Smoothing, decomposition and forecasting of multidimensional and functional time series using regularisation

A thesis submitted for the degree of

Doctor of Philosophy

by

Alexander Dokumentov

M.Sc.(Hons), Moscow State University

Department of Econometrics and Business Statistics

Monash University

Australia

June 2015

Contents

Copyright notice

Abstract

This thesis by publication is built around three articles which are at different stages of publication. All three articles have in common the concept of regularisation and they provide various applications of this concept in the field of functional data analysis.

The thesis consists of five chapters. The first chapter is an introduction and it sets the historical context for concepts such as complexity and regularisation. It also looks at different forecasting problems from the point of view of complexity and considers regularisation as a practical means of reducing complexity in statistical models.

The second part is an article “Bivariate data with ridges: two-dimensional smoothing of mortality rates”, which applies the concept of regularisation to a problem of smoothing mortality rates in particular and to any bivariate data in general. The article proposes an innovative approach for smoothing which allows for data to have abrupt “two-dimensional” changes as well as “ridges” – one dimensional statistically significant effects.

The third part is an article “Low-dimensional decomposition, smoothing and forecasting of sparse functional data”. This article proposes an innovative approach of dealing with bivariate data which allows the data to be sparse. The article demonstrates this approach by applying it to two different problems. The first is related to sparse medical data. The second is related to forecasting where the values, which need to be forecasted, are considered as missing values.

The fourth part is an article “STR: A Seasonal-Trend Decomposition Procedure Based on Regression”. Proposing a new approach of decomposing seasonal time series, it sets a new level of simplicity and generality in this field.

The last part concludes the thesis. It outlines the main ideas which drove my research as well as my understanding of my main contributions. It also discusses new possible research directions.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Alexander Dokumentov

Acknowledgement

I would like to thank Professor Rob J Hyndman for taking me on the exciting journey of research, and for his support and encouragement all this way along.

Preface

The article “Bivariate data with ridges: two-dimensional smoothing of mortality rates” will be resubmitted to the *Journal of Multivariate Analysis*.

The article “Low-dimensional decomposition, smoothing and forecasting of sparse functional data” will be submitted to the *Journal of Computational Statistics & Data Analysis*.

The article “STR: A Seasonal-Trend Decomposition Procedure Based on Regression” will be submitted to the *Journal of American Statistical Association*.

Chapter 1

Introduction

1.1 Complexity

Statistics, as a science, builds a special class of models of the observed (or partially observed) world, which structurally consist of two parts: the first part is a deterministic function, sometimes expressed as an algorithm; the second part is random input for that function or algorithm. It is assumed that the observed sequence of data was generated using that model with some random input.

Usually it is also assumed that any model is an approximation of reality. Such an assumption is often hidden in the randomness of the input. As Box and Draper (1987) mentioned, "In applying mathematics to subjects such as physics or statistics we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless". Or in other words, "all models are wrong, but some are useful".

Understanding of this fact immediately raises the problem of how to compare two (or more) models and choose the most useful (in terms of reflecting current and future experience). This question (not only for statistical models but for models in general) has appeared in various forms from ancient times. For example Aristotle at least once spoke about simplicity of a model with negative connotation: "... we must not carry its reasoning too far back, or the length of our argument will cause obscurity: nor must we put in all the steps that lead to our conclusion, or we shall waste words in saying what is manifest. It is this simplicity that makes the uneducated more effective than the educated when addressing popular audiences ...” (Aristotle, 1959).

One of the first philosophers, who clearly formulated the methodological principle which is the basis for the current most trustworthy scientific approach was William of Ockham. He introduced a principle which was later named “Ockham’s razor”. It can be interpreted for statistical models as the following paradigm: “Avoid unnecessarily complex models” (“entities are not to be multiplied unnecessarily”, see Crombie, 1959).

Acceptance of the “Ockham’s razor” principle raises at least two questions: one is “what is simplicity?” and the other is “why is a simple model better than a complex one?”. Even now both questions are not answered completely (and possibly they will never be answered), although some theoretical background has been developed.

So, what is simplicity? How can it be calculated, or at least formalised? It is quite common to find that an explanation or a theory can “appear” obvious to one person while overly complex to someone else. Is it meaningful at all to talk about simplicity or complexity in regards to a particular theory, explanation or a model?

The theoretical background for complexity (and therefore for simplicity) was established by philosophers and mathematicians Ray Solomonoff, Andrey Kolmogorov and Gregory Chaitin almost at the same time (Solomonoff, 1964a, Solomonoff, 1964b, Kolmogorov, 1963, Chaitin, 1966) (here and further I talk about structural complexity only). The most interesting corollaries from this research are that: complexity can be defined uniquely up to an unknown additive constant; it cannot be computed (although it can be bounded from above and the estimate itself is rather useless since the previously mentioned constant is unknown); probability theory can be defined in terms of algorithmic theory, rather than in terms of measure theory (see for example the proof of the law of the iterated logarithm for random by Kolmogorov sequences (Vovk, 1988)). In many cases we cannot compare the complexity of two models except for some special cases. For example we can easily compare the complexity of two linear models, which take into account different numbers of independent regressors when the number of observations goes to infinity: the model with more regressors is more complex, compared to the model with fewer regressors.

