ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Low-dimensional decomposition, smoothing and forecasting of sparse functional data

Alexander Dokumentov, Rob J Hyndman

May 2014

Working Paper 16/14

Low-dimensional decomposition, smoothing and forecasting of sparse functional data

Alexander Dokumentov
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: alexander.dokumentov@monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: rob.hyndman@monash.edu

29 May 2014

JEL classification: C10,C14,C33

Low-dimensional decomposition, smoothing and forecasting of sparse functional data

Abstract

We propose a new generic method ROPES (Regularized Optimization for Prediction and Estimation with Sparse data) for decomposing, smoothing and forecasting two-dimensional sparse data. In some ways, ROPES is similar to Ridge Regression, the LASSO, Principal Component Analysis (PCA) and Maximum-Margin Matrix Factorisation (MMMF). Using this new approach, we propose a practical method of forecasting mortality rates, as well as a new method for interpolating and extrapolating sparse longitudinal data. We also show how to calculate prediction intervals for the resulting estimates.

Keywords: Tikhonov regularisation, Smoothing, Forecasting, Ridge regression, PCA, LASSO, Maximum-margin matrix factorisation, Mortality rates, Sparse longitudinal data

1 Introduction

In this paper we consider a number of data analysis problems involving sparse functional data, and show that each of these problems can be re-cast in the framework of the following optimization problem:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\left(\left\|W \odot (Y - UV^T)\right\|^2 + \lambda \|KU\|^2 + \theta \|LV\|^2\right), \tag{1}$$

where:

- Y is an $n \times m$ matrix of two-dimensional data;
- U is an $n \times k$ matrix of "scores", $k = \min(n, m)$;
- V is a $k \times m$ matrix of "features";
- $\lambda > 0$ and $\theta > 0$ are smoothing parameters;
- K and L are "complexity" matrices which transform multivariate "scores" U and "features" V into the corresponding "complexity" matrices;
- $\|.\|$ is the Frobenius norm;
- $\odot$ refers to element-wise matrix multiplication; and
- W is an $n \times m$ matrix of weights.

The method for obtaining solutions to problems of the form (1) we call ROPES, meaning Regularized Optimization for Prediction and Estimation with Sparse data. This is also a deliberate allusion to the LASSO (Tibshirani, 1996), which solves a slighly different problem but with obvious similarities. The problem is also closely related to Maximum-Margin Matrix Factorisation (Srebro et al., 2005).

In Section 2.1, we show that this problem can be reduced to a convex optimization problem, and in Section 2.2 we discuss how to solve ROPES numerically. In Section 3, we introduce Canonical ROPES, a special type of solution which exposes the internal structure of the data. Then, in Section 4, we show that ROPES is equivalent to maximum likelihood estimation with partially observed data. This allows the calculation of confidence and prediction intervals, as described in Section 5. Two applications are described in Sections 6 and 7, before we provide some general comments and discussion in Section 8.

In this introduction we will explain the motivation of ROPES by providing brief introductions to the two applications that we will be discussing in detail later. We will also show how the

problem in (1) is connected to other well-known statistical algorithms, principal components and ridge regression.

1.1 Motivation based on sparse longitudinal data

Sparse longitudinal data often have a functional component. It is usually assumed that some unobserved parameters involved in the data generation process have some functional properties like continuity and smoothness along the time dimension. One well-known example is the subset of the data presented in the book by Bachrach et al. (1999). This subset was discussed and used as an example for different methods by James (2010).

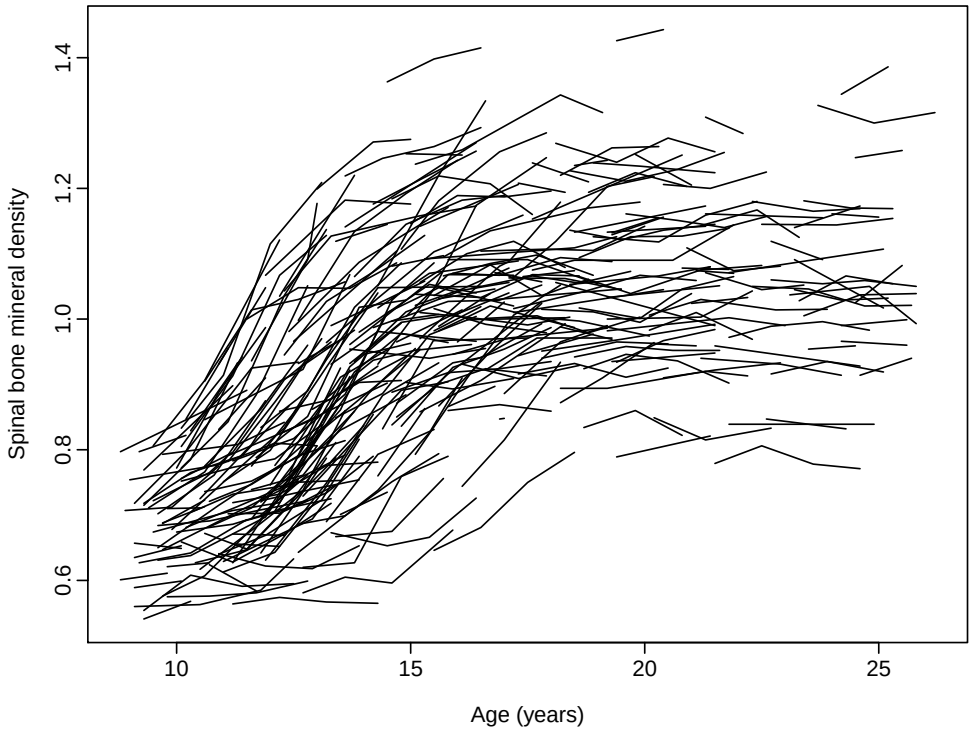

Figure 1: *Measurements of the spinal bone mineral density (g/sm^2) for 280 people.*

The dataset is shown in Figure 1 as a "spaghetti" graph; observations which relate to a particular person are represented as connected points. There are 280 different people and 860 observations in total. Every person has two to four measurements of their spinal mineral bone density (g/sm^2) taken at different periods of their lives.

The interpolation and extrapolation of such data is a difficult task, since various different shapes that vary with gender, race and body type are mixed up in one data set. Obvious candidates like Principal Component Analysis (PCA) are unsuitable because of the sparsity of the data and the

presence of noise. Our proposed method solves these challenges. It accepts sparsity of the data naturally, as well as removing noise, thus smoothing the data.

We present the observations as a matrix Y, where measurements related to each person represent a row in the matrix. Each column will contain observations taken at a given moment. Many cells in the matrix will be missing. We also create a matrix W, which has same dimensions as Y. This matrix will contain the value 1 anywhere where the corresponding person had an observation at the corresponding time, and 0 otherwise.

We will try to find a set of matrices Z which are of the same dimensions as Y, are close to matrix Y at the points where observations are available, and are not very "complex". This can be written as the following optimisation problem:

$$Z_{opt} = \arg\min_{Z} \Big(\|W \odot (Y - Z)\|^2 + \lambda \,\text{Complexity}\,(Z) \Big).$$

Note that the dataset considered has a few different classes of curves with considerably different shapes. Taking this into account, we will represent Z as a multiplication of two matrices $Z = UV^T$. We consider matrix V as a set of "shapes", and U as coefficients which these "shapes" are mixed with. Let us define matrix Z as not being "complex" if the "shapes" in matrix V are "smooth" and the "coefficients" in matrix U are small. Thus, we arrive at the following optimisation problem:

$$\{(U, V)\} = \arg\min_{U,V} (\|W \odot (Y - UV^T)\|^2 + \lambda \|U\|^2 + \theta \|LV\|^2), \tag{2}$$

where:

- Y is an $n \times m$ matrix of two-dimensional data;
- U is an $n \times k$ matrix of "coefficients";
- V is an $k \times m$ matrix of "shapes";
- $\lambda > 0$ and $\theta > 0$ are parameters;
- L takes second derivatives of the columns of the matrix V;
- $\|.\|$ is the Frobenius norm;
- $\odot$ refers to element-wise matrix multiplication;
- W is an $n \times m$ matrix of zeros and ones which has the value 1 at the places where Y has values and 0 otherwise. It allows missed elements of Y to be disregarded.

This can now be seen to be a ROPES problem as (2) can obviously be reduced to problem (1) by setting the matrix K in (1) to the identity matrix.

We note that the target function $J(U,V) = \|(Y - UV^T)\|^2 + \lambda\|KU\|^2 + \theta\|LV\|^2$ in problem (1) is a polynomial of power 4 over the elements of the matrices U and V. It raises questions about the uniqueness of the solution and the methods used to find it (them). It is unclear from the problem's definition whether $J(U,V)$ has one or many local minima, whether it has one or many global minima, or whether there are local minima which are not global minima. The best case scenario from a computational point of view (when the analytical solution is not known) is when there is one single local minimum (which is therefore also the global one when the function domain is a compact set). This is not the case for our problem, since if a given pair (U,V) is a global/local minimum, then for any conforming orthonormal square matrix R, pair (UR, VR) will also be a global/local minimum.

