizing variants from large-scale GWAS meta-analyses as well as biobank genotyping projects.

Voodoo or real inference? ABC meets machine learning with applications to evolutionary epidemiology

Professor Jukka Corander
University of Helsinki

Some statistical models are specified via a data generating process for which the likelihood function cannot be computed in closed form. Standard likelihood-based inference is then not feasible but the model parameters can be inferred by finding the values which yield simulated data that resemble the observed data. This approach faces at least two major difficulties: The first difficulty is the choice of the discrepancy measure which is used to judge whether the simulated data resemble the observed data. The second difficulty is the computationally efficient identification of regions in the parameter space where the discrepancy is low. We give here an introduction to our recent work where we tackle the two difficulties through classification and Bayesian optimization.

Modeling the Impact of Recombination on the Genomic Distribution of Streptococcus Pneumoniae

Doctor Pekka Marttinen, Aalto University

Bacterial genomes can be divided into a core of ubiquitous genes and accessory genes that are present in a fraction of bacterial isolates. The ecological significance of this variation in gene content remains unclear. We develop a simulation model combining diversification in both the core and accessory genome, including mutation, homologous recombination and horizontal gene transfer introducing new genes, to produce a population of interacting clusters of strains with varying genome content. The model is fitted using likelihood-free inference, where the distribution of discrepancies between simulated and observed data is modeled using a Gaussian process (GP), and the parameter estimate is obtained by finding the value that minimizes the mean of the GP. The model fits well to a systematic population sample of 616 pneumococcal genomes, capturing the major features of the population structure with parameter values that agree well with empirical estimates.

Reference: Marttinen, P., Croucher, N.J., Gutmann, M.U., Corander, J. and Hanage, W.P. (2015). Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microbial Genomics, 1, doi:10.1099/mgen.0.000038

Suomen syöpärekisteri – iso aineisto vuodesta 1953

Janne Pitkäniemi
Tilastojohtaja, Suomen Syöpärekisteri

Suomen Syöpärekisteri on Suomen Syöpäyhdistys ry:n ylläpitämä syöpätautien tilastollinen ja epidemiologinen tutkimuslaitos. Suomen Syöpäyhdistys ry ylläpitää Terveyden ja hyvinvoinnin laitoksen alaista valtakunnallista syöpätapausten tietokantaa, syöpärekisteriä, jonka osana ovat myös rinta- ja kohdunkaulasyövän joukkotarkastusrekisterit. Syöpärekisteriin kootaan kattavasti kaikki Suomessa diagnosoidut syöpätapaukset tarkoituksena monitoroida syöpätaakkaa Suomessa. Toiminta on jatkunut yhtäjaksoisesti vuodesta 1953 alkaen. Tietovaranto kasvaa kokoajan ja sen tiivistetyssä versiossa tällä hetkellä noin 1,4 miljoonaa tietuetta.

Syöpärekisteri on viime vuosina uudistanut syöpätiedon tilastojen tuottamista ja siihen liittyvää tilastollista tiedonkäsittelyä. Varsinaisesta tiedonhallinnasta on eriytetty rutiinitilastojen tuottaminen sekä tilastollinen tutkimustoiminta. Uudet tilastoratkaisut perustuvat yhtenäiseen kaksi kertaa vuodessa päivitettävään tiivistettyyn syöpätietovarantoon, johon on liitetty versiointihallinta. Näin on sekä lähtöaineiston dokumentointia että jäljitettävyyttä parannettu.

Tilastolaskenta tehdään R-ohjelmalla ja laskennan toteuttamista varten on kehitetty kaksi R-ohjelman pakettia. Ensimmäinen paketti on tarkoitettu syöpärekisterin omaan sisäiseen käyttöön ja toinen on vapaasti CRANissa levityksessä oleva popEpi paketti (https://cran.r-project.org/web/packages/popEpi/index.html). Sisäisessä paketissa on kiinnitetty erityistä huomioita ison aineiston nopeaan käsittelyyn. PopEpi paketti puolestaan sisältää useita tilastollisia funktioita jotka laskevat väestölähtöisen rekisteritutkimuksen keskeisiä tunnuslukuja, kuten suhteellisen elossaolon arvioita erilaisilla parametrittomilla menetelmillä.

Lisäksi R-ohjelman shiny ympäristöä hyödyntäen on kehitetty vuorovaikutteisen syöpätilasto, josta käyttäjä voi itse hakea ajantasaista syöpään liittyvää tilastotietoa. Uusi internetpohjainen sovellus löytyy osoitteesta: http://tilastot.syoparekisteri.fi/

Uusi sovellus tarjoaa käyttäjälle mahdollisuuden tarkastella koko maan kattavaa syöpien tilastotietoa tietosuojan puitteissa. Käyttäjä voi hakea mm. lukumääriä uusista syöpätapauksista, syöpäkuolemista, elossaolevista syöpäpotilaista tai syöpäpotilaiden elossaoloa diagnoosin jälkeen kuvaavia lukuja. Lisäjaottelua voi tehdä esimerkiksi sukupuolen, alueen ja syöpätaudin suhteen.

”Isohkokin aineisto voidaan muokata sopivilla työkaluilla kaikille saavutettavaan muotoon”, toteaa tilastojohtaja Janne Pitkäniemi.

Verkkosovelluksella voi piirtää sovelluksella hakutuloksia havainnoivia yksinkertaisia kuvaajia sekä karkeita karttapiirroksia voi sovelluksen avulla tuottaa. Uusi palvelu tarjoaa myös mahdollisuuden haetun aineiston tallentamiseen jatkotarkasteluja varten.

Register data –
To be or not to be Big Data

Reijo Sund
Centre for Research Methods, Department of Social Research,
University of Helsinki
e-mail: reijo.sund@helsinki.fi

Data have been produced for thousands of years. The reasons for such production were originally administrative in nature as there was a need for systematically collected numerical facts on a particular subject. Advances in information technology have made it possible to more effectively collect and store larger and larger data sets. As far as there has been data, there has also been a challenge to transform data into useful information and too much data in an unusable form has always been a common complain.

There are more and more ”big data”, but the emphasis has been on technical aspects and not on the information itself. The problem is that data without explanations are useless and that more complex data requires more background information. Big Data are also often secondary data, i.e. not tailored to specific research question at hand.

Finland has a comprehensive collection of nationwide registers composed and maintained for administrative or statistical purposes. Mainly the registers consist of event-based data, such as hospital discharges, reimbursed drug purchases, cancer diagnoses, causes of death etc. Personal identification codes allow deterministic linkages within and between registers. Register data are not Big Data in the sense that the datasets are typically quite small (under 100Gb). On the other hand, register data may (have) be(en) large especially if linked to other data. The register-based data analysis also involves complexity that requires preprocessing that cannot be handled manually – the same problem as with most Big Data.

However, the common belief that big data consist of autonomous, atom-like building blocks is fundamentally erroneous. There are no simple magic tricks to overcome problems arising from the limitations of empirical research and more general aspects of scientific research are needed in order to deal with the related methodological challenges. This is where statistics has a lot to say, as statistics offers not only a set of tools for problem solving, but also a formal way of thinking about the modeling of the actual problem. But rather than trying to squeeze the data into a predefined model or saying too much on what can and cannot be done, data analysis should work to achieve an appropriate compromise between the practical problems and the data.

For more information:

Sund 2003. Utilisation of administrative registers using scientific knowledge discovery. Intelligent data analysis 7 (6), 501–519.
Sund et al. 2014. Use of health registers. In: Handbook of epidemiology, 2nd edition, pp. 707–730.
Sund 2015. Miksi isoon dataan hukutaan? Tieto ja trendit – Talous- ja hyvinvointikatsaus 2/2015, 40–45.