In practice the complexity of a model or data is measured using various, often rather simple methods. The simplest approach is to use various norms, for example the L_2 and L_1 norms. Other, more complex, examples include VC dimensions (Vapnik, 2000) and Rademacher complexity (Bartlett and Mendelson, 2003). Minimum Message Length (MML) and Minimum Description Length (MDL) approaches define complexity through a “compression” algorithm and, in the first case, also a distribution, which have yet to be defined by the researcher in every particular case (see for example Grünwald, Myung, and Pitt, 2005).

The question “why is a simple model better than a complex one?” also does not have an obvious answer. Moreover it is easy to produce an example when a complex explanation is better than the simplest one not just for finite observations, but, for infinitely many observations. Imagine a game where “Nature” and “Human” play against each other: “Nature” produces a sequence of numbers $\{x_\tau\}_{\tau=1}^{t}$, while “Human” at any moment t tries to guess the next number x_{t+1}. Also suppose that “Human’s” strategy to guess the next number x_{t+1} is to “explain” the sequence $\{x_\tau\}_{\tau=1}^{t}$ in the simplest way and to “project” it one step ahead producing estimate $\hat{x}_{t+1}$.

Such a strategy is very logical and directly follows Ockham’s principle of simplicity. Although it can be the worst possible strategy in the case of “Nature” at any moment t, knowing “Human’s” experience $\{x_\tau\}_{\tau=1}^{t}$ will calculate the simplest “explanation” and the prediction $\hat{x}_{t+1}$, but provide an x_{t+1} value different from that calculated, suppose $\hat{x}_{t+1} + 1$. This will make “Human” fail in all predictions (obviously in such a scenario, “Nature” must have more advanced “computational power”, than “Human” and I suppose it has).

When can we rely on the principle of simplicity? Among many cases, one is when linear models are applied to independent observations. Thus Akaike (Akaike, 1974) introduced the famous AIC (“An Information Criterion”, also known as “Akaike’s Information Criterion”), which can be considered as a means of choosing the simplest model. In this case the model is considered the simplest when the combined complexity (complexity of the input data described by the number of predictors, and complexity of the residuals described by the log-likelihood) is the smallest.

Intuitively, the usefulness of complexity reduction techniques for forecasting can be justified as follows: Let us consider two forecasting (or decomposition) methods. Suppose the first one is less complex then the second. The less complex method usually involves a smaller set (in some sense) of procedures. In such a case, applying these two methods to some data we expect them (again intuitively) to separate the training data into two parts: one is noiseless and directed to the algorithmic part of the forecasting procedure, the second part is noise, and it can be discarded or used for defining, for example, prediction intervals. In practice, clean separation never happens and, when a forecasting algorithm is trained, the poisonous noise is absorbed by the algorithm together with meaningful information.

Suppose both forecasting methods make a similar "reasonably good effort" to extract "useful information" from the data. The first method, being more complex than the second method, needs to extract more information to uniquely define itself from a bigger set of procedures, compared to the second, a simpler method. It makes the first method more exposed to noise, which can be confused with "useful information". When the first algorithm uses a mixture of noise and "useful information" for forecasting it makes the forecasts less precise and more fragile (forecasts have higher variance). On the other hand, the more complex method, covering a bigger set of procedures, can have more ability to extract useful information (if the data indeed contains such information). Therefore techniques which measure overall complexity are important. One of them – regularisation – will be discussed in the next section. I will also show how the AIC principle can be reformulated in terms of the regularisation procedure.

1.2 Regularisation

The method of regularisation was introduced by Andrey Tikhonov as a way to solve ill-posed problems (Tikhonov, 1943; Tikhonov, 1963). Consider the problem

$$Ax = b, \tag{1.2.1}$$

where A is a matrix, b is a vector and x is a vector to be determined given A and b. The problem is called ill-posed when it does not have a unique solution (it has many or no solutions). In practice it is often required to have a single solution which obviously cannot be computed using the standard method involving inversion of matrix A:

$$x = A^{-1} b. \tag{1.2.2}$$

The regularisation method proposes to solve the following minimisation problem instead of problem (1.2.1):

$$x = \operatorname*{argmin}_{x} \left(\|b - Ax\|^2 + \|\Gamma x\|^2 \right), \tag{1.2.3}$$

where the norm above is the L_2 norm and Γ is some square matrix.

Approach (1.2.3) has benefits compare to (1.2.1). First of all it is equivalent to (1.2.1) in cases when matrix A is invertible and $\Gamma = 0$. When A is not invertible and Γ is full rank (often $\Gamma = \lambda I$, for some scalar $\lambda > 0$), (1.2.3) provides a single solution which can be interpreted from a statistical, Bayesian point of view as the most probable solution of a linear regression model with normal i.i.d. errors, defined by observations b, regressors A and prior normal distribution defined by Γ. Such an interpretation clearly shows the connection between Tikhonov regularisation and Ridge Regression, discussed in the next section.