1.2 Motivation based on mortality data

Let $m_{x,t}$ denote an observed mortality rate for a particular age x and a particular year t. We define $m_{x,t} = d_{x,t}/e_{x,t}$, where $d_{x,t}$ is the number of deaths during year t for people who died at age x years, and $e_{x,t}$ is the total number of years lived by people aged x during year t.

Mortality rates are used to compute life tables, life expectancies, insurance premiums, and other items which are of interest to demographers and actuaries. As we can see from the definition, mortality rates are two dimensional: one dimension is time and the other dimension is age.

Observed mortality rates are noisy data. To stabilise the variance of the noise, it is necessary to take logarithms. Taking logarithms also makes sense because various features of the data for low mortality rates (for ages 1 to 40) obtain a clearer shape after the transformation. Moreover, different factors affect mortality rates in a multiplicative manner, and after taking logarithms the effects become additive, which is also a good feature for the approach we will consider later.

Figure 2 shows log mortality rates for females in France from 1950 to 1970. The data are taken from the demography package for R (Hyndman, 2012); it was originally sourced from the Human Mortality Database (2008).

Mortality rates are functional data: they are smooth in two dimensions (time and age) although observed with noise. Similarly to the approach presented in Section 1.1, logarithms of mortality rates can be decomposed into "shapes" and "coefficients" by solving the following minimisation

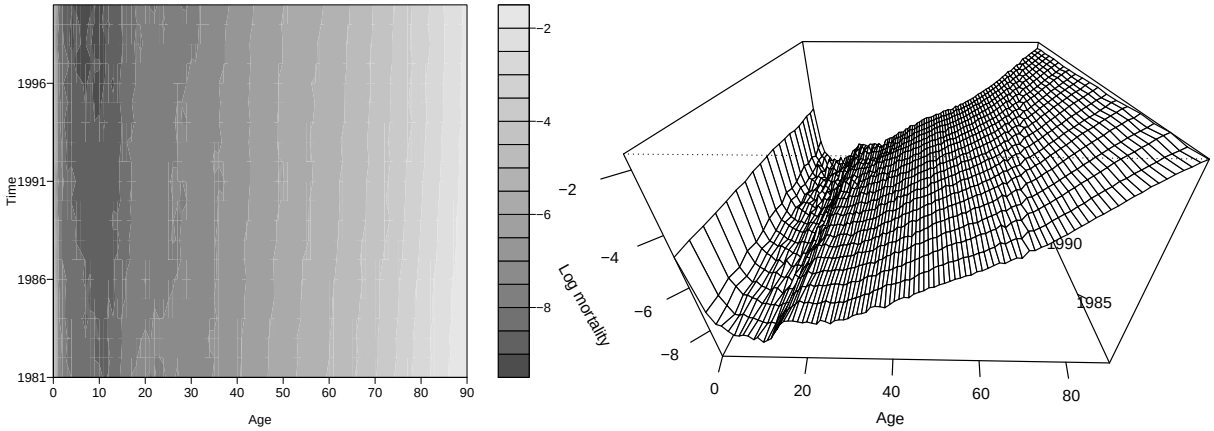

Figure 2: *Natural logarithm of French female mortality rates (ages: 0-90, time: 1981–2000 (in years)).*

problem:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\left(\left\|Y - UV^T\right\|^2 + \lambda\|KU\|^2 + \theta\|LV\|^2\right),$$

where

- Y is an $n \times m$ matrix, with n being the number of years for which mortality data are available and m being the maximum age; $Y_{t,x} = \log(m_{x,t})$;
- U is an $n \times k$ matrix of "coefficients", $k = \min(n,m)$;
- V is a $k \times m$ matrix of "shapes";
- K is a matrix which takes the second derivatives of the columns of matrix U;
- L is a matrix which takes the second derivatives of the columns of matrix V;
- $\lambda > 0$ and $\theta > 0$ are smoothing parameters.

Thus, this is also a ROPES problem. Historical mortality are usually relatively complete, with few missing observations. However, it becomes a sparse estimation problem when we want to forecast as all the data beyond the time of the last observation are missing. We will consider this aspect of the problem in Section 7.

1.3 Connection to PCA

Principal Component Analysis (PCA) is a standard statistical tool which is usually described as a dimension reduction technique. PCA is also a technique which exposes the internal structure of data without any previous knowledge about it.

Suppose that we have a matrix of observations Y, where each row represents a single multivariate observation (often matrix Y is transformed beforehand by removing the observed row mean from each row). PCA can be computed using a Singular Value Decomposition (SVD) in which Y is decomposed as $Y = PDQ^T$, where P and Q are orthonormal matrices and D is a diagonal matrix. PCA presents Y as a multiplication of two matrices, $Y = UV^T$, where $U = PD$ and $V = Q$. U is called a matrix of scores and V is a matrix of eigenvectors. Such names become self-explanatory when we write $Y = \sum_j u_j v_j^T$, where u_j and v_j are columns of U and V respectively. This representation shows that each row of the matrix Y is a linear combination of features (eigenvectors) v_j added with weights (values in a corresponding row of matrix U). It also shows that Y is sum of rank one matrices $u_j v_j^T$.

On the other hand, the solutions of PCA are also the solutions of the following minimisation problem:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\left(\left\|Y - UV^T\right\|^2 + \left\|(J_k - I_k) \odot U^T U\right\|^2 + \left\|VV^T - I_k\right\|^2\right),$$

where

- Y is $n \times m$ matrix which can be interpreted as either multivariate data (each row represents a single observation) or two-dimensional data;
- U is an $n \times k$ matrix of "scores";
- V is a $k \times m$ matrix of "features";
- J_k is a $k \times k$ matrix where all elements are 1;
- I_k is a $k \times k$ identity matrix;
- $\odot$ refers to element-wise matrix multiplication;
- $\|.\|$ is the Frobenius norm.

If we consider the terms $\left\|(J_k - I_k) \odot U^T U\right\|^2 + \left\|VV^T - I_k\right\|^2$ as regularisation terms and apply a different set of regularisation restrictions on the eigenvectors and scores we get:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\left(\left\|Y - UV^T\right\|^2 + \lambda\|KU\|^2 + \theta\|LV\|^2\right),$$

where:

- $\lambda > 0$ and $\theta > 0$ are smoothing parameters;

- L and K are "complexity" matrices which transform the scores U and eigenvectors V into corresponding "complexity" matrices. For example, K and L can be matrices which calculate second differences over the elements of the columns of U and V.

Thus, PCA is a type of ROPES problem.

A similar optimisation problem was mentioned by the winning team of the Netflix Prize competition (Töscher et al., 2009), and a related algorithm called "Basic SVD" was used (among many other algorithms).

1.4 Connection to ridge regression

The ROPES optimisation problem in (1) is also related to ridge regression. Ridge regression involves looking for a solution to the following optimisation problem:

$$\beta = \arg\min_{\beta} \left(\left\| y - X\beta \right\|^2 + \theta \left\| L\beta \right\|^2 \right), \tag{3}$$

where:

- y is a data vector (observations);
- β is a vector representing parameters;
- X is an observed matrix which transforms the parameters;
- $\theta > 0$ is a smoothing parameter;
- L is a "complexity" matrix which transforms β into a corresponding "complexity" vector;
- $\|.\|$ is the Frobenius norm.

Since the function under arg min is quadratic over elements of the vector β, the solution is unique and can easily be expressed analytically.

Suppose matrix X is not known and requires estimation (which changes the meaning of matrix X significantly). Then some restrictions on the set of possible matrices X must be imposed. These can be implemented by adding one more regularisation term to the minimising function in (3). It is also logical to extend our approach to multivariate data and replace the vector β with a matrix B:

$$\{(X, B)\} = \arg\min_{X,B} \left(\left\| Y - XB^T \right\|^2 + \lambda \left\| KX \right\|^2 + \theta \left\| LB \right\|^2 \right), \tag{4}$$

where:

- Y is an $n \times m$ matrix which can be interpreted as either multivariate data (where each row represents a single observation) or two-dimensional data;
- X is an $n \times k$ matrix of estimated "scores";
- B is a $k \times m$ matrix of estimated "features";
- $\lambda > 0$ and $\theta > 0$ are smoothing parameters;
- K and L are "complexity" matrices which transform multivariate "scores" X and "features" B into corresponding "complexity" matrices;
- $\|.\|$ is the Frobenius norm.

As we can see, the minimisation problem in (4) is equivalent to problem (1). Therefore, ridge regression defined by (3), is related to, but is not identical to, a ROPES problem.

2 Numerical solution of ROPES

Problem (1) involves the minimisation of a quartic polynomial of many variables. Since the analytical solution of such problems is unavailable, we use a numerical approach in which we first reduce the problem to convex optimization, and then use a gradient descent algorithm. We discuss these below, and at the same time investigate the behaviours of the optimising function and the optimisation problem.

2.1 Reduction to a convex optimization problem

In this section, we reduce ROPES to a convex optimisation problem and discuss the uniqueness of its solutions. To be precise,

$$Z = UV^T, \tag{5}$$

is unique, with some additional restrictions on matrices W, Y, L and K. Even when the matrix Z is not unique, the set of solutions $\{Z\}$ is convex.