Bayesian Intensity Model for Lexis Diagram

Tommi Härkänen[1], Anna But[2] Jari Haukka[2]
[1]National Institute for Health and Welfare, [2]University of Helsinki

There are various models available to analyze time-to-event data using a single time scale, but usually more than one relevant time scale exists. For instance, observation (calendar) time and age, both appear in many epidemiological studies. The choice of primary (appropriate) time scale can be challenging, and analyzing time-to-event data on two time scales can provide an appealing alternative. We introduce a Bayesian intensity model to analyze two-dimensional point process on Lexis diagram. Furthermore, it can be extended to more general situations such as repeating events or marked point processes. This allows more insightful analyses when compared with, for example, commonly applied stratified Cox's regression models, in which the baseline hazard functions are generally ignored. After a simple discretization of two-dimensional process, we model the intensity by one-dimensional hazard functions. For model parameters, jump points and corresponding hazard levels, we introduce a prior distribution with built-in smoothing feature. We apply the reversible jump Metropolis-Hastings algorithm to sample from posterior distribution, and demonstrate the applicability of the method using simulated survival data. We also perform comparisons with other available methods.

Leo Törnqvist -palkinto:
Paras tilastotieteen
pro gradu -tutkielma
(2015–2016)

A summary of the master's thesis ”Some tools for linear dimension reduction”

Joni Virta
University of Turku,
Department of Mathematics and Statistics

Abstract

Dimension reduction refers to a large family of methods with the common objective of reducing the number of variables in the data while simultaneously losing no information. In *linear dimension reduction* this is accomplished by replacing the original variables with a smaller set of linear combinations of them. Different definitions of “information” then call for different approaches and in the thesis three classical methods of linear dimension reduction are discussed in detail. While seemingly dissimilar, the three methods were also shown to be closely connected under the unifying framework of *simultaneous diagonalization of two scatter matrices.* Finally, two simulation studies comparing the methods in various settings are conducted. The thesis is available online (`http://users.utu.fi/jomivi/publications/`).

1 Dimension reduction and scatter matrices

Nowadays it is becoming less and less uncommon to encounter datasets with the number of variables ranging in tens of thousands. This can naturally lead into problems, both computationally and in interpreting the results of analyses. Perhaps the simplest solution to the problem is given by *linear dimension reduction* where, given a random vector $\boldsymbol{x} \in \mathbb{R}^p$, the objective is to find a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times k}$, $k << p$, such that the mapping

$$\boldsymbol{x} \mapsto \boldsymbol{B}^T \boldsymbol{x},$$

loses no information. The underlying assumption is thus that the information content of $\boldsymbol{x}$, however that is defined, is contained in a few low-dimensional projections onto some suitable subspace.

The above formulation has proved to be rather fruitful in practice, having spanned a great amount of different methods for various data structures. The standard textbook example for i.i.d. data is given by principal component analysis (PCA) which searches for mutually uncorrelated directions having maximal variances. If the uncorrelatedness is replaced with the stronger notion of independence, one arrives to the realm of *independent component analysis* (ICA). Several different methods for solving the so-called independent component problem exist in the literature and the methodology itself is applicable to wide variety of fields, the most prominent being perhaps signal processing. In the thesis the theory behind one particular method of independent component analysis, called the *fourth order blind identification* (FOBI) (Cardoso, 1989), is reviewed and discussed.

Both PCA and FOBI are examples of *unsupervised dimension reduction* where each random variable in $\boldsymbol{x}$ is in a sense considered equal. A fundamentally different class of dimension reduction methods is obtained by introducing an additional response variable y alongside $\boldsymbol{x}$. In *supervised dimension reduction* we are interested in reducing the dimension of the predictor $\boldsymbol{x}$ in the context of regressing y on $\boldsymbol{x}$. A commonly used assumption is then that there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times k}$ such that

$$y \perp \boldsymbol{x} \mid \boldsymbol{B}^T \boldsymbol{x},$$

with k as small as possible (as e.g. $\boldsymbol{B} = \boldsymbol{I}$ of course provides a trivial solution). That is, given the reduced variables in $\boldsymbol{B}^T \boldsymbol{x}$ the original $\boldsymbol{x}$ contains no additional information on y and we can continue the analyses by effectively discarding the original $\boldsymbol{x}$ and retaining only $\boldsymbol{B}^T \boldsymbol{x}$. Two fundamental examples of methods falling into this class are given by *sliced inverse regression* (SIR) (Li, 1991), which is discussed in detail in the thesis, and *sliced average variance estimation* (SAVE) (Cook and Weisberg, 1991).

While the previous examples treated samples of i.i.d. random vectors, linear dimension reduction has been successfully applied to more complex data structures as well. See for example Miettinen et al. (2014) for using *second order blind source separation* (SOBI) to separate mixed time series and Nordhausen et al. (2015) for a blind source separation method applicable to spatially correlated data.

The second major topic of the thesis along the discussion of the three methods, PCA, FOBI and SIR, is their connection to the so-called *simultaneous diagonalization of two scatter matrices*. The use of the technique in statistics originated already in Caussinus and Ruiz-Gazen (1994) but a comprehensive discussion was given in Tyler et al. (2009) where the diagonalization was used to project the data into various *invariant coordinate* *systems.* Later contributions to the theory can be found e.g. in Ilmonen et al. (2012); Liski et al. (2014a). We next shortly review the key points of the technique.

We define a *scatter functional* $\boldsymbol{S}$ to be any $p \times p$ matrix-valued functional on some suitable set of p-variate probability distributions such that

i) $\boldsymbol{S}(F_{\boldsymbol{x}})$ is symmetric and positive-definite,

ii) $\boldsymbol{S}(F_{\boldsymbol{Ax}}) = \boldsymbol{A}\boldsymbol{S}(F_{\boldsymbol{x}})\boldsymbol{A}^T$, for all invertible matrices $\boldsymbol{A} \in \mathbb{R}^{p\times p}$.

The property *ii*) is called affine equivariance. Scatter functionals can be seen as measures of multivariate dispersion and the most familiar example of a scatter functional is the covariance matrix $\boldsymbol{\Sigma} := Cov(\boldsymbol{x})$ with more examples found e.g. in Tyler et al. (2009). The basic idea in the simultaneous diagonalization of two scatter matrices is to take two scatter functionals (matrices) $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ and find a matrix $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p) \in \mathbb{R}^{p\times p}$ such that both

$$\boldsymbol{H}^T\boldsymbol{S}_1(F_{\boldsymbol{x}})\boldsymbol{H} = \boldsymbol{I} \quad \text{and} \quad \boldsymbol{H}^T\boldsymbol{S}_2(F_{\boldsymbol{x}})\boldsymbol{H} = \boldsymbol{D}, \tag{1}$$

where $\boldsymbol{D}$ is a diagonal matrix, hold simultaneously. Assuming that the eigenvalues of $\boldsymbol{S}_1(F_{\boldsymbol{x}})^{-1}\boldsymbol{S}_2(F_{\boldsymbol{x}})$ are distinct the above decomposition is unique up to the signs and order of the columns of $\boldsymbol{H}$. After finding the matrix $\boldsymbol{H}$ a transformation into an invariant coordinate system is given by $\boldsymbol{x} \mapsto \boldsymbol{H}^T\boldsymbol{x}$, see Tyler et al. (2009), and naturally choosing different pairs of scatter functionals $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ leads into different invariant coordinate systems. Also, as suggested by its name, the method is invariant to affine transformations: both $\boldsymbol{x}$ and $\boldsymbol{Ax}$ lead, up to sign and order, to the same resulting invariant coordinates for any invertible matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$.

The next three sections are devoted to summarizing the three linear dimension reduction methods, PCA, FOBI and SIR, and showing how they are connected to the diagonalization of two scatter matrices. The sections are, unless otherwise stated, based on Jolliffe (2002); Oja and Nordhausen (2012); Li (1991), respectively.