In general, regularisation is an approach where a solution of some task is shown as a minimisation problem, and where the function to be minimised is a sum of two terms: the first one expresses how far the solution is from the observed data, and the second constitutes the "complexity" of the solution. As a result the optimal solution "satisfies" reasonably well the restrictions of the task and also it is "simple". This is a direct mathematical implementation of the "Ockham's razor" principle. While the first term of the minimising expression represents the complexity of the data conditional on the solution and the second term represents complexity of the solution itself, the sum of these two terms represents the complexity of the data when the model is used as an explanation. Minimising the expression we find the model which allows the simplest explanation of the observed data:

$$\hat{s} = \operatorname*{argmin}_{s \in S} \left(\text{Complexity}(d|s) + \text{Complexity}(s) \right). \tag{1.2.4}$$

In (1.2.4) s is any solution from some set of allowed solutions S, d is the observed data (dataset) and $\hat{s}$ is the estimated solution.

Interestingly, the AIC model selection approach can be considered as a regularisation approach. AIC can be written (Akaike, 1974) as:

$$\mathrm{AIC} = -2\ln(L) + 2k, \tag{1.2.5}$$

where k is the number of regressors and L is the likelihood function. The first term $-2\ln(L)$ can be considered as a function representing how far the solution is from the observed data, and second term $2k$ can be considered as a function representing the complexity of the solution.

All in all, regularisation involves practical approaches to achieve a balance between complexity and simplicity, both of which are important features for any forecasting (or decomposition) procedure. It must be noted that complexity reduction techniques as they are used in practice by regularisation are mostly quite simple "ordering" procedures, rather than restrictions on structural complexity of the models.

1.3 Ridge Regression

In statistics, the Tikhonov regularisation method is known as Ridge Regression. It is shown (Hoerl and Kennard, 1970) that it has interesting and useful statistical properties in some ill-posed cases, compared to Ordinary Least Squares (OLS).

In particular, let us consider a linear regression problem

$$y = X\beta + \epsilon, \tag{1.3.1}$$

which is defined by a vector of observations y and a matrix of regressors X. The errors ϵ are identically, normally distributed and independent. The goal is to estimate coefficients β. In this case, the OLS solution can be written as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|, \tag{1.3.2}$$

and has the following analytical solution:

$$\hat{\beta} = \left(X'X\right)^{-1} X'y. \tag{1.3.3}$$

It can be shown that in the class of unbiased estimators, estimate (1.3.3) has the lowest variance (see for example Hayashi, 2000). Such an estimate works well in many practical cases except some when the problem is ill-posed. The problem becomes ill-posed when matrix $X'X$ is in some way close to a singular matrix, particularly when the ratio of the maximum and the smallest eigenvalues is high. In this case, estimates (1.3.3) can experience high variance and Ridge Regression can be a solution to reduce it:

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|^2 + \|K\beta\|^2\right), \tag{1.3.4}$$

where K is a regularisation matrix, often $K = \alpha I$ for some $\alpha > 0$.

It is shown (Hoerl and Kennard, 1970) that Ridge Regression reduces variation of the estimates in return for introducing some bias.

The approach can be clarified with the following example. Let the above mentioned regression model have only two regressors and therefore matrix X has only two columns. I also assume that the regressors are highly correlated and that the true model of the data is:

$$y_i = x_{i1} + x_{i2} + \epsilon_i, \tag{1.3.5}$$

where $\{\epsilon_i\}$ are i.i.d. $\mathcal{N}(0,1)$.

Since both regressors are highly correlated the following models are almost indistinguishable for rather a wide range of coefficients α around zero:

$$y_i = (1-\alpha)x_{i1} + (1+\alpha)x_{i2} + \epsilon_i, \tag{1.3.6}$$

and, therefore, even small fluctuations in error terms lead to significant fluctuations in coefficients $\hat{\beta}$.

On the other hand, Ridge Regression, by introducing the regularisation term, reduces the fluctuation in exchange for bias. If the regularisation coefficient is chosen correctly, the final result is better in terms of the variance of $\hat{\beta}^*$.

1.4 LASSO

LASSO (Tibshirani, 1996, and also Hastie, Tibshirani, and Friedman, 2009) is another famous technique (beyond Ridge Regression) to solve problem (1.3.1) which uses regularisation. Ridge Regression uses the regularisation term which involves the L_2 norm for complexity reduction, while the LASSO uses the L_1 norm instead:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_{L_2}^2 + \lambda \|\beta\|_{L_1} \right). \tag{1.4.1}$$

Such a change affects the behaviour of the $\hat{\beta}$ coefficients significantly when λ varies. Tibshirani (1996) shows (heuristically, through some geometrical multidimensional reasoning) that when λ increases many values of vector $\hat{\beta}$ become zero. This allows the use of LASSO as a variable selection procedure.

The regularisation term in both Ridge Regression and LASSO can be associated with a prior distribution over values $\hat{\beta}$. Such a distribution is Gaussian for Ridge Regression and double exponential (Laplace) for LASSO. Differences in the prior distributions makes Ridge Regression and LASSO behave very differently, and even though LASSO uses the same regularisation technique as introduced by Tikhonov it cannot be considered as a special case of Ridge Regresion.