We start by restricting matrices K and L to be square and of full rank. Problem (1) can then be solved by finding:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\left(\left\|W \odot \left(Y - \frac{K^{-1}UV^TL^{-T}}{2(\lambda\theta)^{\frac{1}{2}}}\right)\right\|^2 + \frac{1}{2}(\|U\|^2 + \|V\|^2)\right), \tag{6}$$

and transforming the set of solutions $\{(U,V)\}$ to $\left\{(2\lambda)^{-\frac{1}{2}}K^{-1}U, (2\theta)^{-\frac{1}{2}}L^{-1}V)\right\}$.

Let us note that

$$\|Z\|_* = \frac{1}{2} \min_{UV^T=Z} \left(\|U\|^2 + \|V\|^2 \right),$$

where $\|Z\|_*$ is the nuclear norm of the matrix Z (see Srebro et al., 2005; Jaggi et al., 2010). Then (6) is equivalent to the following problems:

$$\{(U,V)\} = \arg\min_{UV^T \in \{Z\}} \left(\|U\|^2 + \|V\|^2 \right) \tag{7}$$

and

$$\{Z\} = \arg\min_{Z} \left(f(Z) + \|Z\|_* \right), \tag{8}$$

where

$$f(Z) = \left\| W \odot \left(Y - \frac{K^{-1} Z L^{-T}}{2(\lambda\theta)^{\frac{1}{2}}} \right) \right\|^2.$$

Since f is a quadratic function of the the elements of Z, and f cannot be negative, f must be a convex function of Z. Since the nuclear norm is a convex function as well, problem (7) is a convex optimisation problem and the set of its solutions is convex.

Noting also that function in (1) is a smooth function (polynomial) of the elements of K and L, we can conclude that (1) has the same properties without restricting the matrices K and L to be of full rank. It is also clear that if K or L are not square matrices, they can be replaced with square matrices $K_{\text{square}} = (K^T K)^{\frac{1}{2}}$ and $L_{\text{square}} = (L^T L)^{\frac{1}{2}}$ without it having any impact on the result.

Therefore, the new ROPES method of decomposing two-dimensional data is a convex optimisation problem in the space of matrices $Z = UV^T$.

2.2 Numerical solution by gradient descent

The numerical approach described in Section 2.1 is not easy to implement in practice, since matrices K and L can be singular or almost singular.

Another problem appears if we try to use gradient descent (one of the most popular methods) to solve the optimisation problem (8): it is very difficult to find good descent directions for the optimising function. To avoid such difficulties, we solve problem (1) directly.

Since problem (1) is not convex optimisation, it is not clear whether it has only global minima or whether local minima are present as well (Rennie and Srebro, 2005). Theorem 1 below shows that all local minima are global minima as well, and justifies our use of the gradient descent method (although the theorem still does not guarantee its convergence, as there is a small chance that the gradient descent may get stuck in saddle points, for example). Rennie and Srebro (2005)

and our own experiments show that the gradient descent approach works reasonably well. For our calculations, we use R and the method "optim" (L-BFGS-B and CG) in the "stats" R package (R Development Core Team, 2013).

Theorem 1. *For any convex function $f(Z)$ which takes an $n \times m$ matrix as its input and has continuous derivatives everywhere, all local minima of the function*

$$J(U,V) = f(UV^T) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right) \tag{9}$$

are also global minima, where U and V are $n \times k$ and $m \times k$ matrices respectively, and $k = \min(n,m)$.

We prove this theorem by proving a series of propositions.

Proposition 1. *If (U,V) is a local minimum of $J(U,V)$, then $U^TU = V^TV$.*

Proof. Since $J(U,V)$ is differentiable on u_{ij} and v_{ij} (elements of U and V) and the derivatives are continuous, all partial derivatives at local minimum (U,V) are 0:

$$0 = \frac{\partial J(U,V)}{\partial U} = G(UV^T)V + U \tag{10}$$

$$0 = \frac{\partial J(U,V)}{\partial V} = G(UV^T)^TU + V, \tag{11}$$

where

$$G(Z) = \frac{\partial f(Z)}{\partial Z}.$$

After multiplying (10) by U^T from the left, transposing and then multiplying (11) by V from the right, and subtracting the results, we get:

$$U^TU = V^TV. \tag{12}$$

□

Corollary 1. *If (U,V) is a local minimum of $J(U,V)$, then*

$$\|U\|^2 = \|V\|^2$$

and

$$J(U,V) = f(UV^T) + \|U\|^2 = f(UV^T) + \|V\|^2.$$

Definition 1. $\underset{x\in X}{\operatorname{loc\,arg\,min}}(h(x))$ *is the set of all local minima of function* $h(x)$ *over set* X*:*

$$\underset{x\in X}{\operatorname{loc\,arg\,min}}(h(x)) = \{x \in X \mid \exists \epsilon = \epsilon(x) > 0 \ \forall x' \in X \ : \ \|x - x'\| < \epsilon \Rightarrow h(x) \le h(x')\}.$$

Proposition 2. *The set of local minima of the problem*

$$f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right) \underset{U,V:\,(U^TU = V^TV)}{\rightarrow} \min \tag{13}$$

includes the set of local minima of the problem

$$f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right) \underset{U,V}{\rightarrow} \min. \tag{14}$$

Using Definition 1, this statement can also be written as:

$$\underset{U,V}{\operatorname{loc\,arg\,min}}\left(f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right)\right) \subseteq \underset{U,V:\,(U^TU = V^TV)}{\operatorname{loc\,arg\,min}}\left(f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right)\right) \tag{15}$$

Similarly

$$\underset{U,V}{\operatorname{arg\,min}}\left(f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right)\right) = \underset{U,V:\,(U^TU = V^TV)}{\operatorname{arg\,min}}\left(f\left(UV^T\right) + \frac{1}{2}\left(\|U\|^2 + \|V\|^2\right)\right). \tag{16}$$

This proposition states that, by restricting the set of matrices (U,V) to matrices which satisfy (12), we can add more local minima, but none of them will be global. After proving this statement, we can prove Theorem 1 simply by showing that all local minima of problem (13) are global.

Proof. If (U_1,V_1) is a local minimum of problem (14), then, according to Proposition 1, $U^TU = V^TV$. This means that (U_1,V_1) is also a local minimum of problem (13), and proves (15), the first part of Proposition 2.

To prove (16), it is sufficient to show that the set of global minima S_1 of $J(U,V)$ (see (9)) is the same as set S_2 of global minima of $J(U,V)$ when the pairs (U,V) are restricted by the equation $U^TU = V^TV$:

- If $(U_1, V_1) \in S_1$, then (U_1, V_1) is also a local minimum, and according to Proposition 1: $U^T U = V^T V$.

  Since (U_1, V_1) is a global minimum over the unrestricted set and belongs to the restricted set as well, (U_1, V_1) is also a global minimum over the restricted set: $(U_1, V_1) \in S_2$. Therefore $S_1 \subseteq S_2$.

- On the other hand, if pair $(U_2, V_2) \in S_2$, then it is a global minimum of $J(U, V)$ over the restricted set of pairs (U, V). We know that S_1 is the set of points (U, V) where $J(U, V) = \min_{U,V}(J(U, V))$ and $S_1 \neq \emptyset$. By the definition of the S_2 function, $J(U, V)$ has the same value for every $(U, V) \subset S_2$. Since $S_1 \subseteq S_2$ (see above), $J(U_2, V_2) = \min_{U,V}(J(U, V))$. Consequently, $(U_2, V_2) \in \arg\min_{U,V}(J(U, V))$. This means that $(U_2, V_2) \in S_1$, and therefore $S_2 \subseteq S_1$.

□

Next we prove the following "technical" proposition, which clarifies the dependency between matrices U and V when (12) is satisfied.

Proposition 3. *For all $n \times k$ matrices U and for all $m \times k$ matrices V such that*

$$U^T U = V^T V, \tag{17}$$

there exists an $m \times n$ matrix W such that $WW^T = I$ and $V = WU$, where I is an $m \times m$ identity matrix.

Proof. The proof involves the construction of the matrix W. Using the singular value decomposition, we can present V and U as

$$V = PDQ^T \tag{18}$$

$$\text{and} \qquad U = RGS^T, \tag{19}$$

where P, Q, R and S are orthonormal matrices, and D and G are diagonal matrices with positive diagonal values which are sorted by descent.

Substituting V and U into (17), we have

$$QD^2Q^T = SG^2S^T. \tag{20}$$

Since the expressions on the left and right sides of (20) are the same matrix, QD^2Q^T and SG^2S^T must have the same singular values with the same multiplicity. Taking into account the fact that D and G have diagonal values which are positive and sorted, we conclude that $D = G$.

Thus, (20) can be modified further to give $QD^2Q^T = SD^2S^T$ and $(S^TQ)D^2(S^TQ)^T = D^2$. Let us denote

$$\Psi = S^TQ, \tag{21}$$

where Ψ is an orthonormal square matrix. Then Ψ can be presented as

$$\Psi = \prod_{i=1}^{\ell} \Psi_i, \tag{22}$$

where ℓ is the number of different diagonal elements in matrix D and U_i are orthonormal transformations which are "working" (can be non-identity) inside the linear space defined by eigenvectors with the same eigenvalue.

