# On Exploring Hidden Structures of Cervical Cancer Incidence Data in Finland

Niko Lietzén[1⋆], Janne Pitkäniemi[2,3], Sirpa Heinävaara[2], and Pauliina Ilmonen[1]

[1] Department of Mathematics and Systems Analysis, Aalto University School of Science, Helsinki, Finland, {niko.lietzen, pauliina.ilmonen}@aalto.fi
[2] Institute for Statistical and Epidemiological Cancer Research, Finnish Cancer Registry, Helsinki, Finland, {janne.pitkaniemi, sirpa.heinavaara}@cancer.fi
[3] Department of Public Health, University of Helsinki, Finland

**Abstract.** Finding novel latent structures in time related variation in disease incidence may reveal new etiological components and is of great interest in disease epidemiology. We introduce time series version of invariant coordinate selection (tICS) as an exploratory tool in the search of hidden structures in the analysis of population-based registry data. Increasing cancer burden inspired us to consider a case study of age stratified cervical cancer incidence time series data in Finland between the years 1953 and 2014. To our knowledge, tICS has not been previously applied to this data. The latent components, that we uncovered using tICS, support recent findings of calendar time varying contribution of early and late age related components in cervical cancer. Furthermore, we could explain most of the variation of cervical cancer incidence in different age groups by using only two latent tICS components. The second tICS component is particularly interesting, since it clusters the age groups according to the age of menopause. Our case study results imply that the etiology behind cervical cancer differs by age group.

## 1 Introduction

Increasing cancer burden has made researchers worldwide search for factors that explain trends in cancer incidence data [19]. In addition to age, period, and cohort, several other observable factors have an effect on the trends in cancer incidence. Improved diagnostics, organized cancer screening programs, and general awareness have increased the cancer incidence for many cancers. Additionally, some known lifestyle related factors have an effect on the cancer incidence. For example, changes in smoking prevalence have a clear delayed effect on the incidence of lung cancer. However, there are also unknown underlying factors that have effects on cancer incidence rates. Identification and quantification of those unknown factors would further help in understanding the trends in cancer incidence data.

In this paper, we consider a time series version of invariant coordinate selection (ICS) in the context of latent components of calendar time variation in

⋆ Work of N. Lietzén was supported by Emil Aaltonen Foundation (grant 170156 N).

incidence. ICS has been previously applied in numerous medical applications. It has been applied successfully to analyze e.g. EEG data sets of the brain [12] and to cluster mammogram data sets [6]. We use the transformation as an exploratory tool in search of hidden structures of cervical cancer incidence time series data. The data set is from Finland between the years 1953 and 2014 and it is available online [5].

ICS is closely related to the more famous independent component analysis (ICA). Under some assumptions, the ICS procedure provides a solution to the independent component problem. The objective in ICS is to transform the observed data into an invariant coordinate system. Occasionally, the new coordinate system reveals structure from the data that is not originally visible. The clear advantage of ICS when comparing to e.g. the frequently used principal component analysis (PCA) is that the chosen scales and units of measurement have no effect on the results.

## 2 Invariant Coordinate Selection

In this section, we review the scatter matrix based invariant coordinate selection method introduced in [24].

Let $X \in \mathbb{R}^{n\times p}$, where $n > p$. A location functional is a $p$-vector valued statistic $\hat{T}(X)$, that is affine equivariant in the sense that

$$\hat{T}\left(AX + b1_n^\top\right) = A\hat{T}(X) + b$$

for all nonsingular $p \times p$ matrices $A$ and for all $p$ vectors $b$. Likewise, a positive definite $p \times p$ matrix valued statistic $\hat{S}(X)$ is a scatter matrix, that is affine equivariant in the sense that

$$\hat{S}\left(AX + b1_n^\top\right) = A\hat{S}(X)A^\top$$

for all nonsingular $p \times p$ matrices $A$ and for all $p$ vectors $b$.

Elementary examples of a location functional and a scatter matrix are the sample mean vector and the sample covariance matrix. There are several other location functionals and scatter matrices, even families of them, that have different desirable properties, e.g. robustness, efficiency, limiting multivariate normality and computational efficiency [4, 14, 15].