1.5 Convex and non-convex optimisation

All above mentioned procedures such as OLS, Ridge Regression, LASSO can naturally be presented or formulated as optimisation procedures. Therefore it is important to know the features and limitations of such optimisation procedures.

An important class of optimisation problems is convex optimisation problems. The problem

$$F(x) \underset{x \in Z}{\rightarrow} \min \tag{1.5.1}$$

is a convex optimisation problem when function F is convex and set Z is convex.

Convexity of set Z means:

$$\forall x \in Z \ \& \ \forall y \in Z \ \& \ \forall \alpha \in [0,1] \quad \alpha x + (1-\alpha) y \in Z \tag{1.5.2}$$

Function F defined over convex set Z is a convex function when:

$$\forall x \in Z \ \& \ \forall y \in Z \ \& \ \forall \alpha \in [0,1] \quad F(\alpha x + (1-\alpha) y) \leq \alpha F(x) + (1-\alpha) F(y) \tag{1.5.3}$$

The function F is called strictly convex when in (1.5.3) the "$\leq$" sign can be replaced with the "$<$" sign.

The optimisation problems mentioned above are all convex optimisation problems. Convex optimisation problems have a few very important features which simplify the optimisation procedure (see for example Rockafellar, 1970 or Boyd and Vandenberghe, 2004):

- All local minima are global minima;
- The set of global minima is convex;
- When the optimising function is strictly convex, the set of global minima is a single point.

The above properties allow us to use rather simple optimisation procedures such as gradient descent for finding the global minima. Uniqueness of global minima (or convexity of the set of global minima) allows us to avoid difficult situations when gradient descent finds local minima instead of global. There is also a well developed set of more complex approaches designed particularly for convex optimisation problems (see for example Bertsekas, 2015).

It can be noted that regularisation itself does not guarantee convexity of the resulting optimisation problem, therefore non-convex optimisation methods can also be very important. For example variable selection procedure using AIC can be presented as the following minimisation task:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_{L_2}^2 + \lambda \|\beta\|_{L_0} \right), \tag{1.5.4}$$

for some $\lambda > 0$ (if the data generating process is a linear model then λ will depend on the variance of the errors) and where L_0 is a pseudo-norm defined as:

$$L_0(z) = |\{i : z_i \neq 0\}|. \tag{1.5.5}$$

Such a problem appears to be NP-hard (Natarajan, 1995) and often is considered computationally too complex for practical applications. Therefore in many cases such a problem is “relaxed” to the problem:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_{L_2}^2 + \lambda \|\beta\|_{L_p}^p \right), \tag{1.5.6}$$

for some p > 0 and where L_p is defined as:

$$L_p(z) = \left(\sum_i z_i^p \right)^{1/p}. \tag{1.5.7}$$

The LASSO method is a special case with $p = 1$. This is the lowest value of p which makes problem (1.5.6) convex. The LASSO optimisation problem can be solved efficiently by, for example, the Least Angle Regression (LAR) algorithm (Efron et al., 2004), which gives the same computational complexity as OLS (Hastie, Tibshirani, and Friedman, 2009).

There is growing evidence (see Xu et al., 2012 for an overview) that values of p lower than 1 can give better performance and sparsity of (1.5.6) solutions. Despite such problems being NP-hard (Ge, Jiang, and Ye, 2011), these methods have recently captured significant attention. Thus Xu et al. (2012) proposed a solution of (1.5.6) for $p = \frac{1}{2}$. They claim that some practical problems can be solved within reasonable time.

1.6 Smoothing

Smoothing is an approach when data y are presented as a combination, usually a sum of two components:

$$y = s + e \, , \tag{1.6.1}$$

where the above components should have some desired features: the first component s should be "smooth" and the second e should be "small". The first term in (1.6.1) is considered as "smoothed" data, while the second is called an "error" term. Smoothness of the first term can be measured differently and below I will present a few examples of such measurements.

The above informal formulation can be naturally translated to the following minimisation problem:

$$\hat{s} = \underset{s}{\operatorname{argmin}} \left(\|y - s\| + \lambda S(s) \right), \tag{1.6.2}$$

where $e = (y - s)$ is an error term, $\|.\|$ is some function (usually a norm, often the L_2 norm) measuring "distance" between y and s and $S(s)$ is a function measuring the "smoothness" of s. This formulation shows that smoothing can also be presented as a regularised minimisation problem. Depending on the regularisation term $S(s)$, the above problem can become a Tikhonov regularisation problem or its generalisation.