Since Ψ_i are "working" in orthogonal subspaces, $\Psi_i\Psi_j = \Psi_j\Psi_i$ for all $i \in \overline{1,\ell}$ and for all $j \in \overline{1,\ell}$. Moreover, since all diagonal elements of matrix D are the same for such subspaces, $D\Psi_i = \Psi_i D$ for all i.

Using (21) and (22), we can write $S^TQ = \prod_{i=1}^{\ell} \Psi_i$ or

$$S^T = \left(\prod_{i=1}^{\ell} \Psi_i\right) Q^T. \tag{23}$$

Substituting (23) into (19) and taking into account the fact that $D = G$, we get $U = RD(\prod_{i=1}^{\ell} \Psi_i)Q^T$. Since D and Ψ_i are commutative,

$$U = R\left(\prod_{i=1}^{\ell} \Psi_i\right) DQ^T.$$

Recalling that $V = PDQ^T$ according to (18), we can conclude the proof by defining the matrix W as

$$W = P\left(\prod_{i=1}^{\ell} \Psi_i^T\right) R^T.$$

□

As we mentioned earlier, to conclude the proof of Theorem 1, it is sufficient to show the following.

Proposition 4. *All local minima of problem* (13) *are global minima:*

$$\operatorname*{loc\,arg\,min}_{U,V:\,(U^TU=V^TV)}\left(f\left(UV^T\right)+\frac{1}{2}\left(\|U\|^2+\|V\|^2\right)\right)=\operatorname*{arg\,min}_{U,V:\,(U^TU=V^TV)}\left(f\left(UV^T\right)+\frac{1}{2}\left(\|U\|^2+\|V\|^2\right)\right).$$

Proof. Using Corollary 1 and Proposition 3, we can write:

$$\begin{aligned}\operatorname*{loc\,arg\,min}_{U,V:\,(U^TU=V^TV)}\left(f\left(UV^T\right)+\frac{1}{2}\left(\|U\|^2+\|V\|^2\right)\right)&=\operatorname*{loc\,arg\,min}_{U,V:\,(U^TU=V^TV)}\left(f\left(UV^T\right)+\|V\|^2\right)\\&=\operatorname*{loc\,arg\,min}_{U,V:\,(\exists W:\,W^TW=I\,\&\,WV=U)}\left(f\left(UV^T\right)+\|V\|^2\right)\\&=\Omega_1\left(\operatorname*{loc\,arg\,min}_{W,V:\,(W^TW=I)}\left(f\left(WVV^T\right)+\|V\|^2\right)\right)\\&=\Omega_1\left(\operatorname*{loc\,arg\,min}_{W,V:\,(W^TW=I)}\left(f\left(WVV^T\right)+\operatorname{tr}\left(VV^T\right)\right)\right),\end{aligned}$$

where Ω_1 is a function which transforms pair (W,V) to pair $(U,V)=(WV,V)$.

Since VV^T is a symmetric matrix, it can be decomposed as $VV^T=QDQ^T$, where Q is orthonormal and D is a diagonal matrix with non-negative elements. Therefore, we can continue:

$$\begin{aligned}&\operatorname*{loc\,arg\,min}_{W,V:\,(W^TW=I)}\left(f(WVV^T)+\operatorname{tr}(VV^T)\right)\\&\qquad=\Omega_2\left(\operatorname*{loc\,arg\,min}_{W,Q,D:\,(W^TW=I\,\&\,QQ^T=I\,\&\,D=diag(d_i\geq 0))}\left(f(WQDQ^T)+\operatorname{tr}(QDQ^T)\right)\right)\\&\qquad=\Omega_3\left(\operatorname*{loc\,arg\,min}_{P,Q,D:\,(P^TP=I\,\&\,QQ^T=I\,\&\,D=diag(d_i\geq 0))}\left(f(PDQ^T)+\operatorname{tr}(D)\right)\right),\end{aligned}$$

where Ω_2 is a function which transforms the triplet (W,Q,D) into pairs $\{(W,V)\}=\{(W,(QDQ^T)^{\frac{1}{2}})\}$ and Ω_3 is a function which transforms the triplet (P,Q,D) into pairs $\{(W,V)\}=\{(PQ^T,(QDQ^T)^{\frac{1}{2}})\}$.

Then

$$\operatorname*{loc\,arg\,min}_{P,Q,D:\,(P^TP=I\,\&\,QQ^T=I\,\&\,D=diag(d_i\geq 0))}\left(f\left(PDQ^T\right)+\operatorname{tr}(D)\right)=\Omega_4\left(\operatorname*{loc\,arg\,min}_{Z}\left(f(Z)+\|Z\|_*\right)\right)$$

where Ω_4 is a function which transforms the matrix Z into triplets $\{(P,Q,D)\}$ such that $Z=PDQ^T$, P and Q are orthonormal matrices, and D is a diagonal with non-negative elements.

The last problem,

$$f(Z) + \|Z\|_* \underset{Z}{\to} \min,$$

is a convex optimisation problem. It means that all local minima of this problem are global minima. □

This concludes the proof of Theorem 1.

3 Canonical ROPES solutions

Among the many solutions of (7), there is a solution $(U, V) = (P\sqrt{D}, Q\sqrt{D})$, where $Z = PDQ^T$ and PDQ^T is the SVD representation of matrix Z. Let us call it the canonical solution of problem (8).

The solution Z can also be presented as $Z = \sum_{i=1}^{m} d_i p_i q_i^T$, where p_i and q_i are the column vectors of the matrices P and Q, and d_i are scalars and the diagonal elements of the matrix D. If we denote $z_i = p_i q_i^T$, then

$$Z = \sum_{i=1}^{m} d_i z_i. \tag{24}$$

Let us call the decomposition in (24) the canonical decomposition of the solution of (7), and the vectors p_i and q_i and scalars d_i, canonical scores, vectors and values respectively.

If (U, V) is one of the solutions of problem (1) and matrices L and K are not singular, the canonical solution can be calculated using the following procedure:

1. Calculate $Z_* = KUV^T L^T$.
2. Use SVD to write $Z_* = P_* D_* Q_*^T$.
3. The canonical solution for problem (1) will then be $(U_*, V_*) = (K^{-1} P_* \sqrt{D_*}, L^{-1} Q_* \sqrt{D_*})$.

The canonical decomposition of the solution $Z = UV^T$ of the problem (1) will be:

$$Z = \sum_{i=1}^{m} d_{*i} z_{*i}, \qquad \text{where} \quad z_{*i} = K^{-1} p_{*i} q_{*i}^T L^{-T},$$

and p_{*i} and q_{*i} are the column vectors of the matrices P_* and Q_*, and d_{*i} are scalars and the diagonal elements of matrix D_*. The vectors $K^{-1} p_{*i}$ and $L^{-1} q_{*i}$ and the scalars d_{*i} will represent the canonical scores, vectors and values of problem (1), respectively.

Similarly to the Lasso method, we expect that many coefficients d_i and d_{*i} will be negligibly small or zero. Thus, the canonical decomposition represents the solution as the sum of a small number of rank one matrices.

4 The model implied by ROPES

Let us again consider restrictions on problem (1) when K and L are full rank square matrices and W is a matrix taking only 0 and 1 elements. Then, the objective function of the optimisation problem can be rewritten as:

$$\begin{aligned} J(U,V) &= \left\| W \odot \left(Y - UV^T \right) \right\|^2 + \lambda \|KU\|^2 + \theta \|LV\|^2 \\ &= 2\sigma^2 \left(\left\| \frac{1}{\sqrt{2}\sigma} W \odot \left(Y - UV^T \right) \right\|^2 + \left\| \frac{\sqrt{\lambda}}{\sqrt{2}\sigma} KU \right\|^2 + \left\| \frac{\sqrt{\theta}}{\sqrt{2}\sigma} LV \right\|^2 \right) \\ &= \text{Const} \times \left(\left\| \frac{1}{\sqrt{2}\sigma} W \odot E \right\|^2 + \left\| \frac{\sqrt{\lambda}}{\sqrt{2}\sigma} (I \otimes K) \text{vec}(U) \right\|^2 + \left\| \frac{\sqrt{\theta}}{\sqrt{2}\sigma} (I \otimes L) \text{vec}(V) \right\|^2 \right), \end{aligned} \tag{25}$$

where $E = Y - UV^T$.

Equation (25) can be considered as a minus log likelihood function (with some multiplicative and additive constants) of (26), where the observations Y are partially visible (at the places where W has ones):

$$Y = UV^T + E, \tag{26}$$

where

1. E is an $n \times m$ matrix of independent identically distributed errors $N\left(0, \sigma^2\right)$;
2. U is an $n \times k$ random matrix of "scores", $k = \min(n, m)$, which are normally distributed, such that $vec(U)$ has the distribution $\mathcal{N}\left(0, \frac{\sqrt{2}\sigma}{\sqrt{\lambda}} (I \otimes K)^{-1}\right)$; and
3. V is a $k \times m$ matrix of "shapes", which are normally distributed, such that $vec(V)$ has the distribution $\mathcal{N}\left(0, \frac{\sqrt{2}\sigma}{\sqrt{\theta}} (I \otimes L)^{-1}\right)$.