Let $\hat{T}_1(X)$ denote an arbitrary but fixed location functional, and let $\hat{S}_1(X)$ and $\hat{S}_2(X)$ denote arbitrary but fixed and different scatter matrices. The invariant coordinate selection (ICS) transformation $\hat{\Gamma}(X)$ for the data $X$ is defined such that if

$$Z = \hat{\Gamma}(X)\left(X - \hat{T}_1(X)1_n^\top\right),$$

then

$$\hat{T}_1(Z) = 0,\ \hat{S}_1(Z) = I_p,\ \text{and}\ \hat{S}_2(Z) = L = \operatorname{diag}(l_1, ..., l_p),$$

where the diagonal elements of the diagonal matrix $L$ are in decreasing order.

If the data arises from a continuous distribution, then the transformation matrix $\hat{\Gamma}(X)$ is almost surely unique up to the signs of its row vectors. Consequently, it is affine equivariant up to the signs, and it can be used to transform the data to up to signs invariant coordinates. Thus, affine transformations to the original data have no effect on the procedure. The transformation ensures that when examining the transformed data, the findings are true findings and not artefacts of the chosen coordinate system. Note that whereas principal component analysis (PCA) makes data uncorrelated, ICS makes data independent with respect to two measures of linear dependence. ICS transformation can be seen as affine invariant PCA that, on top of first order dependencies, considers second order dependencies as well. Moreover, whereas PCA is highly affected by scaling of the variables, ICS, due to affine invariance, is not affected by scaling at all.

It can be shown that if the chosen location and scatter estimates converge, so do the statistics $\hat{\Gamma}$ and $L$. Moreover, under the assumption of asymptotic normality of the location and scatter estimates, the statistics $\hat{\Gamma}$ and $L$ are also asymptotically normal [9, 10, 11].

The scatter matrix based invariant coordinate selection transformation was first introduced in the context of independent component analysis (ICA) [2]. It was based on the use of the regular covariance matrix and the scatter matrix based on fourth moments. The transformation was named the fourth-order blind identification (FOBI) transformation. Later, the ICS transformation was considered in wider settings [24]. In the independent component model, the elements of a $p$ variate random vector are assumed to be linear combinations of the elements of an unobservable $p$ variate vector with mutually independent components. The aim in ICA is to recover the independent components by estimating an unmixing matrix that transforms the observed $p$ variate vector to independent components [8]. Assuming that the chosen scatter functionals have the independence property, the ICS functional $\hat{\Gamma}$ provides a solution for the ICA problem [2, 10, 17, 18]. Under the assumption of i.i.d. observations, the use of the scatter matrix based ICS transformation has not been limited to ICA. It has been applied in finding hidden underlying structures of data, in constructing affine invariant depth functions, in dimension reduction, in analysing mixture models, and in defining multivariate skewness and kurtosis measures [9, 10, 11, 21, 24].

For time series data, we can obtain transformations similar to ICS, by replacing the second scatter matrix in the transformation by an autocovariance matrix. Depending on the data set, we could also use two autocovariance matrices with different lags. Autocovariance matrix based transformations have been applied to time series ICA in settings where the observations are uncorrelated second order stationary time series [16, 23]. In the context of second order stationary time series data, the procedure is called the algorithm for multiple unknown signals extraction (AMUSE) [23]. Like the scatter matrix based ICA was extended to ICS, we consider applying AMUSE transformation in wider settings. We use

it in dimension reduction and as an exploratory tool in the search of hidden structures in our case study of cancer incidence time series data.

## 3 Invariant Coordinate Selection for Time Series Data

In this section, we consider autocovariance matrix based transformations that have previously been applied to uncorrelated second order stationary time series data [16, 23].

Let $X \in \mathbb{R}^{n\times p}$ be a $p$-variate discrete times series that contains $n$ observations and denote the $i$th row of $X$ by $x_i$. Furthermore, let $\tau \in \{0, 1, \ldots, n-1\}$ and let $\hat{T}(X), \hat{S}_0(X)$, and $\hat{S}_\tau(X)$ denote the sample mean vector, the sample covariance matrix, and the sample autocovariance matrix with lag $\tau$, respectively, that are computed from the data $X$. The sample autocovariance matrix is defined as

$$\hat{S}_\tau(X) = \frac{1}{(n-\tau)} \sum_{j=1}^{n-\tau} \left( \left( x_j - \hat{T}(X) \right) \left( x_{j+\tau} - \hat{T}(X) \right)^\top \right),$$

where the sample covariance matrix is obtained with $\tau = 0$, up to a constant.