Let us consider as an example penalised regression splines (P-splines) (Eilers and Marx, 1996; Ruppert, Wand, and Carroll, 2003). The s component in this case belongs to a linear span of the following functions (P-spline bases):

$$\begin{aligned} h_j(x) &= x^{j-1}, \quad j = 1, \ldots, m \\ h_{m+k} &= (x - \xi_k)_+^{m-1}, \quad k = 1, \ldots, K \quad , \end{aligned} \tag{1.6.3}$$

where $m-1$ is the degree of the splines, $\{\xi_k\}_{k=1}^K$ is a set of K knots and $(x)_+$ is a function which is zero for negative x and identity everywhere else. Then $s(x)$ can be expressed as:

$$s(x) = \sum_{j=1}^{m} \alpha_j h_j(x) + \sum_{j=1}^{K} \beta_j h_{j+m}(x), \tag{1.6.4}$$

for some vectors $\alpha = [\alpha_1, \ldots, \alpha_m]^T$ and $\beta = [\beta_1, \ldots, \beta_K]^T$, and a P-spline estimate is defined as:

$$\hat{s} = \underset{s}{\operatorname{argmin}} \left(\sum_{i=1}^{N} (y_i - s(x_i))^2 + \lambda \|\beta\|^2 \right), \tag{1.6.5}$$

where β and s are related according to (1.6.4) and N is the number of observations.

Another example is cubic splines or their multi-dimensional extension – thin plate splines. Spline functions also allow for presentation through basis functions. A thin plate spline is a function s which is doubly differentiable and which is the solution of the following problem (Green and Silverman, 1994):

$$\hat{s} = \underset{s}{\operatorname{argmin}} \left(\sum_{i=1}^{N} (y_i - s(x_i))^2 + \lambda \int \left[\frac{\partial^2 s}{\partial x_1^2} + 2 \frac{\partial^2 s}{\partial x_1 \partial x_2} + \frac{\partial^2 s}{\partial x_2^2} \right] \mathrm{d}x_1 \, \mathrm{d}x_2 \right). \tag{1.6.6}$$

Both examples are closely related to the first article presented in Chaper 2 where the approach is extended by presenting a minimisation problem which combines features of both P-splines and thin plate splines to achieve better performance for some special types of two-dimensional data.

1.7 Choosing smoothing parameters via cross validation

Another way to control overall complexity of the model is cross validation. This is one of the key tools used in all three articles (chapters 2, 3 and 4) presented in this thesis. Cross validation measures the "correctness" of a forecasting method by splitting data into training and test sets multiple times and measuring the "closeness" of multiple forecasts with the corresponding test sets.

For example, for method $f(x|\Delta)$, transforming regressor x and training set Δ into a prediction, and data set $D = \{(y_i, x_i)\}_{i=1}^{N}$, consisting of independent pairs of variables y_i and regressors x_i, leave one out cross validation can be defined as (see for example Hastie, Tibshirani, and Friedman, 2009):

$$CV(f|D) = \frac{1}{N}\sum_{i=1}^{N} \|y_i - f(x_i|D_{-i})\|, \tag{1.7.1}$$

where $\|.\|$ is some function used for calculating the distance between y_i and $f(x_i|D_{-i})$ and D_{-i} is the data set excluding the ith pair.

If we assume that f depends on some parameter λ and we deal with the set of methods $F = \{f_\lambda|_{\lambda\in\Lambda}\}$ (Λ is the domain for λ), the best parameter λ can be chosen using the following approach:

$$\hat{\lambda} = \underset{\lambda\in\Lambda}{\operatorname{argmin}}(CV(f_\lambda|D)). \tag{1.7.2}$$

According to Stone (1977) this approach is equivalent to minimisation of AIC when choosing the best subset of regressors.

1.7.1 An example when cross validation fails

Let us consider the case when cross validation is used to find the best parameter $0 \le \lambda \le 1$ for a method $f_\lambda(x|D) \in F \overset{def}{=} \{f_\lambda\}_{\lambda\in[0,1]}$.

To define f_λ, first, we introduce a few operations.

Operation $B(\lambda)$ transforms $0 \le \lambda \le 1$ into a sequence of zeros and ones by presenting λ in the binary format: $\lambda = (0.b_1 b_2 \dots b_n \dots)_2$:

$$\lambda \xrightarrow{\mathrm{B}} \{b_i(\lambda)\}_{i=1}^{\infty}. \tag{1.7.3}$$

Operation Split(λ, j) is defined by the following steps:

1. Find the jth prime number p_j.
2. Present λ as a sequence of zeros and ones using operation B: $\lambda \xrightarrow{\mathrm{B}} \{b_i\}_{i=1}^{\infty}$.
3. Define subsequence s_j of the sequence $\{b_i\}_{i=1}^{\infty}$ from step 2 as: $s_j = \{b_{ip_j}\}_{i=1}^{\infty}$.
4. Define result of Split(λ, j) as a number, which has binary representation s_i: $B(\text{Split}(\lambda, j)) = s_j$.

Finally we can define method $f_\lambda(x|D)$ by the following steps:

1. Find the number of elements in dataset D: $n = |D|$.
2. Using Split operation define n values $\{\lambda_i\}_{i=1}^{n}$, as $\lambda_i = \text{Split}(\lambda, i)$.
3. Define result of $f_\lambda(x|D)$ as: $f_\lambda(x|D) = \sum\limits_{i=1}^{n+1} g(\lambda_i) x^{i-1}$, where $g(\lambda)$ can be any function which maps $(0,1)$ to $(-\infty, \infty)$, for example $g(z) = \ln(1/x - 1)$. If any λ_i take values 0 or 1 the result of $f_\lambda(x|D)$ is not defined.