2 Principal component analysis

Principal component analysis is one of the simplest dimension reduction methods and is based solely on the use of second moments in the form of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}$. In PCA the objective is to find mutually uncorrelated directions $\boldsymbol{u}_j^T\boldsymbol{x}$, *principal components*, with highest possible variances. By the properties of the covariance matrix this can be formulated as the following sequence of optimization problems:

$$\boldsymbol{u}_j = \underset{\boldsymbol{u}_j^T \boldsymbol{u}_j = 1}{\operatorname{argmax}} \left(\boldsymbol{u}_j^T \boldsymbol{\Sigma} \boldsymbol{u}_j \right), \quad \text{subject to } \boldsymbol{u}_j^T \boldsymbol{\Sigma} \boldsymbol{u}_l = 0, \quad l = 1, \ldots, j-1,$$

for $j = 1, \ldots, p$. The technique of Lagrange multipliers can then be used to show that the directions $\boldsymbol{u}_j$ are actually the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ and the variances of the principal components $\boldsymbol{u}_j^T \boldsymbol{x}$ are given by the corresponding eigenvalues d_j. That is, the matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p) \in \mathbb{R}^{p \times p}$ is found from the eigendecomposition

$$\boldsymbol{\Sigma} = \boldsymbol{U} \boldsymbol{D} \boldsymbol{U}^T,$$

where $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_p)$ contains the eigenvalues.

Because of both its attractive intuitiveness and computational simplicity PCA is very commonly used in data preprocessing when working with high-dimensional data sets. However, some care should be taken when using it: firstly, it is not affine invariant meaning that any non-trivial affine transformation $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, including marginal scaling, changes the resulting principal components and, secondly, it uses only second moments in finding the components. If the data exhibits some structure hidden only in the higher moments PCA cannot find it. In such a situation an alternative to consider is then the fourth moment-based FOBI of the next section. Third interesting aspect of PCA is the estimation of the true dimension k. Various techniques including the use of scree plots are discussed in Jolliffe (2002) and hypothesis test-based methods can be found in Liski et al. (2014b).

In Liski et al. (2014a) it was shown that PCA can be formulated, if not fully in the guise of ICS, at least in a very similar fashion. Namely, choosing in (1) $\boldsymbol{S}_1(F_{\boldsymbol{x}}) = \operatorname{diag}(\boldsymbol{\Sigma})$ and $\boldsymbol{S}_2(F_{\boldsymbol{x}}) = \boldsymbol{\Sigma}$ is equivalent to performing PCA to marginally standardized $\boldsymbol{x}$. The reason why this fails to be an example of true ICS is of course that $\boldsymbol{S}_1(F_{\boldsymbol{x}})$ is not an actual scatter matrix, failing to be affine equivariant.

3 Fourth order blind identification

In independent component analysis the observations $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, are assumed to be i.i.d. realizations of the random variable $\boldsymbol{x} \in \mathbb{R}^p$ obeying the model

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Omega} \boldsymbol{z}, \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location parameter, the non-singular $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ is the so-called *mixing matrix* and the unobserved random vector $\boldsymbol{z}$ is assumed to have mutually independent components. The objective in ICA is then to estimate some *unmixing matrix* $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma x}$ has mutually independent components.

It can be shown that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ of the independent component model (2) are in the above formulation not identifiable and some additional constraints need to introduced. The usual ones include constraining the independent components to be standardized, $\mathrm{E}(\boldsymbol{z}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{I}$, and limiting the number of Gaussian components in $\boldsymbol{z}$ to at most one, after which the only invariant part of (2) is the order and signs of the independent components (or equivalently, the columns of $\boldsymbol{\Omega}$). For more discussion on these constraints, see for example Hyvärinen et al. (2001).

Perhaps the simplest method for solving the independent component problem is the fourth order blind identification (FOBI) (Cardoso, 1989), requiring just the use of two eigendecompositions. The first step in FOBI is to center and standardize the observed random vector with any inverse square root of $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{x})$, giving the standardized random vector,

$$\boldsymbol{x}_{st} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}]). \tag{3}$$

A standard result in ICA then says that the standardized random vector is only a rotation away from the solution, that is, $\boldsymbol{x}_{st} = \boldsymbol{U z}$ where $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ is some unknown orthogonal matrix. Thus standardization reduces the problem from estimating an inverse of the matrix $\boldsymbol{\Omega}$ to that of estimating an inverse (i.e. transpose) of the orthogonal matrix $\boldsymbol{U}$. Introducing now the *matrix of multivariate kurtosis*,

$$\boldsymbol{\beta}_2(\boldsymbol{x}) := \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}\boldsymbol{x}^T],$$

it is easily seen that $\boldsymbol{\beta}_2(\boldsymbol{x}_{st}) = \boldsymbol{U}\boldsymbol{\beta}_2(\boldsymbol{z})\boldsymbol{U}^T$. Furthermore, $\boldsymbol{\beta}_2(\boldsymbol{z})$ can be shown to be diagonal with the jth diagonal element equal to $\mathrm{E}[z_j^4] + (p-1)$, meaning that the matrix $\boldsymbol{U}$ can be estimated from the eigendecomposition of $\boldsymbol{\beta}_2(\boldsymbol{x}_{st})$. The final solution is then given by the mapping $\boldsymbol{x} \mapsto \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])$.

For the independent components to be estimated properly it is necessary to introduce the additional assumption that the eigenvalues of $\boldsymbol{\beta}_2(\boldsymbol{z})$, or equivalently the fourth moments $\mathrm{E}[z_j^4]$, $j = 1, \ldots, p$, are distinct. Additionally, the eigenvalues can be used to say something about the true dimension k; as the search for independent components can heuristically be seen as the seeking of maximally non-Gaussian directions (Hyvärinen and Oja, 2000), the most interesting components should be the ones having fourth moments differing the most from three, the fourth moment of the standard normal distribution. Also hypothesis testing could be considered but introduces some additional problems now as the most interesting components are not only the ones with the highest eigenvalues but possibly scattered on both ends of the spectrum. Note that these procedures in a sense violate the assumption on maximally one Gaussian component.

Finally, to put FOBI into the two-scatter-matrices context it was observed in Oja et al. (2006); Tyler et al. (2009) that it is equivalent to simultaneously diagonalizing the scatter matrices $\boldsymbol{S}_1(F_{\boldsymbol{x}}) = \boldsymbol{\Sigma}$ and

$$\boldsymbol{S}_2(F_{\boldsymbol{x}}) = \frac{1}{p+2}\mathrm{E}\left[(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])^T\right],$$

the so-called *covariance matrix based on fourth moments.* Note that the fact that FOBI yields a formulation using the diagonalization of two scatter matrices also goes to show that FOBI as a method is actually affine invariant.

4 Sliced inverse regression

The sliced inverse regression is based on the assumption that there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{p\times k}$ with minimal k such that

$$y \perp \boldsymbol{x} \mid \boldsymbol{B}^T\boldsymbol{x}.$$

Note that what is identifiable in the model above is not the parameter $\boldsymbol{B}$ but rather its k-dimensional column space col($\boldsymbol{B}$). So instead of estimating a matrix we are interested in estimating a subspace, called the *effective dimension reduction subspace.*

Assume next that the dimension k is known. It can be shown that under a suitable condition, see Li (1991), the *standardized inverse regression curve* (SIRC),

$$\mathrm{E}[\boldsymbol{x}_{st}|y] = \mathrm{E}[(x_{st,1}, \ldots, x_{st,p})^T|y],$$

where $\boldsymbol{x}_{st}$ is the standardized observation (3), lies entirely in the space col($\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}$). As the inverse regression curve is actually a p-variate random vector the subspace it resides in can be estimated from the eigendecomposition of its covariance matrix; the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, of Cov(E[$\boldsymbol{x}_{st}|y$]) corresponding to the k largest eigenvalues span col($\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}$). Consequently the effective dimension reduction subspace is equal to col($\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}$) where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \in \mathbb{R}^{p\times k}$.

