PRINCIPAL COMPONENT ANALYSIS OF MOBILITY DATA FROM AN OPERATIONAL GPRS NETWORK

Charlotte Dumard, Fabio Ricciato and Thomas Zemen

ftw. Forschungszentrum Telekommunikation Wien
Donau-City-Str. 1/3 A-1220 Wien, Austria
email: {dumard, ricciato, zemen}@ftw.at

ABSTRACT

We present a preliminary analysis of mobility data collected from an operational GPRS network. The input data are time-series counting the number of mobile stations present in each of 126 sample routing areas at equally spaced instants (5 min) during one full week. The time-series were extracted from packet-level traces captured by passively monitoring a subset of the Gb links of the network of mobilkom austria AG & Co KG during October 2004. We apply the principal component analysis (PCA) to this dataset. The PCA offers a simple method for classifying the routing areas into two main groups, residential and business areas, plus a few "atypical" ones. Additionally, we address the problem of robustness of the PCA to temporary local gaps in the input data.

1. INTRODUCTION

Third-generation (3G) cellular networks have extended the reach of internet connectivity with nation-wide wireless coverage. At present the 3G environment is under evolution: the subscriber population and the traffic volume are still in a growing phase (for some details on the network under study see [1]); the relative distribution of terminal types (e.g. laptops vs. handsets) and their capabilities are changing quickly; the service portfolio and tariffs offered by the operators evolves rapidly. In such a fast-evolving environment it is of great importance to continuously monitor the status of the network and user population, and to early detect localized problems as well as global changes and performance drifts.

One important aspect to be monitored in a cellular network is the distribution in space of mobile stations (MS). In a 3G mobile network it is possible to infer the location of each MS at the granularity of Routing Areas (RA) by passively sniffing the signaling exchange between the MS and the SGSN (Serving GPRS Support Node) on the Gb and IuPS links for GPRS/EDGE and UMTS/HSDPA respectively (ref. Fig. 1). An RA is a group of neighboring radio cells (see [2, p. 129] for more details). We show in [3, Sec. 4] how to extract the discrete time-series counting the number of MS present in each RA at periodically sampled instants from from traces captured on the Gb links. The traces had been previously anonymized by eliminating any user-related field value. In this work we consider time-series with a sampling period $\tau = 5$ min and with total length of several days.

Given the large number of RAs it is desirable to reduce the dimensionality of the data at hand and at the same time classify the different RA profiles into a small number of groups. The profiles of individual RAs reflect the daily mobility pattern of the underlying user population across the country, therefore they are expectedly (i) highly correlated and (ii) pseudo-periodic with a primary cycle of 24h and a smaller one at 7 days.

In this work we explore the applicability of the principal component analysis (PCA) to reduce the size of such datasets. One point to be addressed is the robustness of the PCA to errors and gaps in the measured data. In fact, given the operational complexity of monitoring an entire 3G network, and certain technical issues in the signaling tracking procedure, it can not be always expected to obtain perfectly completed data for all RAs and for periods of several days. Temporary gaps in the data must be considered as a "normal" aspect. The contribution of this work is to present a preliminary investigation of the applicability of PCA to incomplete network data.

The PCA method was applied to network data by Lakhina et al. [4] for the analysis of origin-destination aggregate flows in an IP network. It is shown that such data has an intrinsically low dimensionality: the dataset can be very well approximated by only five dominant components. The authors also find that intermediate components might be helpful to reveal hidden anomalies (e.g. traffic spikes or dips) that are not evident in the original signals.