It is easy to see that:

$$\min_{\lambda}(CV(f_\lambda|D)) \equiv 0. \tag{1.7.4}$$

In other words f is able to fit any dataset perfectly since it can "encode" the information of the whole dataset inside its parameter λ. Therefore the estimated parameter $\hat{\lambda} = \operatorname*{argmin}_{\lambda}(CV(f_\lambda|D))$ does not contain any "generalisation" of the dataset D and any method, which extracts meaningful information from D, for example a linear regression $L(x|D)$, having cross validation score strongly greater than zero, can be better for forecasting than $f_{\hat{\lambda}}(x|D)$, which has "perfect" cross validation score equal to zero.

Finally by adding this linear regression L to the set of methods F we get a new set of methods where cross validation always chooses the worst forecasting method (with the assumption that dataset D contains some information, which can be extracted and used for prediction with a linear model L).

The example, provided above, is far from practice, although it shows the weakness of cross validation under circumstances when the set of estimated functions is extremely "rich". The problem of the above example (which most practical cases do not have) is that the parameter λ represents infinitely many parameters and therefore complexity of the dataset easily "flows" into λ without "extracting" any information useful for forecasting.

In practice we have one or few parameters λ which can not absorb as much information from the dataset and therefore using cross validation is as safe as using a linear regression with one or few predefined regressors for forecasting. On the other hand if the number of parameters λ increases, cross validation can be worse, reminiscent of the situation with linear regression, when it forecasts poorly in the case when too many regressors are used.

Ng (1997) discusses overfitting problems with cross validation and proposes ways to solve them.

1.8 Problems discussed in the thesis

The thesis consists of three rather independent articles although reflecting a common theme by presenting the corresponding problem as a regularized optimization task. The articles are written in collaboration with Professor Rob J. Hyndman.

In the first article we explore some bivariate smoothing methods with partial differential regularizations designed to handle smooth bivariate surfaces with occasional ridges. We apply our technique to smoothing mortality rates. We propose three new practical methods of smoothing mortality rates over age and time. Although our methods are designed to smooth logarithms of mortality rates, they are generic enough to be applied to any bivariate data with occasional ridges.

In the second article 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).

In the last article we propose novel generic models and methods for decomposing seasonal data: STR (a Seasonal-Trend decomposition procedure based on Regression) and Robust STR. In some ways, STR is similar to Ridge Regression and Robust STR can be related to LASSO. Our new methods are much more general than any alternative time series decomposition methods.

References

Akaike, H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aristotle (1959). *Ars Rhetorica*. Ed. by WD Ross. Oxford.

Bartlett, PL and S Mendelson (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482.

Bertsekas, DP (2015). *Convex Optimization Algorithms*. Athena scientific Belmont.

Box, GE and NR Draper (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Boyd, S and L Vandenberghe (2004). *Convex optimization*. Cambridge University Press.

Chaitin, GJ (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM* **13**(4), 547–569.

Crombie, AC (1959). *Medieval and Early Modern Science: Science in the Middle Ages, V-XIII centuries*. Vol. 167. Doubleday Anchor Books.

Efron, B, T Hastie, I Johnstone, R Tibshirani, et al. (2004). Least angle regression. *The Annals of Statistics* **32**(2), 407–499.

Eilers, PH and BD Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 89–102.

Ge, D, X Jiang, and Y Ye (2011). A note on the complexity of L_p minimization. *Mathematical programming* **129**(2), 285–299.

Green, PJ and BW Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Grünwald, PD, IJ Myung, and MA Pitt (2005). *Advances in minimum description length: Theory and applications*. MIT press.

Hastie, T, R Tibshirani, and J Friedman (2009). *The elements of statistical learning*. Springer.

Hayashi, F (2000). *Econometrics*. Princeton University Press.

Hoerl, AE and RW Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.

Kolmogorov, AN (1963). On tables of random numbers. *The Indian Journal of Statistics, Series A*, 369–376.

Natarajan, BK (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing* **24**(2), 227–234.

Ng, AY (1997). Preventing" overfitting" of cross-validation data. In: *ICML*. Vol. 97, pp.245–253.

Rockafellar, RT (1970). *Convex analysis*. Princeton university press.

Ruppert, D, MP Wand, and RJ Carroll (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Solomonoff, RJ (1964a). A formal theory of inductive inference. Part I. *Information and control* **7**(1), 1–22.

Solomonoff, RJ (1964b). A formal theory of inductive inference. Part II. *Information and control* **7**(2), 224–254.

Stone, M (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44–47.

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

Tikhonov, AN (1963). Regularization of incorrectly posed problems. In: *Soviet Math. Dokl.* Vol. 4. 6, pp.1624–1627.

Tikhonov, AN (1943). On the stability of inverse problems. In: *Dokl. Akad. Nauk SSSR*. Vol. 39. 5, pp.195–198.

Vapnik, V (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

Vovk, VG (1988). The law of the iterated logarithm for random Kolmogorov, or chaotic, sequences. *Theory of Probability & Its Applications* **32**(3), 413–425.