We can also note the following:

- The requirement that matrices K and L be square is not a restriction. If K and L are not square but still have full rank, they can be replaced with the square matrices $\left(K^T K\right)^{\frac{1}{2}}$ and $\left(L^T L\right)^{\frac{1}{2}}$ respectively without any change in $J(U,V)$.

- More generic cases can also be considered. For example, the errors E can be correlated and the values of the matrix W can be outside the set of $\{0,1\}$.

5 The confidence and prediction intervals

Assuming that the data are described by (26) and that the parameters λ, θ and σ^2 are known (in most cases, they can be estimated using cross validation; at present we do not take their variability into account), the confidence intervals for the solution $Z = UV^T$ can be calculated using the following procedure.

Let us denote a single solution of the minimisation problem (1) by $Z(Y) = UV^T$. It is truly a single solution, since the matrices L and K are not singular. Let us also denote the residuals of the last solution by $E(Y) = Y - Z(Y)$. We take $Z_{ij}(Y)$ to refer to a single element of the matrix $Z(Y)$ at row i and column j.

Given a matrix of observations Y and indexes $0 < i < n,\ 0 < j < m$, let us define values $\ell_{ij}(Y)$ and $u_{ij}(Y)$ such that

- $\ell_{ij}(Y)$ is the largest value which satisfies
  $\underset{\Omega}{\text{Prob}}\left(Z_{ij}(Y+\Omega) < \ell_{ij}(Y)\right) \le \frac{1}{2}(1 - p_{\text{conf}})$
- and $u_{ij}(Y)$ is the lowest value which satisfies
  $\underset{\Omega}{\text{Prob}}\left(Z_{ij}(Y+\Omega) > u_{ij}(Y)\right) \le \frac{1}{2}(1 - p_{\text{conf}})$,

where $0 < p_{\text{conf}} < 1$ is a specified coverage probability and the elements of Ω are i.i.d $\mathcal{N}\left(0, \sigma^2\right)$.

Let us also define set of matrices $\Delta(i,j,Y)$ such that $\delta \in \Delta(i,j,Y) \iff Z_{ij}(Y+\delta) \ge \ell_{ij}(Y)$ & $Z_{ij}(Y+\delta) \le u_{ij}(Y)$. This definition implies that $\text{Prob}\left(D \in \Delta(i,j,Y)\right) \ge p_{\text{conf}}$.

We denote the "true" model by $Z_0 = U_0 V_0^T$ and $Y = Z_0 + E$, and consider the set $Z_0 - \Delta(i,j,Y)$. Our observation Y belongs to this set with a probability greater than or equal to p_{conf}:

$$\text{Prob}\left(Y \in Y_0 - \Delta(i,j,Y)\right) \ge p_{\text{conf}}.$$

Thus, $\text{Prob}\left(Y_0 \in Y + \Delta(i,j,Y)\right) \ge p_{\text{conf}}$, and

$$\left[\inf_{D \in \Delta(i,j,Y)} \left(Z_{ij}(Y+D)\right), \sup_{D \in \Delta(i,j,Y)} \left(Z_{ij}(Y+D)\right)\right]$$

is the confidence interval for element Z_{ij}.

The above ideas allow us to propose the following Monte-Carlo style algorithm in order to find p-confidence intervals for the true model Z_0.

1. Take ℓ draws of $n \times m$ matrices Δ_k, which have elements i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. We denote a set of ℓ draws by $\Delta = \bigcup_{k=1}^{\ell} \{\Delta_k\}$.
2. Create a set of "distorted" observations $Y_\Delta = Y + \Delta$, then find a set of solutions $Z(Y_\Delta)$ for them.
3. For every $0 < i < n,\ 0 < j < m$, the $\left(\frac{p}{2}\right)$ and $\left(1 - \frac{p}{2}\right)$ quantiles of the set $Z_{ij}(Y_\Delta)$ will be the approximate p-confidence intervals for element Z_{ij}.

It should be ensured that ℓ is big enough to be able to calculate interval boundaries with the required level of precision.

Prediction intervals can be found using similar ideas. The algorithm for the approximate calculation of the prediction intervals is described below.

1. Take ℓ draws of the $n \times m$ matrices Δ_k and ℓ draws of the $n \times m$ matrices Υ_k, which have elements i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. We denote these two sets of ℓ draws as $\Delta = \bigcup_{k=1}^{\ell} \{\Delta_k\}$ and $\Upsilon = \bigcup_{k=1}^{\ell} \{\Upsilon_k\}$.
2. Create a set of "distorted" observations $Y_\Delta = Y + \Delta$ and then find a set of solutions $Z(Y_\Delta)$ for them.
3. Create a set of "distorted" solutions $Z(Y_\Delta)$ using a set of random draws Υ: $Y_{\Delta\Upsilon} = \bigcup_{k=1}^{\ell} \{Z(Y_\Delta)_k + \Upsilon_k\}$.
4. For every $0 < i < n,\ 0 < j < m$, the $\left(\frac{p}{2}\right)$ and $\left(1 - \frac{p}{2}\right)$ quantiles of the set $(Y_{\Delta\Upsilon})_{ij}$ will be the approximate p-forecasting intervals for element Z_{ij}.

This technique can have many variants. For example, since calculating $Z(Y_\Delta)$ is expensive, but calculating $Y_{\Delta\Upsilon}$ knowing $Z(Y_\Delta)$ is cheap, calculating $Y_{\Delta\Upsilon}$ can be done a few times with different random draws Υ and keeping the original value $Z(Y_\Delta)$ without change.

It only remains for us to mention that, in many cases, λ and θ can be estimated using cross validation, and the variance of the elements of the matrix E in model (26) can be estimated as $\hat{\sigma}^2 = \frac{\|E(Y)\|^2}{\|W\|-1}$ (since the elements of W are 0 or 1).

6 Interpolation and extrapolation of spinal bone mineral density

We will demonstrate our new ROPES method on a subset of the spinal bone mineral density dataset used by Bachrach et al. (1999). Our aim is to interpolate and extrapolate the sparse longitudinal data presented in Figure 1 over the time dimension.

We prepare the data by subtracting the average of all curves (and will add it back in the end of this procedure), then solve the minimisation problem (2) in order to calculate the prediction. In particular, we solve the following problem, which can be reduced easily to (2):

$$\{(U,V)\} = \arg\min_{U,V}\Big(\|W\odot(Y-UV^T)\|^2+\|U\|^2+\|\mathrm{DIFF}_2(\lambda_2)V\|^2+\|\mathrm{DIFF}_1(\lambda_1)V\|^2+\|\mathrm{DIFF}_0(\lambda_0)V\|^2\Big), \tag{27}$$

where

- Y is an $n\times m$ matrix of observations, with n being the number of people tested and m being the number of points in the "features" dimension (4 points per year);
- W is an $n\times m$ matrix. $W_{p,t}=1$ where test data are available for person p at moment t and $W_{t,x}=0$ otherwise. W "masks" values which we do not know and are trying to predict;
- U is an $n\times k$ matrix of "scores", $k=\min(n,m)$;
- V is an $k\times m$ matrix of "features";
- $\mathrm{DIFF}_i(\alpha)$ is a linear operator which represents differentiation i times and multiplication of the result to the conforming vector α: $\mathrm{DIFF}_i(\alpha)=\alpha\odot(D^{(1)}...D^{(i)})$, where $D^{(.)}$ are conforming differentiation matrices

$$D^{(.)} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \dots & & & & & \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

Problem (27) can be reduced to (2) by combining (stacking) matrices $\mathrm{DIFF}_2(\lambda_2)$, $\mathrm{DIFF}_1(\lambda_1)$, and $\mathrm{DIFF}_0(\lambda_0)$ into matrix L as

$$L = \begin{bmatrix} \mathrm{DIFF}_2(\lambda_2) \\ \mathrm{DIFF}_1(\lambda_1) \\ \mathrm{DIFF}_0(\lambda_0) \end{bmatrix}.$$

We do not estimate the smoothing parameters, but select them so as to obtain reasonable curve shapes. The estimation of smoothing parameters, while working well for long term forecasts, is a difficult task here, since each person was tested over only a short period of time (3–4 years and 2–4 times only).

The method described in Section 5 is used to obtain the forecasting intervals. Since we do not take the variability of the smoothing coefficients into account, the prediction intervals are narrower than in the case where the smoothing parameters were estimated.

Some of the results of the interpolation and extrapolation (with 90% prediction intervals) for the chosen parameters are shown in Figure 3.