To estimate the standardized inverse regression curve SIR used *slicing.* That is, assuming a standardized sample $(\boldsymbol{x}_{1,st}, y_1), \ldots, (\boldsymbol{x}_{n,st}, y_n)$, we divide the observed range of the response variable y into H slices with p_h denoting the proportion of observations falling in the hth slice, $h = 1, \ldots, H$. The slice means $\boldsymbol{m}_h \in \mathbb{R}^p$ of the vectors $\boldsymbol{x}_{i,st}$ then provide point-wise estimates for the standardized inverse regression curve in the corresponding slices and an estimate for the covariance matrix of the standardized inverse regression curve is given by

$$\boldsymbol{C} := \sum_{h=1}^{H} p_h \boldsymbol{m}_h \boldsymbol{m}_h^T.$$

Collecting the eigenvectors of $\boldsymbol{C}$ corresponding to the k largest eigenvalues as the columns of the matrix $\hat{\boldsymbol{V}} \in \mathbb{R}^{p \times k}$ the estimated reduced predictors are then $\hat{\boldsymbol{V}}^T \boldsymbol{x}_{i,st}$, $i = 1, \ldots, n$.

Some possible issues concerning the use of SIR include choosing the number of slices and the estimation of non-monotonous relationships. The former is discussed in Li (1991) with the conclusion that while the number of slices can affect the asymptotic variances of the estimated directions, root-n consistency is still achieved, no matter what the number of slices is. An alternative for the slicing is given if one instead uses splines to estimate the standardized inverse regression curve. This approach was discussed in Kent (1991); Zhu and Yu (2007) and in the thesis a slightly differing spline-based approach is given. That is, instead of estimating the covariance matrix of the SIRC we estimate the covariance matrix of the residuals, $\boldsymbol{x}_{st} - \mathrm{E}[\boldsymbol{x}_{st}|y]$, of the fitted SIRC. As a consequence the directions spanning the dimension reduction subspace are now the eigenvectors corresponding to the k *smallest* eigenvalues.

The second issue stems from the fact that if the slice means do not vary enough in some particular direction that direction may go unnoticed by SIR, even if it contains a clearly visible structure, see the thesis for an example of this. In such a case one possibility is to use instead SAVE (Cook and Weisberg, 1991) which operates not on slice means but slice variances.

Also SIR can be stated in the form of diagonalization of two scatter matrices as shown in Liski et al. (2014a), implying that it too is affine invariant. Namely, choosing again $\boldsymbol{S}_1(F_{\boldsymbol{x}}) = \boldsymbol{\Sigma}$ and either $\boldsymbol{S}_2(F_{\boldsymbol{x}}) = \mathrm{Cov}(\mathrm{E}[\boldsymbol{x}|y])$ or $\boldsymbol{S}_2(F_{\boldsymbol{x}}) = \mathrm{Cov}(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}|y])$ leads to regular SIR and the SIR based on residuals of the standardized inverse regression curve, respectively. Note that $\boldsymbol{S}_2(F_{\boldsymbol{x}}) = \boldsymbol{S}_2(F_{\boldsymbol{x},y})$ is now a *supervised scatter functional*, see Liski et al. (2014a).

5 Additional remarks

The thesis concludes with two simulation studies. The first one compares using the three discussed methods as pre-processing steps in regression. The results indicated that while SIR is by definition the only method actually utilizing the joint distribution of $\boldsymbol{x}$ and y in the estimation of the reduced predictors, it too can sometimes fail in estimating them correctly. That is, in the study SIR failed to find a circle-shaped dependency between $\boldsymbol{x}$ and y for the exact reasons described in the previous section. Additional comparison was still done using the "Boston" data in the R-package *MASS* (Venables and Ripley, 2002) with the predictable conclusion that SIR was far superior in finding the relevant predictors as compared to PCA and FOBI.

The second simulation study compares the standard SIR with the proposed residual-based spline-SIR (SSIR) using two different models. The indication of the study was that while SSIR gave more accurate estimates than SIR and introduced additional tuning parameter in the degree of the splines used, also SSIR suffered from the inability to estimate dependencies of certain form. Thus, rather than tinkering with various parameters one is likely better off by trying multiple different dimension reduction methods instead.

References

Cardoso, J.-F. (1989). Source separation using higher order moments. In *Proceedings of IEEE international conference on acoustics, speech and signal processing*, pages 2109–2112. IEEE.

Caussinus, H. and Ruiz-Gazen, A. (1994). Projection pursuit and generalized principal component analysis. *New Directions in Statistical Data Analysis and Robustness*, pages 35–46.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley-Interscience.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430.

Ilmonen, P., Oja, H., and Serfling, R. (2012). On invariant coordinate system (ICS) functionals. *International Statistical Review*, 80(1):93–110.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Kent, J. T. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):pp. 336–337.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Liski, E., Nordhausen, K., and Oja, H. (2014a). Supervised invariant coordinate selection. *Statistics*, 48(4):711–731.

Liski, E., Nordhausen, K., Oja, H., and Tyler, D. (2014b). Estimation of dimension reduction subspace in PCA. Manuscript.

Miettinen, J., Nordhausen, K., Oja, H., and Taskinen, S. (2014). Deflation-based separation of uncorrelated stationary time series. *Journal of Multivariate Analysis*, 123:214–227.

Nordhausen, K., Oja, H., Filzmoser, P., and Reimann, C. (2015). Blind source separation for spatial compositional data. *Mathematical Geosciences*, 47(7):753–770.

Oja, H. and Nordhausen, K. (2012). Independent component analysis. In El-Shaarawi, A.-H. and Piegorsch, W., editors, *Encyclopedia of Environmetrics*, pages 1352–1360. John Wiley & Sons, , Chichester, UK, 2nd edition.

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35(2):175–189.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):549–592.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Zhu, L.-P. and Yu, Z. (2007). On spline approximation of sliced inverse regression. *Science in China Series A: Mathematics*, 50(9):1289–1302.

ILTAPÄIVÄSEMINAARI 11.6.2015
Arkaluonteiset asiat tilastollisessa tutkimuksessa

Aika: 11.6. klo 14.00–17.00.

Paikka: Helsingin yliopiston päärakennuksen vanhan puolen auditorio XV.

Ohjelma:

- *associate professor Jouni Kuha (London School of Economics):* Arkaluonteiset aiheet: Uusia satunnaistettuja kyselymenetelmiä
- *vanhempi tutkija Markku Heiskanen (Heuni):* Arkaluonteiset kyselyt
- *yliaktuaari Päivi Hokka (Tilastokeskus):* Seksiaiheinen kyselytutkimus
- *professori emeritus Seppo Laaksonen:* kommenttipuheenvuoro.

Arkaluonteisten asioiden selvittäminen on ajankohtainen tutkimusaihe. Jouni Kuha julkaisi viime vuonna tutkimuksen, jossa pystyttiin selvittämään varastettuun tavaraan ryhtymisen yleisyyttä. (The item count method for sensitive survey questions: modelling criminal behavior. *Journal of the Royal Statistical Society*, *Applied Statistics*.) Menetelmää voidaan käyttää muidenkin arkaluonteisten asioiden selvittämiseen.

Markku Heiskanen on kerännyt ja analysoinut surveyaineistoja arkaluonteisista aiheista kuten seksistä, alkoholin käytöstä ja turvallisuudesta. Hän on erikoistunut rikosuhritutkimuksiin ja naisten ja miesten kokemaan väkivaltaan. Julkaisuja: MH ja Minna Piispa (1998) Usko, toivo, hakkaus. Kyselytutkimus miesten naisille tekemästä väkivallasta. Tilastokeskus, Oikeus 1998:12, MH (2002) Väkivalta, pelko, turvattomuus. Survey-tutkimusten näkökulmia suomalaisten turvallisuuteen. Tilastokeskus, tutkimuksia 236 sekä MH ja Ruuskanen, Elina (2011) Men’s experiences of Violence in Finland 2009. HEUNI, Publication Series No. 71.