Xu, Z, X Chang, F Xu, and H Zhang (2012). $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems* **23**(7), 1013–1027.

Chapter 2

Paper “Bivariate data with ridges: two-dimensional smoothing of mortality rates”

Bivariate data with ridges: two-dimensional smoothing of mortality rates

Alexander Dokumentov
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

4 January 2016

Bivariate data with ridges: two-dimensional smoothing of mortality rates

Abstract

In this article we explore some bivariate smoothing methods with partial differential regularizations designed to handle smooth bivariate surfaces with occasional ridges. We apply our technique to smoothing mortality rates.

Mortality rates are typically smoothed over two dimensions: age and time. Occasional ridges occur due to period effects (e.g., deaths due to wars and epidemics) and cohort effects (e.g., the effects of wars and epidemics on the survivors).

We propose three new practical methods of smoothing mortality rates over age and time. The first method uses bivariate thin plate splines. The second uses a similar procedure but with lasso-type regularization. The third method also uses bivariate lasso-type regularization, but allows for both period and cohort effects. In these smoothing methods, the logarithms of mortality rates are modelled as the sum of four components: a smooth bivariate function of age and time, smooth one-dimensional cohort effects, smooth one-dimensional period effects and random errors. Cross validation is used to compare these new smoothing methods with existing approaches.

Although our methods are designed to smooth logarithms of mortality rates, they are generic enough to be applied to any bivariate data with occasional ridges expected in a few predefined directions.

Keywords: Bivariate data, nonparametric smoothing, mortality rates, graduation, cohort effects, period effects.

1 Introduction

Mortality rates are used to compute life tables, life expectancies, insurance premiums, and other items of interest to demographers and actuaries. However, mortality rates are noisy (for example for young ages when mortality rates are low, or for very old ages when the population is small), and so it is useful to smooth them in order to obtain better estimates with smaller variance. (In demography, smoothing mortality rates is known as "graduation".) Mortality rates also contain occasional non-smooth features, usually due to wars and epidemics, which are manifest as period effects (additional deaths in a particular year) and cohort effects (changed mortality rates of the survivors). These appear as ridges in the otherwise smooth surface, and any effective smoothing methods applied to mortality data need to allow for these features.

Several nonparametric smoothing approaches have been proposed in the past (e.g., Schuette, 1978; Portnoy, 1997; Hyndman and Ullah, 2007; Kirkby and Currie, 2010), but none to our knowledge that exploit all the features of mortality rates. In this paper, we propose several new bivariate smoothing methods for mortality data, the last of which also allows for both period and cohort effects. We compare our new methods, and some existing methods, using a cross-validation procedure.

The methods we propose are extensions and combinations of quantile smoothing splines (see, for example, Koenker et al., 1994; Portnoy, 1997; He et al., 1998) with partial differential regularizations (Sangalli, 2014). While our methods are generally applicable to any bivariate data with ridges, we restrict our discussion here to their application to smoothing natural logarithms of human mortality rates.

Let $M_{x,t}$ denote an observed mortality rate $M_{x,t}$ for a particular age x and for a particular year t, defined as $M_{x,t} = D_{x,t}/E_{x,t}$, where $D_{x,t}$ is the number of deaths during year t for people who died being x years old, and $E_{x,t}$ is the total number of years lived by people aged x during year t. In practice, $E_{x,t}$ is usually approximated by the mid-year population of people aged x in year t.

As we can see from the definition, mortality rates are two dimensional: one dimension is time and the other dimension is age. Our aim is to smooth the bivariate surface in both the age (x) and time (t) dimensions, and to allow occasional ridges due to period effects (along x for a specific t) and cohort effects (along $x = t + k$ for a specific k).

To stabilize the variance of the noise, and to make the smoothness more uniform, it is necessary to take a transformation. While deaths can be considered Poisson (Brillinger, 1986), mortality rates are usually modelled in logarithms: $m_{x,t} = \log(M_{x,t})$ (the choice of natural logarithms is historical, although base 10 logarithms would have better interpretation). See, for example, Lee and Carter (1992), Portnoy (1997) and Hyndman and Ullah (2007). Moreover, features of the data for low mortality rates (for ages from 1 to 40) have clearer shape after taking logarithms. Taking logarithms also makes sense from the point of view that different factors affect mortality in a multiplicative manner, and after taking logarithms the effects are then additive. Finally, mortality rates range over several orders of magnitude, so taking logarithms allow different parts of the data to play a comparable role.

Mortality data in practice is usually available at regular grid points. In this article we have observations at every year in time and age dimensions. The data is available as $q \times T$ matrix m:

$$m = [m_{j,k}],$$

where $j \in [1 \dots q]$ and $k \in [1 \dots T]$, $\;q$ is the number of age groups and T is the number of years where the data is available.

Equivalently it can be presented as a set of points:

$$\{x_i, t_i, m_{x_i,t_i}\}_{i=1}^{n},$$

where $n = qT$, $\;m_{x_i,t_i}$ is the logarithm of mortality rate which corresponds to x_i age group and t_i year.