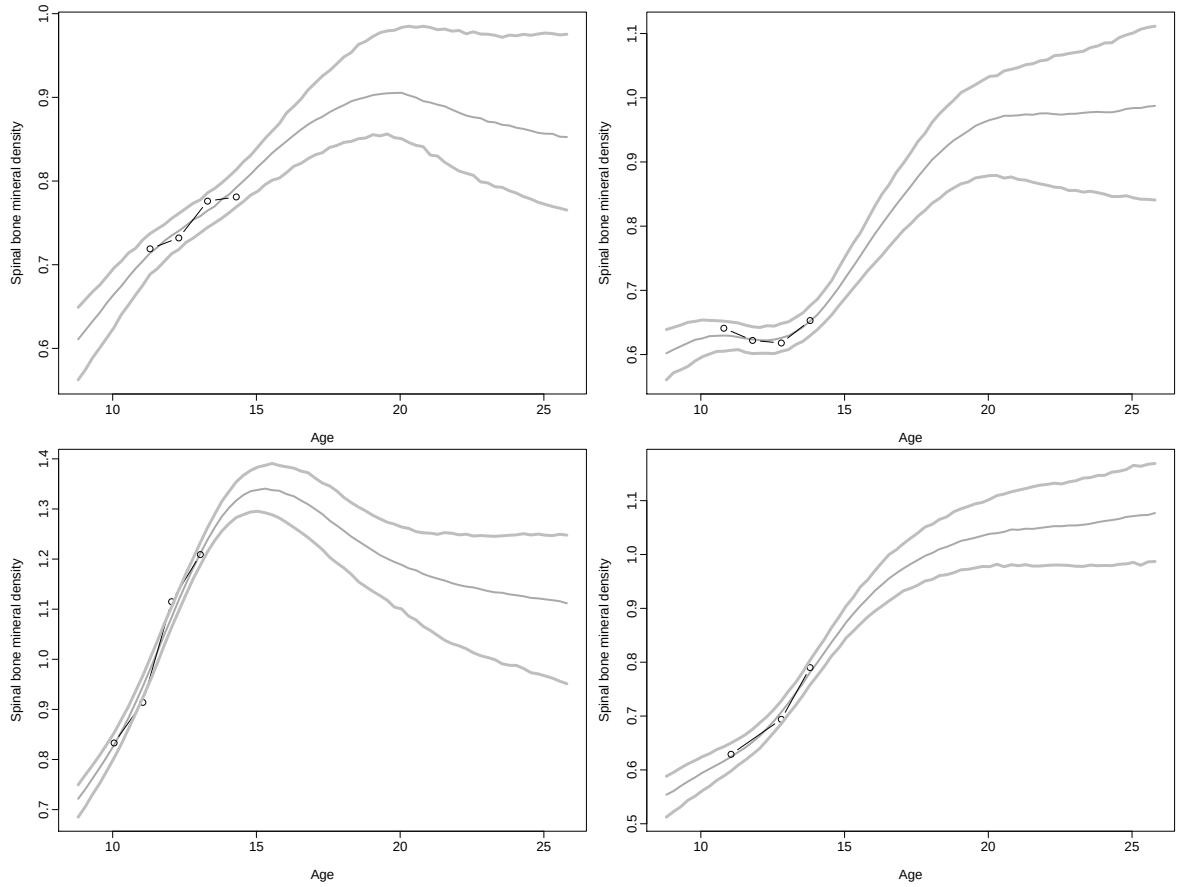

Figure 3: *Interpolation and extrapolation of the spinal bone mineral density (g/sm^2) for cases 1, 3, 4 and 190.*

The canonical decomposition has two significant terms, as can be seen from Figure 4.

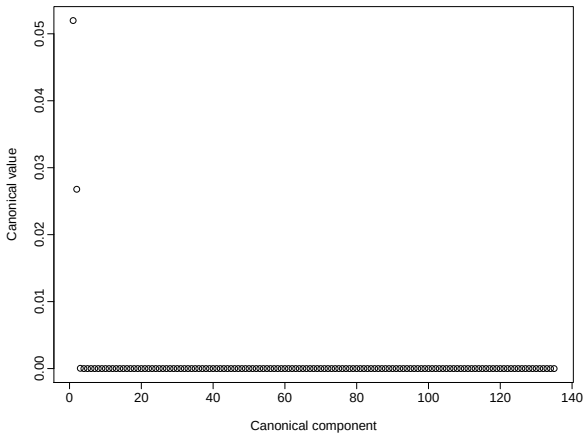

Figure 4: *Canonical values of the spinal bone mineral density dataset.*

Canonical vectors and their scores are presented in Figure 5.

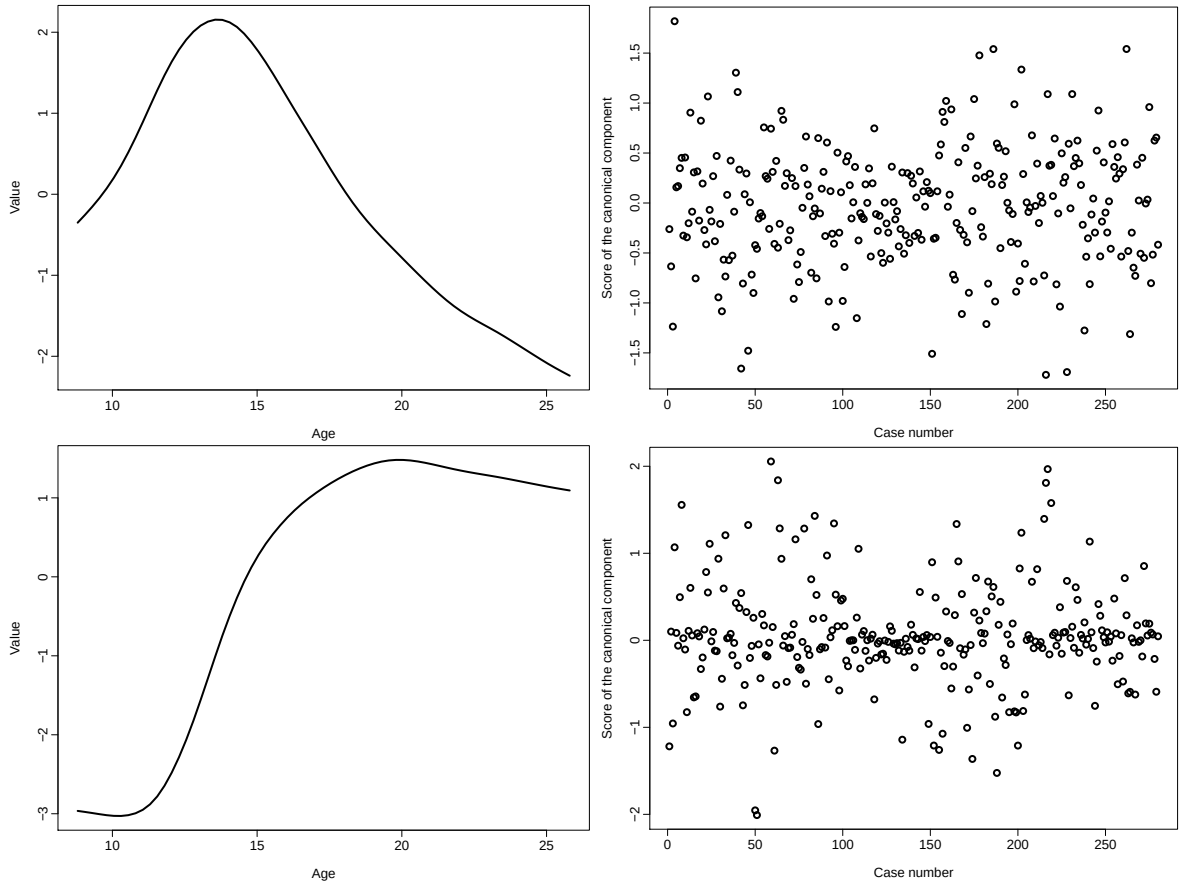

Figure 5: *The first two canonical components and their scores for the spinal bone mineral density dataset.*

The scatter plot of the first two canonical scores is presented in Figure 6.

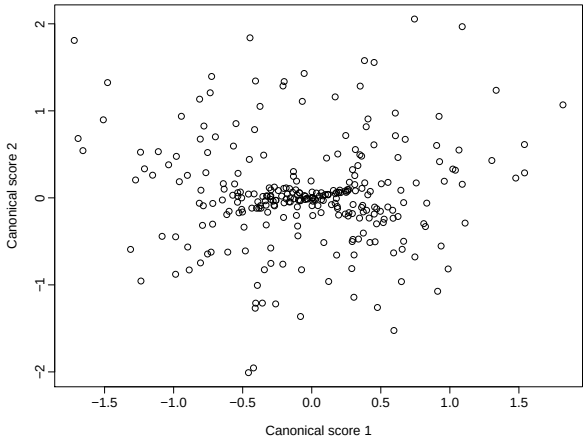

Figure 6: *Scatter plot of the first two canonical scores of the spinal bone mineral density dataset.*

7 Forecasting of mortality rates

Our aim is to forecast the bivariate surface of logarithms of mortality rates (Figure 2) in the time (t) dimension. Several different approaches have been proposed to date (see Shang et al., 2011, for an overview and comparison of the different approaches). In this paper, we use the new Linear Prediction Method (LPM) to predict mortality rates at a horizon of h and a training period of n years.

For forecasting purposes, we treat the future observations as missing. So the data are "sparse" where the sparsity occurs at the future points at which predictions need to be made.