Fabio Ricciato was supported by the Kplus programme in the ftw. project DARWIN. The work of Charlotte Dumard and Thomas Zemen was funded by the Wiener Wissenschafts- Forschungs- und Technologiefonds (WWTF) in the ftw. project "Future Mobile Communications Systems (Math+MIMO)".

f

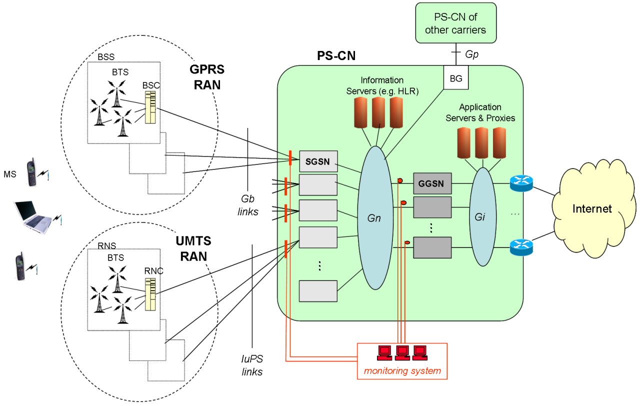

Fig. 1. 3G network structure.

In the present work we analyze a dataset that was used in a previous contribution on network optimization [3]. The dataset reports the number of MSs present in different GPRS RAs, sampled every 5 min, for a total measurement period of exactly 1 week (from Monday 00:00 to Sunday 24:00). The measurements were taken in October 2004 on the Gb links (ref. Fig. 1) of the operational GPRS network of mobilkom austria AG & Co KG, the leading mobile operator in Austria, EU, within the framework of an applied-research project on 3G traffic monitoring [1]. A number of 126 RAs are included in the METAWIN/Gb dataset, covering a fraction $X < 1$ of the total network (X non disclosed). To the best of our knowledge no previous work has reported large-scale measurements of spatial user distribution from a real operational 3G mobile network.

Notation: We denote a column vector by $\boldsymbol{a}$ and its i-th element with $a[i]$. Similarly, the columns of a matrix $\boldsymbol{A}$ are denoted through $\boldsymbol{A_i}$. The transpose of $\boldsymbol{A}$ is given by $\boldsymbol{A}^{\mathrm{T}}$. A diagonal matrix with elements $a[i]$ is written as diag($\boldsymbol{a}$) and the $Q \times Q$ identity matrix as $\boldsymbol{I}_Q$. The norm of $\boldsymbol{a}$ is denoted through $\|\boldsymbol{a}\|$.

2. THE PRINCIPAL COMPONENT ANALYSIS

In this section we summarize the theory of the principal component analysis (PCA) [5, Sec. 6.8.1]. The PCA is a coordinate transformation that maps the measured data onto a set of axes spanned by eigenvectors in order to obtain the "principal components". Each eigenvector points to the direction of maximum variation or energy (with respect to the Euclidean norm) remaining in the data, given the energy already accounted for in the preceding components. As such, the first principal component captures the total energy of the original data to the maximal degree possible on a single axis. The next principal components then captures the maximum residual energy among the remaining orthogonal directions. In this sense, the principal axes are ordered by the amount of energy in the data they capture [4].

Given a set of N_r data vectors $\boldsymbol{d}_r \in \mathbb{R}^{N_t}$ for $r \in \{1, \dots, N_r\}$, we define the data matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \dots, \boldsymbol{d}_{N_r}]$. In our case $\boldsymbol{d}_r$ represents the number of MS present in an area r sampled at time $t \in \{1, \dots, N_t\}$ after centering (zero-mean) and normalization to unit energy. We apply this normalization to prevent that areas with large user population dominate the profile of the eigenvectors. Using the singular value decomposition, we can write:

$$\boldsymbol{D} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\,, \tag{1}$$

where

$$\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_{N_r}] \in \mathbb{R}^{N_t \times N_r} \tag{2}$$

and

$$\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_{N_r}] \in \mathbb{R}^{N_r \times N_r} \tag{3}$$

are orthonormal matrices (i.e. $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_{N_r}$). $\boldsymbol{S}$ is a diagonal matrix of positive real values (the singular values) defined by $\boldsymbol{S} = \mathrm{diag}([s[1], \dots, s[N_r]]^{\mathrm{T}})$ ordered such that $s[1] \geq s[2] \geq \dots \geq s[N_r]$.