Note that $\hat{S}_\tau$ does not necessarily produce symmetric estimates. The symmetrized version of the autocovariance matrix is obtained by

$$\hat{S}_\tau^{\mathrm{S}} = \frac{1}{2} \left( \hat{S}_\tau + \hat{S}_\tau^\top \right),$$

which is a more convenient estimator since the population quantities of scatter are usually assumed to be symmetric.

The time series invariant coordinate selection transformation matrix, i.e. the unmixing matrix, $\hat{\Gamma}(X)$ for the data $X$ is now defined such that if

$$Z = \hat{\Gamma}(X) \left( X - \hat{T}(X) 1_n^\top \right),$$

then

$$\hat{T}(Z) = 0 \quad , \quad \hat{S}_0(Z) = I_p \quad \text{and} \quad \hat{S}_\tau^{\mathrm{S}}(Z) = \Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_p),$$

where $|\lambda_1| \geq ... \geq |\lambda_p| > 0$.

For general time series data, we call the transformation time series ICS or shortly tICS. The tICS transformation transforms time series data to invariant coordinates and it may be used in dimension reduction and/or as an exploratory tool in the search of hidden structures from time series data. We can think that ICS is an extension to PCA and tICS is incorporating the time series structure to the ICS transformation.

The efficiency of the tICS procedure depends strongly on the choice of the parameter $\tau$. The approach proposed in literature is to try different values of $\tau$ and choose the parameter such that the estimate $\Lambda$ has as distinct diagonal elements as possible [3].

Note that,

$$\hat{S}_0(X)^{-1}\hat{S}_\tau^{\mathrm{S}}(X)\hat{\Gamma}(X)^\top = \hat{\Gamma}(X)^\top \Lambda,$$

i.e. like in the case of regular ICS, the diagonal elements of $\Lambda$ are the eigenvalues of $\hat{S}_0(X)^{-1}\hat{S}_\tau^{\mathrm{S}}(X)$, and the column vectors of $\hat{\Gamma}(X)^\top$ are the corresponding eigenvectors. And again, if the diagonal elements of $\Lambda$ are distinct, then the solution is unique up to the signs of the eigenvectors. If the data arises from a continuous distribution, then the transformation matrix $\hat{\Gamma}(X)$ is almost surely unique up to the signs of its row vectors. Consequently, it is affine equivariant up to the signs, and it can be used to transform time series data to invariant coordinates.

After deriving the estimate $\hat{\Gamma}(X) = \hat{\Gamma} \in \mathbb{R}^{p\times p}$, the observed centered curves can be estimated using the inverse $\hat{\Gamma}^{-1}$ such that

$$\hat{x}_t(i) = \sum_{k=1}^{q} \left[\hat{\Gamma}^{-1}\right]_{ik} z_t(k),\ t \in 1, ...., n \tag{1}$$

where $i \in \{1, \dots p\}$ is the $i$th column of $X$, $q$ is the number of underlying components we use to estimate the original time series, $z_t(k)$, $k \in \{1, \dots p\}$, is the $k$th column of the transformed data, i.e. the $k$:th estimated tICS component at time point $t$, and $\hat{x}_t(i)$ is the resulting estimate of the observed $i$th time series at a point of time $t$. Note that if we use all the tICS components, i.e. choose $q = p$, the estimates are then exactly the centered version of the original time series.

## 4 Underlying Trends in Cancer Incidence

In this section, we apply time series ICS transformation to time series data of age stratified cervical cancer incidence rates between years 1953 and 2014 in Finland. These long-term cervical cancer data from 1953–2014 in Finland were obtained from the population-based cancer registry with excellent quality and coverage of registration of solid tumors [13, 22]. The data is available on the web page of the NORDCAN project [5].

During this 60–year period, cervical cancer incidence has been affected mostly by the nationwide screening program [7]. The organized cervical cancer screening program was introduced in 1963 and it reached full coverage in the early 70s. Screening invitations are sent to females between 30 and 60 years of age in every five years. In some municipalities, invitations are extended also to females aged 25 and/or 65 years. Thus, we have divided the data into five year age groups. There have been clear changes, mostly in the years and the age groups subject to organized screening. We search for underlying structures that can be used in describing other changes in cancer incidences over the years.