7.1 The Linear Prediction Method (LPM)

The Linear Prediction Method is a practical method of forecasting two-dimensional functional data which uses (approximate) linearity in the data in the time dimension and takes smoothness of the data in the dimension of "features" into account.

First, let us consider a forecasting method which can be represented as a solution of the following optimisation problem:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\Big(\left\| W \odot (Y - UV^T) \right\|^2 + \|\mathrm{DIFF}_2(\mu_a)U\|^2 + \|\mathrm{DIFF}_1(\theta_a)U\|^2 + \|\mathrm{DIFF}_0(\lambda_a)U\|^2 + \|\mathrm{DIFF}_2(\mu_y)V\|^2 + \|\mathrm{DIFF}_1(\theta_y)V\|^2 + \|\mathrm{DIFF}_0(\lambda_y)V\|^2 \Big), \tag{28}$$

where

- Y is an $(n+h)\times m$ matrix of observations, with n being the period for which data is available for training, h being the prediction horizon, and m being the number of points in "features" dimension;
- W is an $(n+h)\times m$ matrix. $W_{t,x}=1$ for $1\le t\le n$ and $W_{t,x}=0$ for $n+1\le t\le n+h$. W "masks" future values (values which we do not know but are trying to predict);
- U is an $n\times k$ matrix of "scores", $k=\min(n+h,m)$;
- V is a $k\times m$ matrix of "features";
- $\mathrm{DIFF}_i(\alpha)$ is a linear operator which represents differentiation i times and multiplication of the result by the conforming vector α: $\mathrm{DIFF}_i(\alpha)=\alpha\odot(D^{(1)}...D^{(i)})$, where $D^{(.)}$ are conforming differentiation matrices

$$D^{(.)}=\begin{bmatrix}1 & -1 & 0 & ... & 0 & 0\\ 0 & 1 & -1 & ... & 0 & 0\\ ... & & & & & \\ 0 & 0 & 0 & ... & 1 & -1\end{bmatrix}.$$

Let us note that problem (28) can be reduced to a ROPES problem (1) by combining (stacking) matrices $\mathrm{DIFF}_2(\mu_a)$, $\mathrm{DIFF}_1(\theta_a)$ and $\mathrm{DIFF}_0(\lambda_a)$ into matrix K as

$$K=\begin{bmatrix}\mathrm{DIFF}_2(\mu_a)\\ \mathrm{DIFF}_1(\theta_a)\\ \mathrm{DIFF}_0(\lambda_a)\end{bmatrix},$$

and similarly, matrices $\mathrm{DIFF}_2(\mu_y)$, $\mathrm{DIFF}_1(\theta_y)$ and $\mathrm{DIFF}_0(\lambda_y)$ into matrix L

$$L=\begin{bmatrix}\mathrm{DIFF}_2(\mu_y)\\ \mathrm{DIFF}_1(\theta_y)\\ \mathrm{DIFF}_0(\lambda_y)\end{bmatrix}.$$

Here, vectors μ_a and μ_y control the smoothness of the scores and vectors. In our method, we use the following values.

1. μ_a is set to some average value controlling the smoothness in the features dimension.
2. μ_y is set to a very high value, in order to make scores almost linear in the time dimension.

3. θ_a, θ_y and λ_y are set to very small values greater than zero. There are two ideas behind this: first, they are kept greater than 0 in order to reduce the number of degrees of freedom in the decomposition; second, they are kept small in order to avoid them having much influence on the solution in terms of matrix Z (5).

4. λ_a is set to a value of 1 in order to avoid "rebalancing". "Rebalancing" is the behaviour observed in minimisation problem (28) in the case when all $\theta_\cdot$ and $\lambda_\cdot$ coefficients are zero. In such a case, the solution in terms of matrix Z (5) does not change for any $\alpha > 0$, and when μ_a is replaced with $\alpha\mu_a$ and μ_y is replaced with $\frac{\mu_y}{\alpha}$. Setting λ_a to 1 and μ_y to a very high value allows us to apply "pressure" selectively and make the scores almost linear, but not the feature vectors.

We are now ready to define LPM. LPM is a method which solves problem (28) approximately. Note that if the $\theta_\cdot$ and $\lambda_\cdot$ coefficients are zero, the solution of the problem is linear in the time dimension in areas where the matrix W has zero values. Moreover, as we described above, we choose the $\theta_\cdot$ and $\lambda_\cdot$ coefficients in (28) so as to make scores linear everywhere, including the places where the matrix W has the value 1. We therefore construct our approximate solution by solving problem (28) over available data only, and then continue obtained scores linearly into the future to get the forecast (which is obtained by summing existing feature vectors multiplied by the new scores).

7.2 The forecasts and forecast intervals

We used the method described in Section 7.1 to forecast French female mortality rates. We took the years 1981–2000 as our training set and forecast 10 years ahead for the year 2010.

The fitted (first 20 years) and forecast (last 10 years) surface is shown in Figure 7.

The residuals are shown in Figure 8. The results of the 10-year forecasting with 90% prediction intervals are shown in Figure 9.

The canonical decomposition has two significant terms, as can be seen from Figure 10. The first two canonical vectors and their scores are presented in Figure 11.

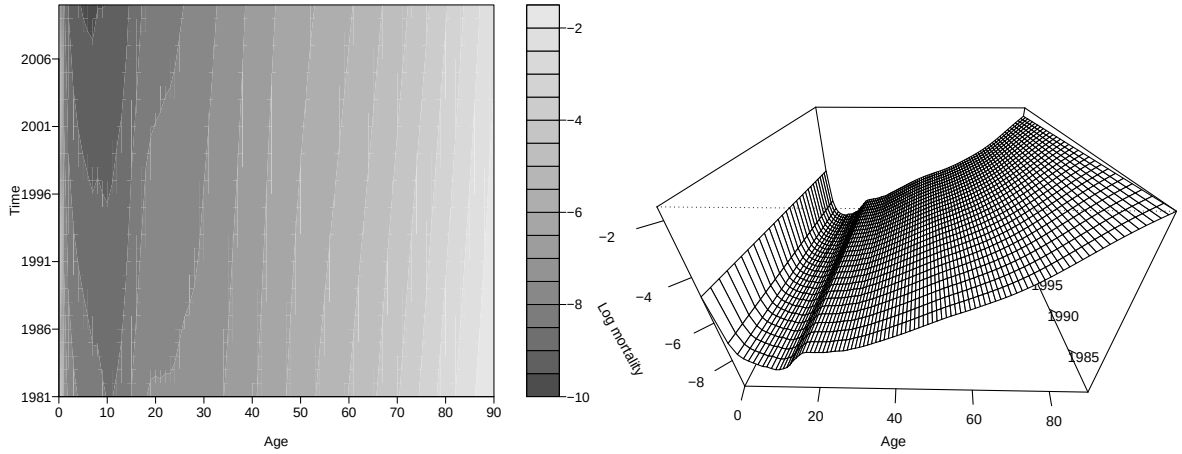

Figure 7: *The fitted (first 20 years) and forecast (last 10 years) surface for French female log mortality rates using LPM.*

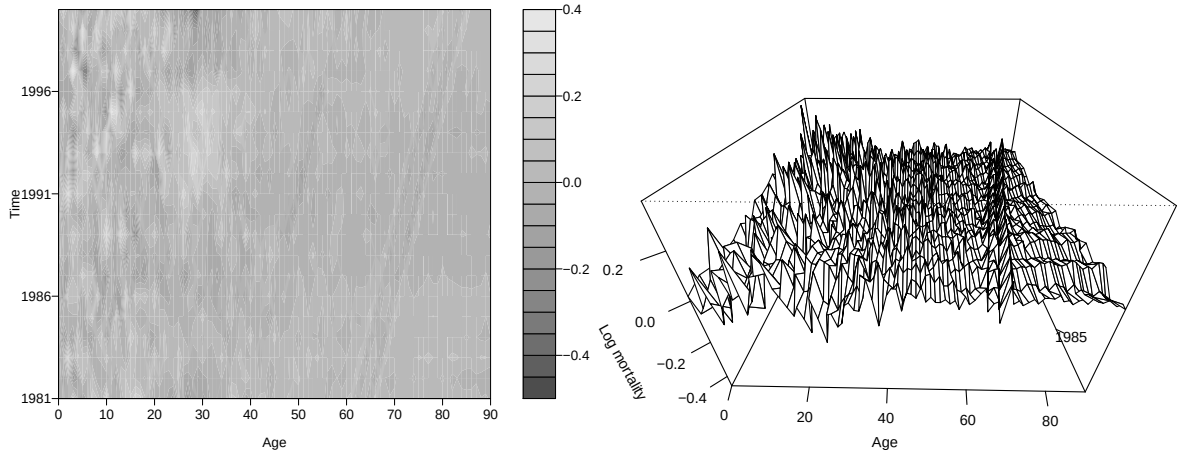

Figure 8: *The residuals for the fitted surface for French female log mortality rates using LPM.*