The sample covariance matrix $\boldsymbol{R_D}$ of $\boldsymbol{D}$ can be written as $\boldsymbol{R_D} = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^{\mathrm{T}}$. The square singular values of $\boldsymbol{D}$ are the eigenvalues of the covariance matrix $\boldsymbol{R_D}$. The eigenvalues represent the amount of energy in the *eigenvector* $\boldsymbol{u}_i$ (which is called eigenflows in [4]). From (1) we can write each data vector $\boldsymbol{d}_r$ as

$$\boldsymbol{d}_r = \boldsymbol{U}\boldsymbol{w}_r = \sum_{i=1}^{N_r} \boldsymbol{u}_i w_r[i]\,, \tag{4}$$

where $[\boldsymbol{w}_1, \dots, \boldsymbol{w}_{N_r}] = \boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$. Each data vector $\boldsymbol{d}_r$ can be expressed as a linear combination of the eigenvectors $\boldsymbol{u}_i$ weighted by the principal components $w_r[i]$. Additionally, $\boldsymbol{w}_r = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{d}_r$. For every data vector $\boldsymbol{d}_r$ we define the low-rank approximation vector $\tilde{\boldsymbol{d}}_r(K)$ obtained by truncating the sum in (4) to the first K components (with $K \ll N_r$):

$$\boldsymbol{d}_r \approx \tilde{\boldsymbol{d}}_r(K) = \sum_{i=1}^{K} \boldsymbol{u}_i w_r[i]\,. \tag{5}$$

For each RA r we measure the distance between the data vector and its approximation by the quadratic reconstruction error

$$z_r(K) = \frac{1}{N_t}\|\boldsymbol{d}_r - \tilde{\boldsymbol{d}}_r(K)\|^2\,. \tag{6}$$

3. RESULTS

We apply the PCA to the METAWIN/Gb dataset. Recall that each data vector $\boldsymbol{d}_r$ is one week long and has been previously centered (zero-mean) and normalized.

[1] METAWIN project home page http://userver.ftw.at/∼ricciato/metawin

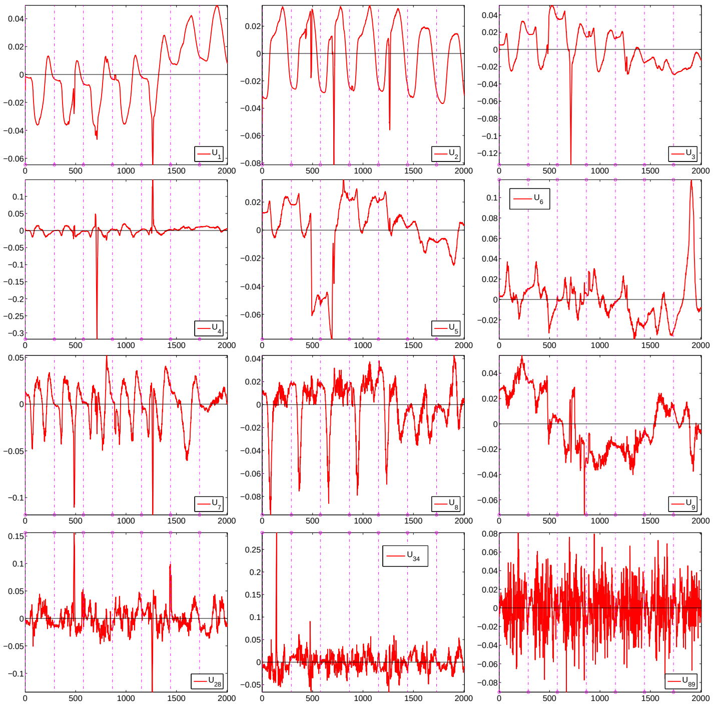

Fig. 2. Sample eigenvectors $\boldsymbol{u}_i$ for $i \in \{1-9, 28, 34, 89\}$.