The age groups of younger than 35 have been combined into a single group. Likewise, age groups of older than 74 have been combined into a single group. Thus, our data set contains 10 separate age groups, resulting in a 10 dimensional

time series with 62 observations. Annual incidences of the different age groups are presented in Fig. 1. Furthermore, the sample mean time series of the cervical cancer incidence is presented as a black curve in Fig. 1.

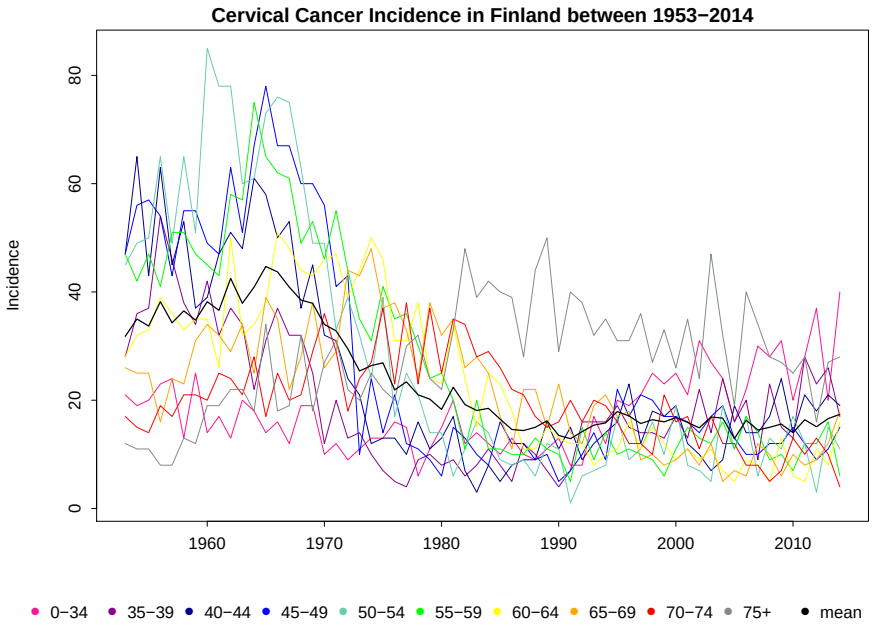
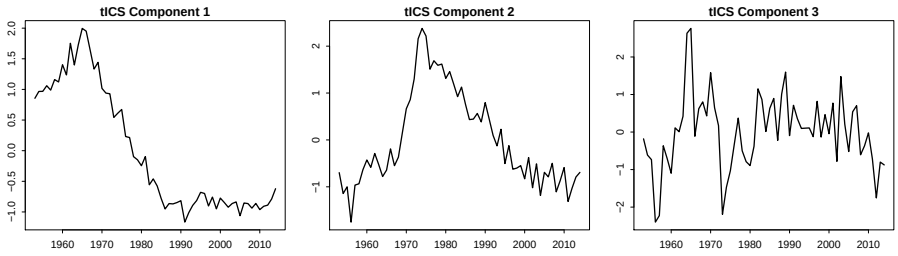



**Fig. 1.** Age stratified cervical cancer incidence in Finland between 1953 and 2014.

We performed the tICS transformation using the sample covariance matrix as the first scatter matrix and the sample autocovariance matrix with lag parameter $\tau = 1$ as the second scatter matrix. The first three estimated tICS components are presented in Fig. 2. We want to emphasize that the scales of the tICS components are not relevant. Instead, we seek for curves that have interesting shapes. The first three components have the largest corresponding absolute diagonal values on the estimated matrix $\Lambda$ and thus they are the most important. The remaining seven tICS components reveal no interesting structure and resemble noise. The remaining seven components are omitted from this paper.

The shape of the first component is similar to the mean curve time series, compare the black curve in Fig. 1 and the curve in Fig. 2a. We name the first component as "the average". The first component represents the average cervical cancel incidence.

The shape of the second tICS component is the most interesting, see Fig. 2b. It represents increasing trend from 1953 until mid–70s, and a decreasing trend after that. We call the second component "the turning point".




(a) First tICS comp (b) Second tICS comp. (c) Third tICS comp.

**Fig. 2.**

The third tICS component in Fig. 2c is less interesting when compared to the first tICS components. Like the last seven components, the third component has a great deal of resemblance to random variation.