Päivi Hokka työskentelee Tilastokeskuksen Tiedonhankinta-yksikössä. Hän toimii projektinvetäjänä tiedonkeruissa ja suunnittelee ja kehittää haastattelu- ja kyselytutkimuksia. Hän on ollut toteuttamassa Juomatapatutkimuksen, Päihdetutkimuksen ja Finsex-tutkimuksen aineistojen keruita.

Seksiaiheinen kyselytutkimus

Päivi Hokka
yliaktuaari
Tilastokeskus, Tiedonhankintayksikkö

Seksuaalisuus lukeutuu arkaluontoisimpiin aiheisiin, joita kyselytutkimuksissa on selvitetty. Suomessa seksistä kysyttiin väestöotokselta ensimmäistä kertaa vuonna 1971 (Sievers, Koskelainen, Leppo, 1974). Tämän jälkeen aihetta on tutkittu laajasti Finsex-hankkeen survey-tutkimuksilla vuosina 1992, 1999, 2007 (Kontula, 2008) sekä vuonna 2015. Tutkimussarjassa seurataan suomalaisten parisuhteissa, seksuaaliasenteissa, seksuaalisessa käyttäytymisessä ja seksuaaliongelmissa tapahtuvia muutoksia. Finsex-tutkimusten tiedonkeruut on toteuttanut Tilastokeskus.

Tilastokeskuksen ammattihaastattelijat keräsivät vuoden 1992 Finsex-tutkimuksen tiedot käyntihaastatteluina. Kaikkein arkaluontoisimmat kysymykset kysyttiin erillisellä lomakkeella, jonka vastaaja täytti haastattelijan näkemättä. Vuosina 1999 ja 2007 tiedot kerättiin postikyselyllä. Vuonna 2015 käytettiin internetkyselyn ja postikyselyn yhdistelmää eli mixed-mode –tiedonkeruutapaa.

Itsetäytettävät kyselyt sopivat parhaiten arkaluonteisten aiheiden kysymiseen. Kääntöpuolena on niiden alhaisempi vastausprosentti verrattuna haastatteluihin. Vuoden 1992 käyntihaastatteluissa saavutettiin 76 prosentin vastausosuus, mutta seitsemän vuotta myöhemmin postikyselyssä vastausprosentti oli vain 46. Vastaushalukkuuden aleneminen on ollut yleinen kehitys haastattelu- ja kyselytutkimuksissa, mutta posti- ja web-kyselyt ovat kärsineet siitä eniten. Aiheen ollessa arkaluonteinen ongelma vielä korostuu.

Finsex-hankkeen viimeisin tiedonkeruu, *Tutkimus ihmissuhteista, seksuaalisista elämäntyyleistä ja asenteista 2015,* tehtiin Tilastokeskuksen tiedonhankintayksikössä Väestöliiton toimeksiannosta. Tutkimusaineiston analysoinnista vastaa Väestöliiton tutkimusprofessori Osmo Kontula. Tässä esityksessä kerrotaan kyselyn toteutuksesta tiedonkeruun näkökulmasta. Tutkimukseen poimittiin 6000 henkilön satunnaisotos 18–79 -vuotiaasta Suomen väestöstä. Tutkimussisältö oli pitkälti samanlainen kuin aikaisemmissa sarjan tutkimuksissa, täydennettynä uusilla kysymyksillä.

Tiedonkeruussa otokseen valittuja lähestyttiin maksimissaan neljä kertaa. Ensimmäisellä kerralla lähetettiin kirje, jossa oli ohjeet ja tunnukset internetlomakkeella vastaamista varten. Kirjeen mukana oli myös erillinen tutkimuksesta kertova esite. Pelkästä internetvastausmahdollisuudesta ensimmäisellä kierroksella on etua arkaluonteisessa aiheessa, koska vastaaja ei heti näe kysymyksiä eikä siten voi ”pelästyä” niiden sisältöä. Seuraavilla kolmella kierroksella lähetettiin paperilomake, jonka kannessa olevassa saatekirjeessä oli myös tunnukset ja ohjeet vaihtoehtoiseen vastaamiseen internetissä. Poikkeuksena tähän olivat alle 36-vuotiaat, joille lähetettiin paperilomake vasta kolmannella yhteydenottokerralla, sekä yli 65-vuotiaat, joille lähetettiin paperilomake heti ensimmäisellä kerralla. Näin tehtiin siksi, että kokemusten mukaan nuoret valitsevat useimmin nettivastaamisen, ja iäkkäämmät ihmiset paperilomakevastaamisen.

Koska posti- ja webkyselyissä haastattelija ei ole suostuttelemassa vastaajaa, kyselyn saatekirje on keskeisessä roolissa. Nykyisin Tilastokeskuksen tiedonhankintayksikössä vakiintuneena käytäntönä ovat eri vastaajaryhmille kohdennetut saatekirjeet, joiden sisältöä muokataan vastaanottajaryhmälle sopivaksi. Finsex-tutkimuksessa kirjeitä eriytettiin lisäksi vastaajan sukupuolen mukaan. Ensimmäisellä kierroksella lähettiin viisi erilaista kirjettä seuraavalla jaottelulla: alle 36-vuotiaat naiset, alle 36-vuotiaat miehet, 36-65-vuotiaat naiset, 36–65-vuotiaat miehet ja kaikki yli 65-vuotiaat.

Nuorten kirjeissä käytettiin rennompaa otetta ja sinuttelua, kun taas vanhempia lähestyttiin virallisemmin ja teititellen. Ylipäätään virallisuus ja asiallisuus lisäävät luottamusta arkaluontoisessa aiheessa. Naisten ja miesten kirjeissä painotettiin hiukan erilaisia asioita. Viimeisellä kierroksella kaikille vielä vastaamattomille lähetettiin tutkimusprofessori Kontulan allekirjoittama ja hänen kuvallaan varustettu henkilökohtaisen sävyinen, vetoava kirje. Kaikissa kirjeissä kerrottiin, että kyselyyn vastanneiden kesken arvotaan kolme taulutietokonetta.

Lopputuloksena vuoden 2015 Finsex-tutkimuksen tiedonkeruussa saatiin kaikkiaan 2 150 vastausta. Vastausosuus oli tutkimussarjan historian alhaisin, 36 prosenttia. Kuten yleensä kyselyissä, naiset vastasivat miehiä innokkaammin. Ainoastaan 75–79-vuotiailla miesten vastausosuus oli naisia korkeampi. Eniten kyselyyn vastasivat alle 35-vuotiaat naiset (45 %) − ihmissuhteet ja seksuaalisuus ovat tätä ryhmää kiinnostava aihe. Myös 55–74-vuotiaat naiset vastasivat melko hyvin. Alle 25-vuotiaiden miesten vastausosuus oli heikoin (23 %) − vastaava ilmiö on tuttu kyselytutkimuksista yleisemminkin. Miehistä yli 54-vuotiaat vastasivat innokkaammin kuin heitä nuoremmat miehet. Vanhimmissa ikäryhmissä kyselyihin vastaaminen koettaneen useammin velvollisuudeksi kuin nuorimmissa. Tosin 75 vuotta täyttäneiden naisten vastausosuus oli vain 28 prosenttia – oletettavasti tutkimusaihe tuntui heistä erityisen arkaluontoiselta.

Vastauksista 57 prosenttia saatiin internetissä ja 43 prosenttia paperilomakkeella. Mitä nuorempi vastaaja, sitä todennäköisemmin vastaus annettiin netissä. Yli 65-vuotiaat valitsivat yleisimmin paperilomakevastaamisen.

Tutkimusaiheen arkaluontoisuus vaikutti vastausosuuteen selvästi. Tilastokeskuksen tiedonhankintayksikössä väestöotoksille tehdyissä posti- ja webkyselyissä on viime vuosina saatu 45–50 prosentin vastausosuuksia. Tulevaisuudessa entistä suurempana haasteena on ihmisten motivoiminen vastaamaan kyselytutkimuksiin ylipäätään, ja erityisesti arkaluontoisimpiin aiheisiin.