3.1. Eigenvectors and Eigenvalues

Some representative eigenvectors of the dataset $\boldsymbol{D}$ are shown in Fig. 2. As expected the first two eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ expose a strong daily (pseudo)-periodicity. The profile in the weekdays (Mon-Fri) is markedly different from the weekend (last two days). The dotted vertical lines mark midnight for each day.

The large dips are caused by short-term lack of data affecting multiple RAs at the same time. It can be seen that this lack of data is already captured in the very first eigenvectors. Similarly the large gap in the fifth eigenvector $\boldsymbol{u}_5$ is due to a temporary gap in the data between samples 487 and 695, affecting multiple RAs. As expected, higher index eigenvectors become more and more noisy (and have higher frequency content). Note that the variability range is modulated by a 24h periodic envelope (e.g. $\boldsymbol{u}_{89}$), since the absolute value of the signal variance increases with the signal level.

In Fig. 3-left we plot the distribution of the eigenvalues (square singular values) in loglog scale. The first two eigenvalues are rather close to each other. From the 2nd eigenvalue onwards the distribution fits very well a power-law with slope -1.25. This power-law fit is remarkable and can not be considered incidental. We do not have a full explanation for such a behavior: power-law eigenvalue spectra are known to be a feature of certain classes of random graphs (see [6] and references therein) but there is no trivial connection to our case. In some preliminary experiments with

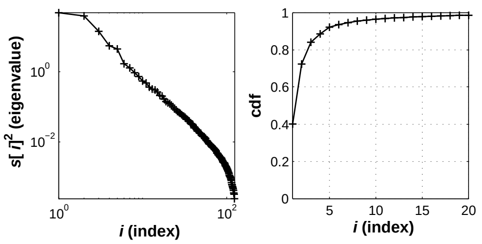

Fig. 3. Distribution of the squared singular values $s[i]^2$ in log-log scale (left). Cumulative distribution function (right)

synthetic data we noticed that power-law spectra emerge also when the dataset is built from independent combinations of sinusoids, however we leave the exploration of this point to further study.

In Fig. 3-right we plot the cumulative distribution function of eigenvalues $s[i]^2$. This corresponds to the total energy (or variability) captured by the first i eigenvectors. It can be seen that more than 95% of the energy can be captured by the first five eigenvectors only.

3.2. Reconstruction and Classification

In Fig. 4 we show some representative data vectors along with the low-rank approximation using the first two principal components ($K = 2$). The profiles of the top two graphs (RA 8, 52) are typical for a business environment: daily peaks placed at working-hours, and a considerable lower level during the week-end. On the contrary the bottom two profiles (RA 51, 62) are typical for residential areas: daily peaks are located in the evening, and the number of active users is larger on the week-end than on weekdays.

We can see that the original data is very well approximated with only $K = 2$ eigenvectors in all the vectors shown in Fig. 4. In fact, we have chosen on purpose RAs with highly “representative” profiles (i.e. recurrent across the whole dataset). It turns out that the first two eigenvectors have well captured the behavior of these two RA groups.

In Fig. 5 we report the reconstruction error $z_r(K)$ for $K = 2$ (left) and $K = 5$ (right). In each graph the threshold delimits the eight highest values of the reconstruction error.

The profile of the eight RAs with the highest reconstruction error $z_r(2)$ are reported in Fig. 6. In four out of eight cases the reconstructed signal failed in capturing the large data gap centered at sample 600. From Fig. 4 we know that the gap was not present before the fifth eigenvector. In fact the reconstructed signals $\tilde{\boldsymbol{d}}_r(K)$ with $K = 5$ (not shown here) follow very closely the original vector, as can be inferred by the lower values of the reconstruction error in the right-hand plot of Fig. 5.