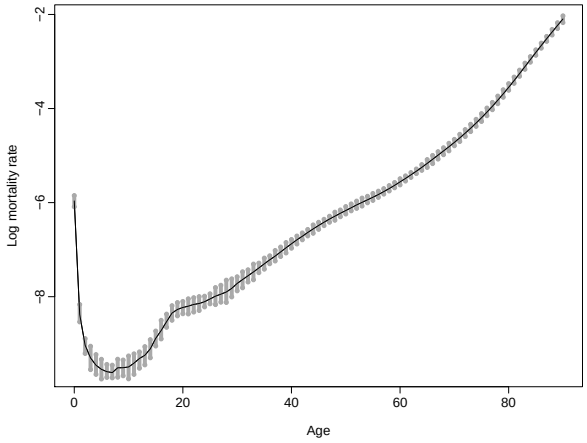

Figure 9: *The forecast of French female log mortality rates for year 2010 (with 90% prediction intervals).*

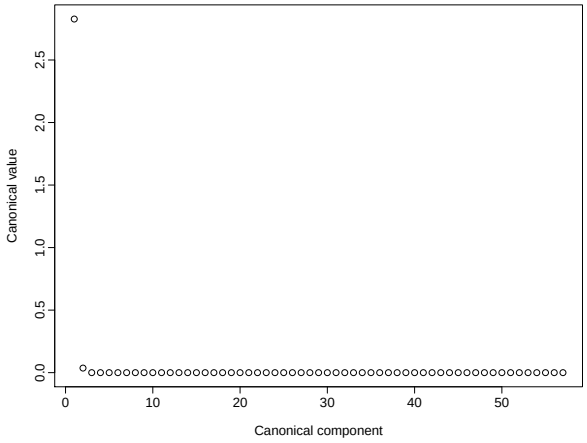

Figure 10: *Canonical values of French female log mortality rates, years 1981–2000.*

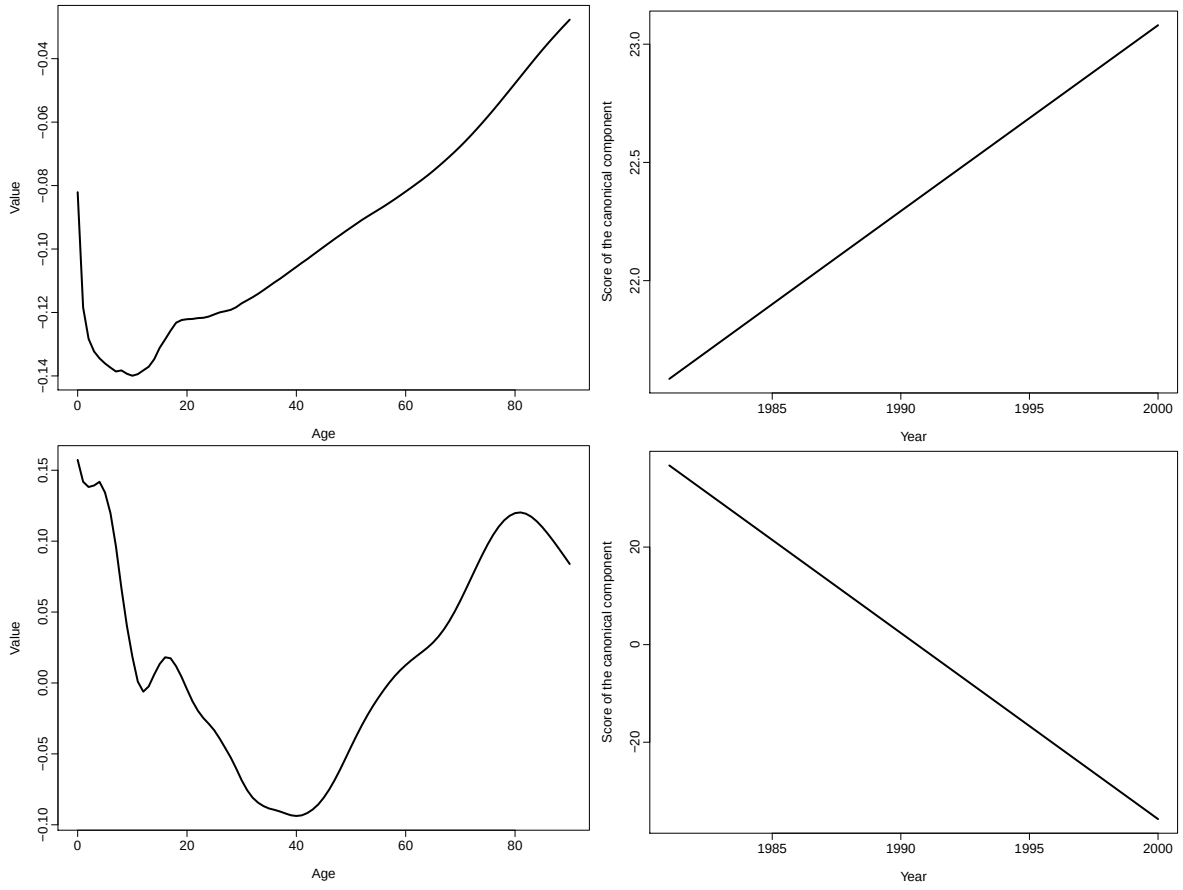

Figure 11: *The first two canonical components and their scores for French female log mortality rates, years 1981–2000.*

8 Discussion, limitations and extensions

In this paper we have introduced the ROPES method for decomposing, smoothing and forecasting two-dimensional sparse data. ROPES has features that are similar to many well-known

methods including Ridge Regression, the LASSO, Principal Component Analysis and Maximum-Margin Matrix Factorisation.

One of the main advantages of our new approach is that it allows data to be very sparse. It simplifies many different tasks from cross validation to forecasting and also imposes fewer restrictions on observed data. Another important feature of the ROPES method is that it works with noisy data directly and implements the smoothing procedure naturally as part of the optimisation process.

We have demonstrated this new approach by applying it to the interpolation and extrapolation of spinal bone mineral density data and to the forecasting of mortality rates. In the first case we dealt with extremely sparse dataset and in the second case the forecasting was implemented by treating future values as missing observations.

We have found that the new approach of decomposing, smoothing and forecasting two-dimensional data is practical and can be used for both smoothing and forecasting. In the case of forecasting it gives reasonable out-of-sample forecasts due to the ability to linearly project smoothed data.

One of the main limitations of the method can be difficulties in selecting the smoothing parameters used for estimation. The method can have quite a few such parameters and it is not always obvious how to select appropriate values. In addition to the unknown smoothing parameters, the weight matrix W can vary as well and must be specified. All this requires a good understanding of how every smoothing parameter and the weight matrix affect the estimates or forecasts before proceeding.

During our experiments we also noted that optimisation of the smoothing parameters using cross-validation can fail if data is very sparse. In such cases, alternatives to cross-validation should be used.

A further practical limitation of the method is that it can be relatively slow. As currently implemented, the whole optimisation procedure (with known smoothing parameters) can take 20 or more minutes on moderate sized data sets consisting of a few thousand observations. In some cases the optimisation method ("optim" function in R package "stats" version 3.0.2) did not report convergence, although the final result of the optimisation was very reasonable.

Further improvements of the method can include various generalisations. For example, for $p \geq 1$, problem (29) can also be reduced to a convex optimisation problem:

$$\{(U,V)\} = \arg\min_{U,V} \left(\left\| W \odot (Y - UV^T) \right\|_{L_p}^p + \lambda \|KU\|^2 + \theta \|LV\|^2 \right). \tag{29}$$

Another interesting question which we have not attempted to answer is whether the confidence intervals can be calculated analytically. In this work we estimate them using a Monte-Carlo style method.

A further problem which could be investigated is the issue of correlated and/or non-Gaussian errors when using the Linear Prediction Method which was applied to mortality data in Section 7.2. Accounting for distribution and correlation nuances should lead to more reliable prediction intervals.

We leave the investigation of these problems to later research papers.

The R code used in this article can be made provided on request.

9 Acknowledgements

The subset of the bone mineral density dataset (Bachrach et al., 1999) was kindly provided by Prof. Gareth James.

References

Bachrach, L. K., Hastie, T., Wang, M.-C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian youth: A longitudinal study. *Journal of Clinical Endocrinology & Metabolism*, 84(12):4702–4712.

Human Mortality Database (2008). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 20 Feb 2008. http://www.mortality.org.

Hyndman, R. J. (2012). *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.16. With contributions from Heather Booth, Leonie Tickle and John Maindonald. http://cran.r-project.org/package=demography.

Jaggi, M., Sulovsk, M., et al. (2010). A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 471–478.

James, G. (2010). Sparseness and functional data analysis. In Ferraty, F. and Romain, Y., editors, *The Oxford Handbook of Functional Data Analaysis*, pages 298–323. Oxford University Press.

R Development Core Team (2013). *R: A language and environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Rennie, J. D. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 713–719. ACM.

Shang, H. L., Booth, H., and Hyndman, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, 25(5):173–214.

Srebro, N., Rennie, J. D., and Jaakkola, T. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, pages 1329–1336.

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

Töscher, A., Jahrer, M., and Bell, R. M. (2009). The BigChaos solution to the Netflix grand prize. Technical report, Commendo Research & Consulting. http://www.commendo.at/UserFiles/commendo/File/GrandPrize2009_BigChaos.pdf.