Kirjallisuus:

Sievers Kai, Koskelainen Osmo & Leppo Kimmo (1974): Suomalaisten sukupuolielämä. Porvoo: WSOY.

Kontula Osmo (2007): Halu ja intohimo. Tietoa suomalaisesta seksistä. Helsinki: Otava.

The item count method for sensitive survey questions

Jouni Kuha
Department of Statistics, London School of Economics and
Political Science Houghton St, London WC2A2AE, Iso-Britannia.
j.kuha@lse.ac.uk

One of the many challenging problems in survey research is asking question about sensitive topics such as illegal behaviour, sexual activity, or socially undesirable opinions and prejudices. A number of methodological approaches may be used to try to elicit truthful information on topics like these. One of them is to ask questions in such a way that the respondents can be confident — if they understand the instructions correctly — that their individual answers cannot be known to the interviewer. An increasingly commonly used method of this kind is the *item count* or "list experiment" question (Miller, 1984). Its properties, assumptions and analysis were discussed in this seminar.

The topic of an item count question is whether or not the respondent has engaged in some sensitive behaviour. For instance, in the application considered by Kuha and Jackson (2014) this was whether the respondent had bought stolen goods in the past year. The question, however, is not directly and only about this behaviour. Instead, it presents a list of activitities ("items"), typically 4–10 of them, and asks the respondent to report only the total number of how many of these activitities he or she has done in (say) the past year. One of the items is the sensitive one that we are actually interested in, and the rest (the *control items*) are activities which we would not expect to be sensitive to the respondents. The control items are of no separate interest for the survey, and they serve only to hide the specific answer to the sensitive item. Each respondent is randomly (and silently) assigned either to the *control group* who receive a list of the control items only, or to the *treatment group* whose item count list includes also the sensitive item.

Data from an item count question are best analysed by treating them as an instance of incomplete categorical data. Let Y denote a respondent's (unknown) answer to the sensitive item, with the values 0 for No and 1 for Yes, and Z the total number of the J control items to which the answer is Yes, with the possible values $0, 1, \ldots, J$. The reported total count that we receive as the answer to the item count question is then $S = Z$ for respondents in the control group and $S = Z + Y$ in the treatment group. What we want to estimate are the parameters of a model for Y given a (possibly empty) set of explanatory variables X. In addition, it is also necessary to specify a model for the control total Z given X and Y. Both of these models are identifiable from the item count data, because of the randomized allocation of respondents to the treatment and control groups. Methods of estimating these models are described by Imai (2011) and Kuha and Jackson (2014).

The principles and analysis of the item count method are straightforward in theory. In practice, however, its validity and efficiency depend crucially on how the respondents react to the question, and how well the assumptions of the analysis are satisfied. Many of the assumptions concern the list of the control items. These are a peculiar feature of the item count technique, and potentially its Achilles' heel. Even though they are of no direct interest themselves, the control items still need to be included in the question and the analysis, and errors associated with them can distort conclusions about the sensitive item of interest. For validity, the model for the control total should be sufficiently correctly specified. For efficiency, it would be best if the activitities described by the control items were independent of the sensitive behaviour — but if they were too obviously different in content, this might cause the respondents to react badly to the question because of the apparently senseless nature of the list of items. Because of such difficulties, designing a well-behaved item count question presents a formidable practical challenge.

In the seminar, possible choices in the modelling of the control items were described, drawing on the discussion in Kuha and Jackson (2014). Also reported were some preliminary results from a cross-national survey which included an item count question combined with an experiment with different sets of control items.

References

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association 106*, 407–416.

Kuha, J. and J. Jackson (2014). The item count method for sensitive survey questions: Modelling criminal behaviour. *Journal of the Royal Statistical Society, Series C 63*, 321–341.

Miller, J. D. (1984). *A New Survey Technique for Studying Deviant Behavior*. PhD thesis, The George Washington University.

Gunnar Modeen -minnesmedaljen

Jukka Hoffrén

Statistiska Samfundet i Finland r.f. har i samband med de nordiska statistikda-garna traditionsenligt delat ut Gunnar Modeen -minnesmedaljen till särskilt meri-terade statistiker. Praxisen har varit att dela ut medaljen till en representant för det land där statistikdagarna hålls.

Gunnar Modeen -minnesmedaljen beviljas för en betydande livsgärning inom sta-tistikbranschen. Meningen är att den person som belönas är en framstående se-nior expert inom statistikbranschen, som uttryckligen utmärkt sig i det praktiska statistikarbetet och som uppskattas av sina kolleger.

Styrelsen för Statistiska Samfundet väljer den person som får medaljen och me-daljen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga överlåts medaljen till en betydande nordisk statistiker från det land som respek-tive år arrangerar mötet. Den första medaljen överläts vid det nordiska statisti-kermöte som hölls i Finland år 1989.

Bakgrunden till och kriterier för GM-minnesmedaljen

Efter Gunnar Modeens bortgång år 1988 grundades en medaljfond till hans minne. Medaljen utarbetades på basis av den medaljong som Gunnar Modeens familj gett konstnären Matti Haupt i uppdrag att utforma till Modeens 70-årsdag år 1965. Mottagaren av medaljen väljs av styrelsen för Statistiska Samfundet i Finland och medaljen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga överlåts medaljen till en betydande nordisk statistiker från det land som respektive år arrangerar mötet. Den första medaljen överläts vid Nor-diska Statistikermötet i Finland år 1989. Priset utdelas vart tredje år till en meri-terad statistiker från det land där Nordiska Statistikermötet anordnas.

Allmänna kriterier för Gunnar Modeen -minnesmedaljen:

- priset beviljas för en betydande livsgärning inom statistikbranschen.

Den person som tilldelas medaljen:

- är en expert inom statistikbranschen, som uttryckligen utmärkt sig i det prak-tiska statistikarbetet
- är en nordisk, framstående senior expert som uppskattas av sina kolleger,
- har akademisk examen (magister, licentiat eller doktor) och
- är villig att ta emot GM-medaljen

Mottagare av GM-minnesmedaljen

Den första medaljen tilldelades Mauno Koivisto, Finlands dåvarande president, som en särskild hedersbetygelse. År 1989 var han beskyddare av Nordiska Sta-tistikermötet i Finland som firade 100-årsjubileum för nordisk statistik. Ytterli-gare en medalj delades ut på mötet och mottagare var professor Eino H. Laurila. Övriga mottagare av medaljen:

År 1992 tilldelades medaljen inte.
År 1995 direktör Poul Jensen, Danmarks Statistik.
År 1998 professor Sven Nordbotten, Universitetet i Bergen.
År 2001 professor Emeritus Gunnar Kulldorf, Umeå universitet.
År 2004 direktör Asta Manninen, Helsingfors stads faktacentral.
År 2007 generaldirektör Hallgrímur Snorrason, Hagstofa, Island.
År 2010 direktör Lars Thygesen, Danmarks Statistik.
År 2013 Liv Hobbelstad Simpson, pensionerad från Statistisk sentralbyrå (SSB) som Head of National accounts och past chair of IARIW
År 2016 Eva Elvers, PhD, pensionerad från Design and Plan & Build and Test som Process owner

Scandinavian Journal of Statistics

Recognised as a leading journal in its field, the Scandinavian Journal of Statistics is an international publication devoted to reporting significant and innovative original contributions to statistical methodology, both theory and applications. The journal specializes in statistical modelling showing particular appreciation of the underlying substantive research problems. Scandinavian Journal of Statistics is published on behalf of the Danish Society for Theoretical Statistics, the Finnish Statistical Society, the Norwegian Statistical Society and the Swedish Statistical Society. Journal is currently edited by professors Peter Dalgaard and Niels Richard Hansen. National editor for Finland is Jukka Corander (University of Helsinki). The chairman of the board is Thomas Scheike and other members of the board are Juha Karvanen, Jukka Corander, Geir Olve Storvik, Rolf Larsson, J.Hjelmborg and Hans Karlsen.