In Fig. 7 we plot the first two principal components $w_r[1]$ and $w_r[2]$ for all RAs. The two subgraphs show exactly the same points but on the left (resp. right) graph the

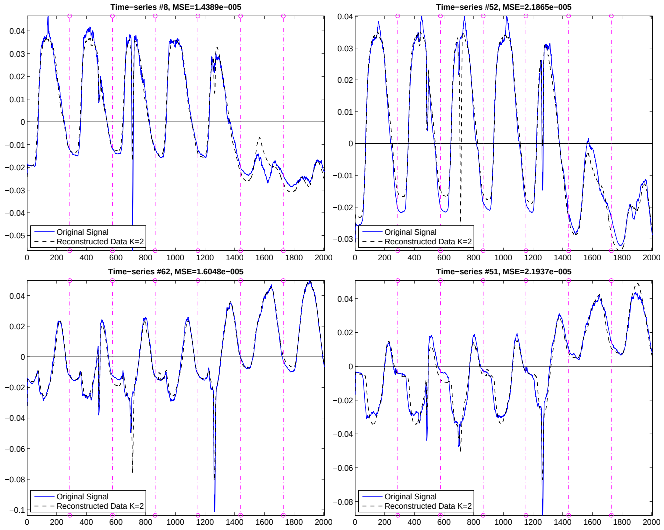

Fig. 4. Samples of well-reconstructed vectors using two principal components only ($K = 2$).

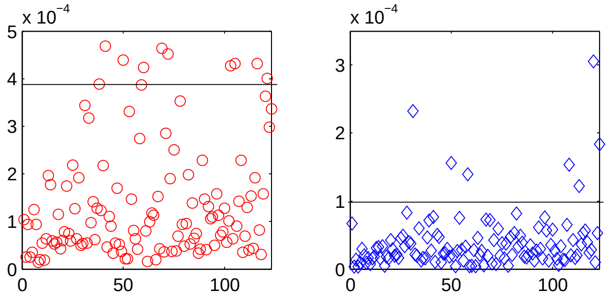

Fig. 5. Reconstruction error $z_r(K)$ for $K = 2$ (left) and $K = 5$ (right) versus RA r. The straight line shows the threshold delimiting the worst 8 values

black '+' mark the RA with the lowest (resp. highest) reconstruction error $z_r(2)$. The first observation is the presence of two large clusters that include most of the RAs. The leftmost cluster has negative values of $w_r[1]$ and $w_r[2]$ above 0.5, while the rightmost cluster has positive values of $w_r[1]$ and $w_r[2]$ between 0 and 1. The two clusters discriminate the two main types of RAs identified visually in Fig. 4: residential and business areas. The second interesting observation is that all the "best reconstructed" signals (with lowest reconstruction error, black '+' marks in the left-hand graph of Fig. 7) falls within such clusters, while all the 8 "worst reconstructed" ones lay outside the clusters.

In other words the most common ("typical") RA profiles are also well-approximated by the first two eigenvectors, while the RAs with atypical behavior are not. This result suggests that the PCA can be used for a first classification of RA profiles, discriminating the "typical" and "atypical" behavior based on the reconstruction error $z_r(2)$ of the first two eigenvectors. The "typical" can be further clustered into two groups (residential and business areas) based on the principal components $w_r[1]$ and $w_r[2]$. The few "atypical" RAs should be instead analyzed separately. For instance, one could apply again PCA on the subset of atypical RAs, seeking for further clustering and classification.