The incidence curves in different age groups can be roughly estimated using only the first two components, see Eq. 1 and the green curves in Fig. 3. Most of the estimated incidence curves are relatively close to the observed curves. The age group of 0–34 has the least observations. Thus, random variation has a larger effect in this age group, which could be the reason for the estimate being worse when compared to the other age groups.

In order to visually observe cluster structures, the scores of the components, i.e. the curves

$$c_{ik} = \left[\hat{\Gamma}^{-1}\right]_{ik} z_t\left(k\right), \quad k \in \{1,2,3\}, \; i \in \{1,\ldots,p\},$$

are presented in Fig. 4, where $p = 10$ is the number of age groups in our case study. We refer to the value $\left[\hat{\Gamma}^{-1}\right]_{ik}$ as the loading related to the age group $i$ and tICS component $k$. If the absolute value of the loading is large, that specific tICS component has a high impact in explaining the variation of the specific age group. The curves with the highest absolute loadings are the top and bottom curves in Fig. 4. Likewise, low absolute loading values indicate that the specific tICS component has a low impact in explaining the variation of the specific age group. The curves with low absolute loadings are the middle curves in Fig. 4. The curves $c_{ik}$ provide visual clustering based on the first three components.

The first set of curves, $c_{\cdot 1}$, are ordered based on the trend of cancer incidence in a specific age group. The curves representing age groups, where the incidence has been decreasing, i.e. the behavior is similar to the mean curve, have a positive loading in this component. The largest positive loading is for the age group 50–54, which is the age group where the incidence has decreased the most. The age groups that have a negative loading with respect to the first component, have the first tICS component mirrored in Fig. 4. The age group of older than 75 has the largest negative loading with respect to the first component. The behavior