Members of the Finnish Statistical Society entitled to discount prices when ordering the Scandinavian Journal of Statistics. For further information please see webpage:

http://www.wiley.com/bw/subs.asp?ref=0303-6898&site=1

ISI Journal Citation Reports® Ranking: 2015: 60/123 (Statistics & Probability) **Impact Factor:** 0.908

Suomen Tilastoseuran hallitus vuonna 2015

Board members of the Finnish Statistical Society 2015

Puheenjohtaja Chair	Jyrki Möttönen	Filosofian toht. PhD
Varapuheenjohtaja Vice Chair	Ari Jaakola	Filosofian maist. M.Sc
Rahastonhoitaja Treasurer	Emma Kämäräinen	Valtiotiet. kand. B.Soc.Sc.
Sihteeri Secretary	Paula Bergman	Luonnontiet. kand. B.Sc.
Jäsen Member	Tommi Härkänen	Filosofian tohtori PhD
Jäsen Member	Tara Junes	Valtiotiet. maist. M.Soc.Sc.
Jäsen Member	Marjo Kaasila	Filosofian maist. M. Sc.
Jäsen Member	Marianne Laalo	Valtiot. yo Student of Soc.Sc.
Jäsen Member	Pekka Pere	Doctor of Philosophy DPhil
Jäsen Member	Marjo Pyy-Martikainen	Filosofian toht. PhD

Suomen Tilastoseuran hallitus vuonna 2016

Board members of the Finnish Statistical Society 2016

Puheenjohtaja	Jyrki Möttönen	Filosofian toht.
Chair		PhD
Varapuheenjohtaja	Ari Jaakola	Filosofian maist.
Vice Chair		M.Sc.
Rahastonhoitaja	Emma Kämäräinen	Valtiotiet. kand.
Treasurer		B.Soc.Sc.
Sihteeri	Paula Bergman	Luonnontiet. kand.
Secretary		B.Sc.
Jäsen	Tommi Härkänen	Filosofian tohtori
Member		PhD
Jäsen	Tara Junes	Valtiotiet. maist.
Member		M.Soc.Sc.
Jäsen	Pihla Oksanen	Valtiotiet. yo
Member		Student of Soc.Sc.
Jäsen	Pekka Pere	Doctor of Philosophy
Member		DPhil
Jäsen	Johanna Seppänen	Filosofian tohtori
Member		PhD

Suomen Tilastoseuran julkaisuja

Publikationer utgivna av Statistiska Sammanfundet

Publications issued by the Finnish Statistical Society

1. Monikielinen väestötieteen sanakirja, suomenkielinen laitos, Helsinki 1962.
   Multilangual Demographic Dictionary, Finnish section, Helsinki 1962.

2. Suomen Tilastoseura – Statistiska Sammanfundet i Finland 1920-1970, Porvoo – Borgå 1970.

3. Pohjoismainen tilastosanasto, toinen tarkistettu laitos.
   Nordisk statistik nomenklatur, andra reviderade upplagan.
   Nordic statistical nomenclature, 2nd revised edition. Jyväskylä 1975 (loppuunmyyty)

4. Aikasarja-analyysin menetelmiä, Helsinki 1977.

5. Pekka Tavaila: Leo Törnqvist Posti- ja lennätinhallituksen liiketaloudellisen tutkimuslaitoksen esimiehenä 1949–1977, Helsinki 1982.

6. Otanta teoriassa ja käytännössä. Vesa Kuusela ja Leif Nordberg (toim.). Helsinki 1986.

7. Suomen Tilastoseura 70 vuotta. Statistiska Sammanfundet i Finland 70 år.
   The Finnish Statistical Society 70 years. Helsinki 1991.

Tilastotieteellisiä tutkimuksia

Statistiska undersökningar

Statistical Research Reports

ISSN 0356–3499

1. Pentti Manninen: Puolueiden kannatusosuuksien estimoinnin tarkkuus Demingin vyöhykepoiminnassa. [The Accuracy of Party Support Estimation in Deming Zone Selection.] In Finnish with English Summary. Helsinki 1976.

2. Timo Hakulinen: On Competing Risks of Death. Helsinki 1977.

3. Lars-Erik Öller: Time Series Analysis of Finnish Foreign Trade. Helsinki 1978.

4. Pekka Laippala: The Empirical Bayes Two-Action Rules with Floating Optimal Sample Size and Exponential Conditional Distributions. Helsinki 1980.

5. Markku Nurminen: Some Developments in Quantitative Methods of Epidemiology. Helsinki 1982.

6. Pentti Saikkonen: Comparing Asymptotic Properties of Some Tests Used in the Specification of Time Series Models. Helsinki 1985.

7. Lauri Tarkkonen: On Reliability of Composite Scales. Helsinki 1987.

8. Juni Palmgren: Models for Categorical Data with Errors of Observation. Helsinki 1987.

9. Ari Veijanen: On Estimation of Parameters of Partially Observed Random Fields and Mixing Processes. Helsinki 1989.

10. Ritva Luukkonen: On Linearity Testing and Model Estimation in Non-Linear Time Series Analysis. Helsinki 1990.

11. Hely Salomaa: Factor Analysis of Dichotomous Data. Helsinki 1990.

12. Kenneth Nordström: Contributions to the Comparison of Linear Models and to the Löwner-Ordering Antitonicity of Generalized Inverses. Helsinki 1990.

13. Seppo Laaksonen: Handling Household Survey Nonresponce Data. Helsinki 1992.

14. Mervi Eerola: On Predictive Causality in the Statistical Analysis of a Series of Events. Helsinki 1993.

15. Mikael Linden: Studies in Integrated and Co-Integrated Economic Time Se-ries. Helsinki 1995.

16. Tadeusz Dyba: Precision of Cancer Incidence Predictions Based on Poisson Distributed Observations. Helsinki 2000.

17. Kimmo Vehkalahti: Reliability of Measurement Scales. Helsinki 2000.

18. Sirpa Heinävaara: Modelling survival of patients with multiple cancers. Helsinki 2003.

Suomen Tilastoseuran vuosikirja

Årsbok för Statistiska Sammanfundet i Finland

The Yearbook of the Finnish Statistical Society

ISBN 0355–5941

1975, Helsinki 1976
1976, Helsinki 1977
1977, Helsinki 1978
1978, Helsinki 1979
1979, Helsinki 1980
1980, Helsinki 1981
1981, Helsinki 1982
1982, Helsinki 1983
1983, Helsinki 1984
1984, Helsinki 1985
1985, Helsinki 1986
1986, Helsinki 1987
1987, Helsinki 1988
1988–1989, Helsinki 1990
1990, Helsinki 1991
1991, Helsinki 1992
1992, Helsinki 1993
1993, Helsinki 1994
1994, Helsinki 1995
1995, Helsinki 1996
1996, Helsinki 1997
1997, Helsinki 1998
1998, Helsinki 1999
1999–2000, Helsinki 2000
2001, Helsinki 2002
2002, Helsinki 2003
2003, Helsinki 2004
2004, Helsinki 2005
2005, Helsinki 2006
2006, Helsinki 2007
2007, Helsinki 2008
2008, Helsinki 2009
2009, Helsinki 2010
2010, Helsinki 2011
2011–2012, Helsinki 2012
2013–2014, Helsinki 2014

Tilastoseuran julkaisuja voi tiedustella sihteeriltä sähköpostitse osoitteesta suomentilastoseura@gmail.com.
Joidenkin julkaisujen painokset ovat tosin jo loppuneet.