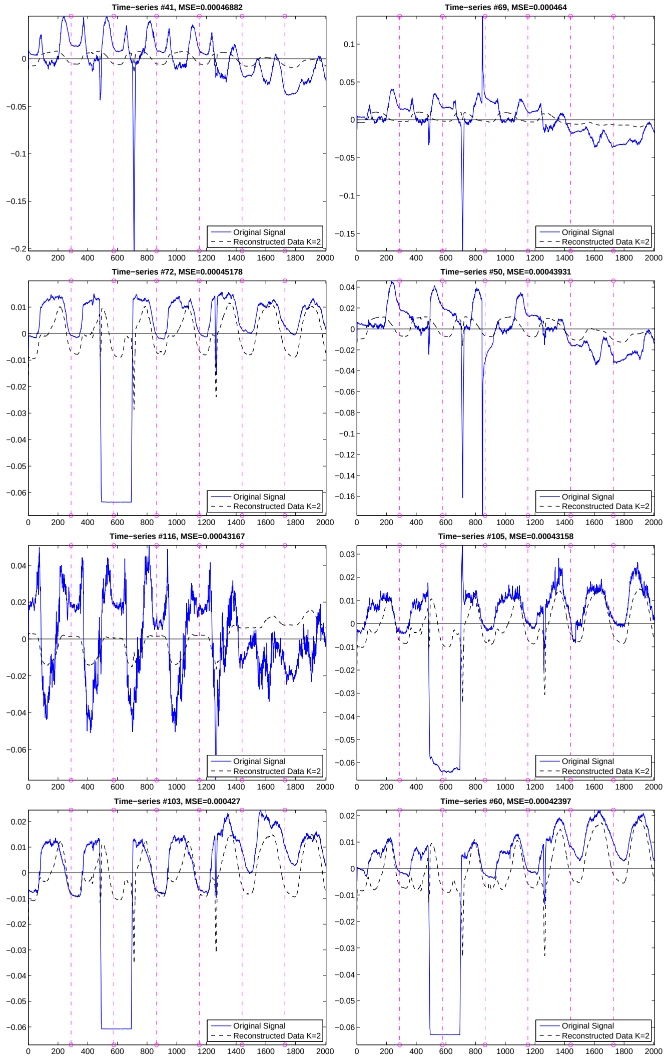

Fig. 6. Atypical time-series and associated reconstructed signals $\tilde{\boldsymbol{d}}_r(K)$ with $K = 2$.

3.3. Local Data Gaps Cause Global Distortion

The PCA decomposes the original set of data vectors into a new basis. More specifically, it selects the new basis subject to a constraint, i.e. orthonormality, and with an optimality criterium, namely the minimization of the residual reconstruction error at each step. Both the constraint and the optimality metric are defined as sums in time of inner product between vectors. That means that the profile of each eigenvector in a certain time interval is *not* independent from the

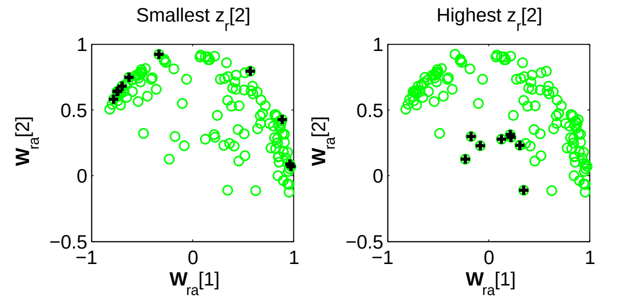

Fig. 7. Scatter-plot of the first two principal components $w_r[1]$ and $w_r[2]$. On the left (right) side the eight RA with the lowest (highest) reconstruction error are marked with +.

value of the original dataset outside that interval. As a direct consequence, a large error or gap in the data in any time interval, say $[t_1, t_2]$, will have a global effect on the whole profile of the eigenvector. In other words a local (in time) error in the data will propagate to a global effect. That puts in question the utility of applying PCA to non-ideal datasets, i.e. incomplete or partially corrupted.