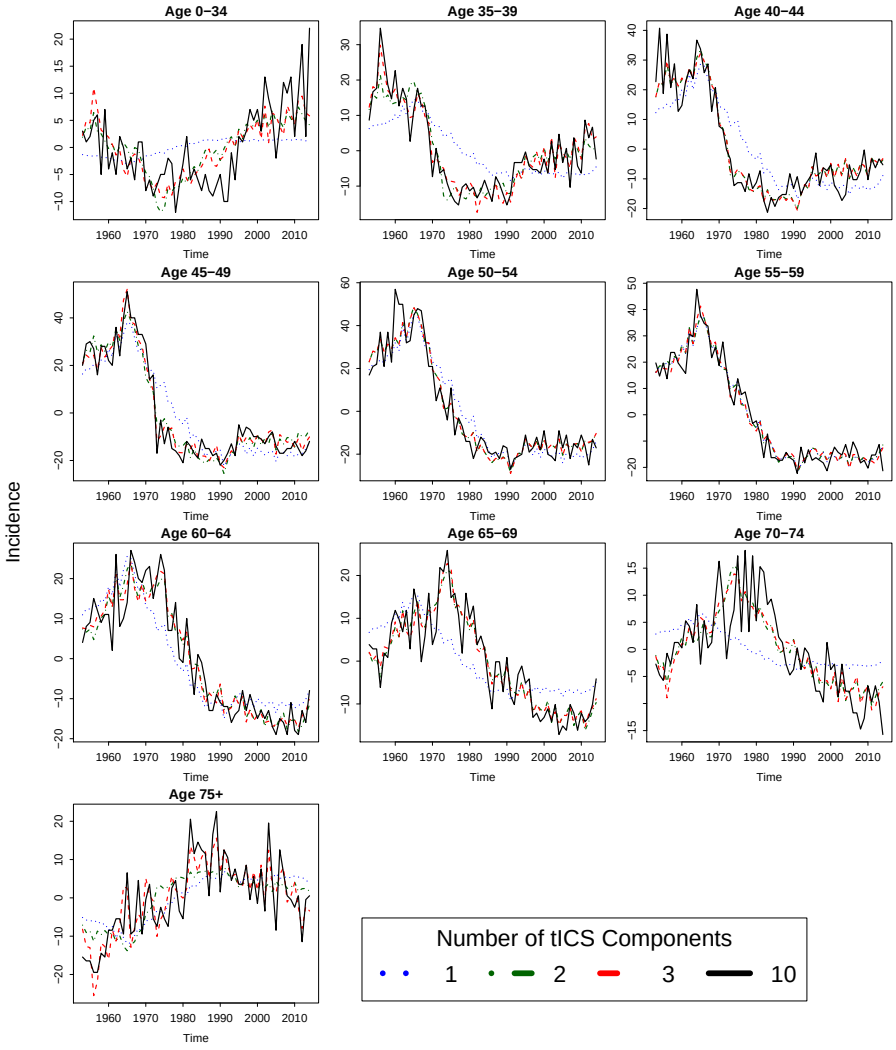



**Fig. 3.** Cervical cancer incidence in Finland between 1953 and 2014 in terms of estimated components. The estimation is performed using Eq. 1.

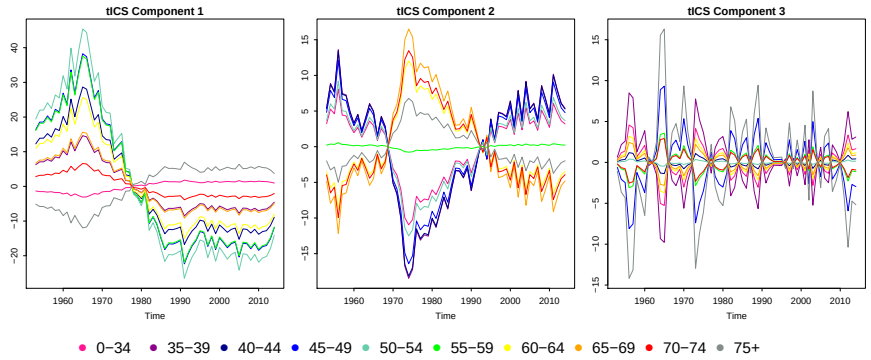



**Fig. 4.** Clustered age stratified cervical cancer incidences for the first three tICS components. The curves are the tICS components multiplied with the corresponding loadings.

of the incidence in this age group is the opposite compared to the mean cancer incidence, see Fig. 1.

The second set of curves, $c_{\cdot 2}$, provide clustering based on the second component in Fig. 4. Visual clustering reveals that the second component splits the age groups according to the age of menopause. Age groups of older than 60 have a positive loading, the age group of 55–59 has a loading close to zero and the age groups younger than 50 have a negative loading.

Visual clustering based on the remaining components, including the third tICS component in Fig. 4, reveal nothing interesting, as was expected from random variation.

## 5 Discussion

In our case study of cervical cancer incidence in Finland, tICS procedure produced interesting findings. The underlying structures found by tICS procedure support recent findings discovered using other methods [20]. The second tICS component was the most interesting one. The first component clustered the age groups with respect to trend. It separated the age groups where cancer incidence has been decreasing from those where the incidence has been increasing or has stayed relatively same. The information provided by the clustering of the first component could also be easily be verified from Fig. 1 and thus the clustering provided by this component is not particularly interesting. The components after the second one seemed to be random variation, i.e. uninteresting noise. The tICS components 4–10 were omitted from this paper for this reason.

The second tICS component clusters the age groups according to the approximate age of menopause. The behavior of this component supports the findings by Seppä et al. [20] of the calendar time varying contribution of early and late age related components in cervical cancer. The first crossing point of the curves

is soon after starting the cervical cancer screening. Hidden structures in the incidence in age groups close to menopause are different from those in the age groups far away from menopause. Thus, it seems that hormonal changes might have a significant role in etiology of cervical cancer.

We performed the tICS procedure with several different lags parameters $\tau$, and the best separation was obtained using lag $\tau = 1$. The corresponding lag produced the most distinct values for the diagonal elements of the estimated matrix $\Lambda$ and the best estimated curves in Fig. 3. However, the most interesting findings — the cluster structures visible in Fig. 4 — stayed almost identical with several different lags that were close to $\tau = 1$.

Furthermore, we applied the second order blind identification (SOBI) [1] procedure for this data. In SOBI, the second diagonalization is replaced with a joint diagonalization with respect to multiple autocovariance matrices with distinct lags. This makes the choice of the lag parameter less decisive. The shapes of the first two components were similar to our findings. However, the first two SOBI components, with every set of lags that we tried, had considerably worse performance in explaining the variation using only two or three components, whereas the first two tICS components explain the variation of the original time series relatively well, see Fig. 3.