Muita julkaisuja

Andra publikationer

Other publications

Suomen Tilastoseura 1920–1945, Helsinki 1946

Statistiska Sammanfundet i Finland 1920–1945, Helsingfors 1946

Pohjoismainen tilastosanasto – Nordisk statistisk nomenklatur, Kööpenhamina 1954

13:e Nordiska statistikermötet i Helsingfors 14–16 juni 1973, Jyväskylä 1974

The 13th Joint Meeting of the Nordic Statistical Societies in Helsinki June 1973, Jyväskylä 1974

Det 18:e nordiska statistikmötet i Esbo, Hundraårsjubileum, Helsingfors 1990

The Joint Conference of the Nordic Statisticians in Espoo, Finland 1989, Helsinki 1990

Scandinavian Journal of Statistics
Theory and Applications

The Scandinavian Journal of Statistics (SJS) is an international statistical journal which welcomes contributions from all countries. The language is English.

The Main purpose of the journal is to publish research papers in theoretical and applied statistics. It also welcomes statistically motivated papers on relevant aspects of probability and other fields, as well as papers on innovative applications of statistical methodology.

Scandinavian Journal of Statistics is published under the auspices of
the Danish Society for Theoretical Statistics
the Finnish Statistical Society
the Norwegian Statistical Society
the Swedish Statistical Association

Scandinavian Journal of Statistics is published quarterly in March, June, September and December by Blackwell Publishers, 108, Cowley Road, Oxford OX4, 1JF, UK or 238 Main Street, Cambridge, MA 02142, USA.

Myönnetyt Leo Törnqvist -palkinnot

1978 **Rene Tigerstedt, Helsingin yliopisto.** En modell för valbeteende i trafiken.

1979 **Pirkko Kirjavainen, Turun kauppakorkeakoulu.** Mallin rakentaminen ja ennusteen laatiminen Suomen sähkön kulutukselle kahta aikasarja-analyysimenetelmää käyttäen.

1980 **Esa Läärä, Helsingin yliopisto.** Ikä-, aika- ja kohorttitekijöiden vaikutukset Suomen miesten keuhkosyöpäsairastavuudessa vuosina 1953–76.

1981 **Arvi Suvanto, Tampereen yliopisto.** Kausivaihtelu aikasarjamalleissa.

1982 **Maija Salo, Helsingin yliopisto.** Yritys prioritiedon käytöstä alkoholi-juomien kulutusta selittävän kysyntämallin tukena. Jamel Boucelham, Jyväskylän yliopisto: Tunnustuspalkinto.

1983 **Vesa Vihriälä, Helsingin yliopisto.** Aikasarjojen välisen riippuvuuden mittaus ja testaus: sovellus suomalaisiin rahatalouden sarjoihin.Pirkko Welin, Tampereen yliopisto: Tunnustuspalkinto.

1984 **Jari Palsio, Turun kauppakorkeakoulu.** Skenaarioiden rakentaminen ristivaikutusanalyysimallia käyttäen.

1985 **Kenneth Nordström, Helsingin yliopisto.** Gauss-Markov-mallien erikoisongelmista.

1986 **Tapio Nummi, Tampereen yliopisto.** APL-pohjainen ohjelmisto GMANOVA-mallille.

1987 **Ari Veijanen, Helsingin yliopisto.** Pickardin kentän soveltamisesta kuva-analyysissä.Kari Nissinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1988 **Jaason Haapakoski, Helsingin yliopisto.** Binomijakautuneiden muuttujien muutospisteongelma.

1989 **Pasi Korhonen, Helsingin yliopisto.** Kemometrian tilastollisista menetelmistä.

1990 **Päivi Partanen, Jyväskylän yliopisto.** Suljetun populaation koon estimointi merkintä-takaisinpyynti-menetelmällä: log-lineaarinen lähestymistapa. Markku Nurhonen, Tampereen yliopisto: Tunnustuspalkinto.

1991 **Elina Järvinen, Helsingin yliopisto.** Rajoitettujen, stokastisten ja konveksien estimaattoreiden käytöstä polynomisen viipymämallin parametrien estimoinnisssa simulointikokeiden valossa.

1992 **Jouni Kuha, Helsingin yliopisto.** Binääristen regressiomallien selittäjien mittausvirheet ja parametriestimaattien mittausvirhekorjaukset. Juha Heikkinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1993 Palkintoa ei jaettu (yhtään ehdotusta ei saatu).

1994 **Ilkka Taskinen, Jyväskylän yliopisto.** Äärelliset Markovin ketjut ja annelointi.

1995 **Mika Rautakorpi, Teknillinen korkeakoulu.** Application of Markov chain techniques in certification of software. Tuija Jäppilä, Jyväskylän yliopisto: Tunnustuspalkinto.

1996 **Veli-Matti Suppola, Jyväskylän yliopisto.** Robustit menetelmät. Jakaumien vinouden vaikutuksesta korrelaatiomatriisin estimointiin.

1997 **Albert Höglund, Teknillinen korkeakoulu.** An Anomaly Detection System for Computer Networks.

1998 **Samuli Visuri, Oulun yliopisto.** Robustista kovarianssimatriisin esti-moinnista ja sen sovelluksista signaalinkäsittelyssä.

1999 **Jani Raitanen, Tampereen yliopisto.** Jalkapallo-ottelun lopputuloksen tilastollinen mallintaminen.

2000 **Reijo Sund, Helsingin yliopisto.** Tilastollisia menetelmiä dynaamisten potilaspopulaatioiden mallintamiseen. Tapahtumahistoria-analyysia hoitoilmoitusrekisterin skitsofreenikoille.

2001 **Samu Mäntyniemi, Oulun yliopisto.** A Hierarchial Bayes Model for Assessing Salmon (Salmo salar L.) Parr and Smolt Populations.

2002 **Ilmari Juutilainen, Oulun yliopisto.** Teräslevyjen lujuuden ennustaminen regressio- ja neuroverkkomalleilla.

2003 **Leena Kalliovirta, Helsingin yliopisto.** Mar-malli.

2004 **Mikko Myrskylä, Jyväskylän yliopisto.** Estimation of Class Frequencies with Micro Level Auxiliary Information.

2005 **Antti Liski, Tampereen yliopisto.** Lonkkamurtumapotilaiden hoitokustannusten vertailu vastaavuuspistemäärään perustuvalla menetelmällä.

2006 **Karri Seppä, Oulun yliopisto.** Suomalaisten paksusuolisyöpäpotilaiden ennusteen analyysi suhteellisen elossapysymisen ja syykohtaisen kuolleisuuden malleilla käyttämällä suurimman uskottavuuden ja Bayesin menetelmiä. ja Jukka Siren, Helsingin yliopisto. Populaatioiden geneettisen rakenteen spatiaalinen mallintaminen.

2007 **Outi Ahti-Miettinen, Helsingin yliopisto.** Kaksivaiheisen potenssikiintiöinnin käyttö otoksen tehostamisessa – Esimerkkinä otoksen suunnittelu työvoimakustannusindeksin tietojen keruulle.

2008 **Paul Catani, Svenska handelshögskolan.** Enhetsrottest och initialvärdet Tillämpning på arbetslösheten i Finland

2009 **Elina Ahola, Jyväskylän yliopisto.** Eksponenttisen perheen tila-avaruusmallien sovellus alkoholikuolleisuusaineistoon Matias Leppisaari, Aalto yliopiston teknillinen korkeakoulu: Tunnustuspalkinto.

2010 **Sanna Peltomäki, Tampereen yliopisto.** Estimation of Below Threshold Intra-EU Trade.

2011–2012
Tytti Pasanen, Tampereen yliopisto. Two-Level Structural Equation Modeling with Non-Normal Observed Variables for Assessing Poverty in Laos.

2013–2014
Joni Virta, Turun yliopisto. Some tools for linear dimension reduction.

Myönnetyt väitöskirjapalkinnot

2009–2012
Jukka Sirén, Helsingin yliopisto. Statistical models for inferring the structure and history of populations from genetic data.

ISSN 0355 – 5941