To see such effects on our dataset we performed the following additional analysis. Recall that in our dataset there is a large gap in the data affecting multiple RAs between day 2 and 3 (see Fig. 6). In order to bypass it, we applied PCA to a subset of the original data where only the samples for the last 4 days were considered. We artificially create *erroneous* datasets by zeroing the samples in the time interval $[1600, 1700]$ for a fraction Y of randomly selected RAs. In Fig. 8 we report the first four eigenvectors of the original subset (dotted line) along with the ones obtained for the artificial *erroneous* data with $Y = 15\%$ (full line) and $Y = 45\%$ (dashed lines). In case of $Y = 15\%$ the (artificial) data gap impacts only the eigenvectors above rank 2. For $Y = 45\%$ also the first eigenvector is modified. Interestingly, the second eigenvector remains stable also for very high Y (e.g. 60%). These results show that an anomaly (e.g. missing data) affecting a fraction of the RAs might impact the global eigenvectors.

4. CONCLUSIONS

We applied PCA to time-series representing the number of MS in different RAs. We showed that the projection onto the first two eigenvalues reveals two large clusters corresponding to residential and business areas, leaving only a few “atypical” elements out of the clusters. We found that the whole set can be well-approximated by only five principal components, while “typical” profiles can be approximated by just the first two.

We showed that the interpretation of the eigenvectors must be handled with care since the presence of local data gaps can have a global impact on the whole eigenvector profiles. This raises some concerns about the robustness of PCA to incomplete datasets. We are currently exploring alternative decomposition methods for data reduction, seeking for better robustness properties to missing data.

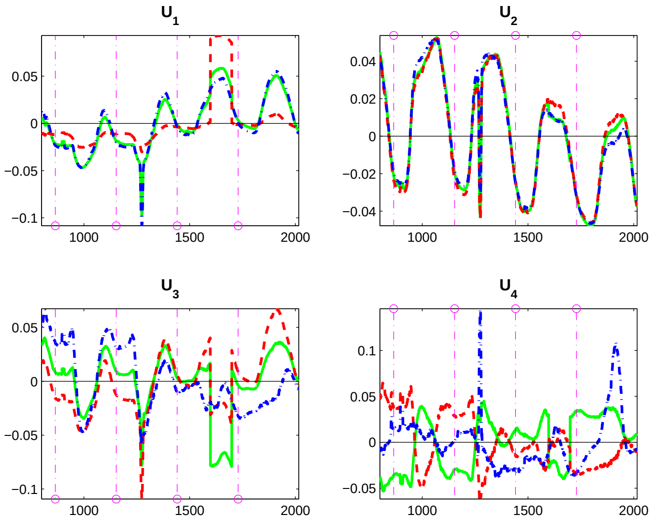

Fig. 8. First four eigenvectors for the truncated datasets for original data (—·), 15% erroneous data (–) and 45 % erroneous data (- -).

5. REFERENCES

[1] P. Svoboda, F. Ricciato, E. Hasenleithner, R. Pilz, “Composition of GPRS/UMTS traffic : snapshots from a live network,” *4th Int’l Workshop on Internet Performance, Simulation, Monitoring and Measurement, Salzburg (IPS-MOME’06)*, February 2006.

[2] J. Bannister, P. Mather, S. Coope, *Convergence Technologies for 3G Networks*, Wiley, 2004.

[3] F. Ricciato, R. Pilz, E. Hasenleithner, “Measurement-based Optimization of a 3G Core Network: a Case Study,” *6th Int’l Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN’06), St. Petersburg,*, May 2006.

[4] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, “Structural Analysis of Network Traffic Flows,” *ACM SIGCOMM*, June 2004.

[5] Todd K. Moon and Wynn C Stirling, *Mathematical methods and algorithms for signal processing*, Prentice Hall, Upper Saddle River (NJ), USA, 2000.

[6] Milena Mihail and Christos H. Papadimitriou, “On the eigenvalue power law,” in *6th International Workshop on Randomization and Approximation Techniques (RANDOM)*. 2002, vol. 2483 of *Lecture Notes In Computer Science*, pp. 254 – 262, Springer Verlag.