Since the tICS procedure is affine invariant, it ensures that the findings are not simply artifacts of the used coordinate system. Exploratory tools such as PCA are not affine invariant. Affine transformations of the original data would yield completely different results in PCA whereas tICS would remain unaffected. Furthermore, we applied the PCA transformation to this data set and the first two principal components were similar to the first two tICS components. However again, visual inspection of Fig. 3 revealed that the first two and three principal components did not perform as well in estimating the original time series.

In this paper, we stratified the analysis according to age groups, but other types of stratifications are also possible. We could e.g. consider stratification according to cohort and for other cancer sites also to gender. In the Western countries we are facing increasing cancer burden [19]. Risk factors of specific cancer have been studied for a long time, and many of them have been established. Even then, attributable fraction to known risk factors is often quite low leaving us with the need of further understanding etiology of a studied cancer. Identification of latent components of cancer incidence may open new possibilities for this. Further, if age related latent components of cancer are modifiable, it is important for future efforts of reducing cancer burden in Finland.

# Bibliography


[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444, 1997.

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

[3] A. Cichocki and S. Amari. *Adaptive blind signal and image processing: learning algorithms and applications*, volume 1. John Wiley & Sons, 2002.

[4] L. Davies. Asymptotic behavior of s-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15:1269–1292, 1987.

[5] G. Engholm, J. Ferlay, N. Christensen, A. Kejs, R. Hertzum-Larsen, T. Johannesen, S. Khan, M. Leinonen, et al. Nordcan: Cancer incidence, mortality, prevalence and survival in the nordic countries, version 7.3 (8.7.2016). *Association of the Nordic Cancer Registries. Danish Cancer Society. Available from http://www.ancr.nu, accessed on 4.9.2017*, 2016.

[6] R. Gallardo-Caballero, C. García-Orellana, A. García-Manso, H. González-Velasco, and M. Macías-Macías. Independent component analysis to detect clustered microcalcification breast cancers. *The Scientific World Journal*, 2012, 2012.

[7] M. Hakama, U. Joutsenlahti, A. Virtanen, and U. Räsänen-Virtanen. Mass screenings for cervical cancer in finland 1963-71. organization, extent, and epidemiological implications. *Annals of clinical research*, 7(2):101–111, 1975.

[8] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, New York, 2001.

[9] P. Ilmonen. On asymptotic properties of the scatter matrix based estimates for complex valued independent component analysi. *Statistics and Probability Letters*, 83:1219–1226, 2013.

[10] P. Ilmonen, J. Nevalainen, and H. Oja. Characteristics of multivariate distributions and the invariant coordinate system. *Statistics and Probability Letters*, 80(23-24):1844–1853, 2010.

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

[12] C. A. Joyce, I. F. Gorodnitsky, and M. Kutas. Automatic removal of eye movement and blink artifacts from eeg data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004.

[13] M. K. Leinonen, M. Rantanen, J. Pitkäniemi, and N. Malila. Coverage and accuracy of myeloproliferative and myelodysplastic neoplasms in the finnish cancer registry. *Acta Oncologica*, 0(0):1–5, 2016.

[14] H. P. Lopuhaä. On relations between S-estimators and M-estimators of multivariate location and scatter. *Annals of Statistics*, 17:1662–1683, 1989.

[15] R. A. Maronna, R. D. Mardin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Chichester, 2006.

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

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

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

[19] D. M. Parkin, L. Boyd, and L. Walker. 16. the fraction of cancer attributable to lifestyle and environmental factors in the uk in 2010. *British journal of cancer*, 105:S77–S81, 2011.

[20] K. Seppä, J. Pitkäniemi, N. Malila, and M. Hakama. Age-related incidence of cervical cancer supports two aetiological components: a population-based register study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 123(5):772–778, 2016.

[21] R. J. Serfling. Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, 22:915–936, 2010.

[22] L. Teppo, E. Pukkala, and M. Lehtonen. Data quality and quality control of a population-based cancer registry: experience in finland. *Acta oncologica*, 33(4):365–369, 1994.

[23] L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 1784–1787, 1990.

[24] D. E. Tyler, F. Chritchley, L. Dümbgen, and H. Oja. Invariant coordinate selection. *Journal of Royal Statistical Society Series B*, 71:549–592, 2009.