We also assume that the data is of the form:

$$m_{x,t} = f(x,t) + \epsilon_{x,t},$$

where $m_{x,t}$ is logarithm of mortality rate for age group x and time t, $f(x,t)$ is a smooth function apart from ridges and $\epsilon_{x,t}$ is “noise”.

Figure 1 shows log mortality rates for females in France from 1950 to 1970 (the choice of country and the gender is rather arbitrary, and we provide calculations for 11 countries and both genders in the Appendix). The data is taken from the demography package for R (Hyndman, 2014); it was originally sources from the Human Mortality Database (2008).

After taking logs the following features of the log mortality surface become evident:

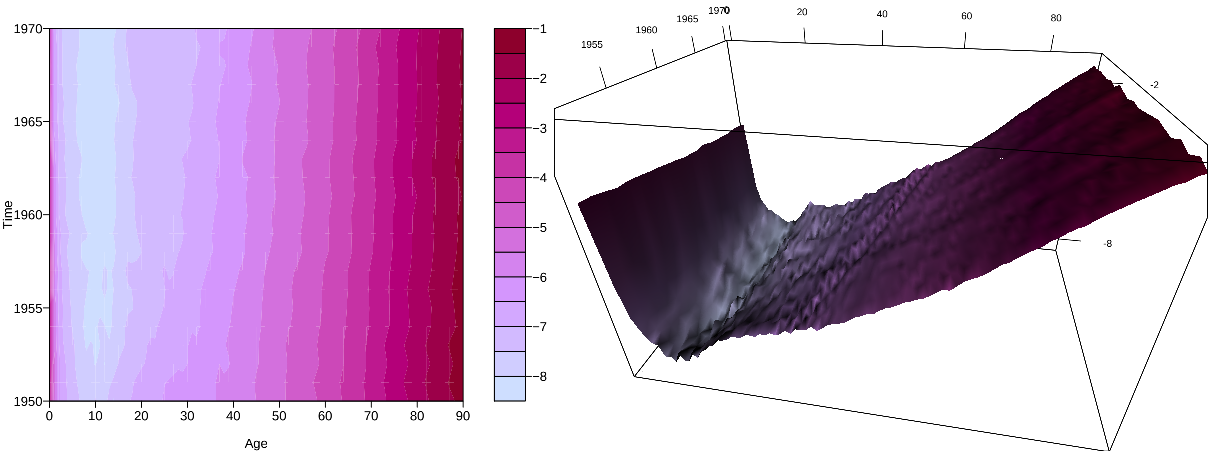


Figure 1: *Natural logarithms of French female mortality rates.*

- In the age dimension, the log mortality decreases rapidly for the early ages and reaches a minimum at age about 10 years. There is a "bump" around age 20, after which it increases almost linearly to the very old ages.
- For almost every age, mortality has decreased with time over the period of these data. It decreases more steeply (on the log scale) for younger ages then for older ages.
- The highest mortality is reached for older ages.
- There are diagonal ridges due to cohort effects. While these are not easy to see in Figure 1, later graphs will highlight their existence. Such patterns may be due to some extreme events experienced by a cohort of people when they were born, or in their early years.
- There are horizontal patterns (for a fixed year) due to period effects. Such patterns are usually due to some extreme environmental event such as a war or pandemic, affecting all people (with different magnitude) during a particular year. These are more evident during the periods 1914–1918 and 1939–1945 than in the years shown in Figure 1.

Each of these features should be preserved when smoothing mortality data because they represent "real" effects.

A classic method for smoothing mortality rates is to use a parametric non-linear function of age dependent on only a few coefficients. Heligman and Pollard (1980) were among the first to propose a formula which covers all living ages. Having such a function, the smoothing can done by simply estimating the coefficients using, for example, least squares. This approach is one-dimensional (in x) but can be extended to two dimensions relatively easily. We do not follow this path because we do not want to be restricted to features easily described parametrically; we want the freedom to model any data features that appear. Forfar et al. (1988) discusses some alternative one-dimensional parametric approaches for smoothing mortality rates.

Hyndman and Ullah (2007) proposed a one-dimensional non-parametric approach for smoothing, based on penalized regression splines with a monotonicity constraint. We discuss this approach in more detail in Section 2.

In normal circumstances (when wars and pandemics are relatively uncommon), it is reasonable to assume that mortality rates are smooth in two dimensions: time and age, and so the main idea behind bivariate smoothing is to use both dimensions for estimating mortality at some two-dimensional point defined by time and age. The one-dimensional approach is (usually) to find a smooth function $f_t(x) = \mathrm{E}[m_{x,t}]$ for each t. In contrast, bivariate smoothing looks for a smooth bivariate function $f(x,t) = \mathrm{E}[m_{x,t}]$. By allowing the assumption of smoothness in both dimensions, better performance should be possible due to the additional information included in the estimation.

Currie et al. (2004) proposed using two-dimensional P-splines for smoothing and forecasting mortality rates. Camarda (2012) implemented this approach along with a one-dimensional version in the R package MortalitySmooth. The latter authors acknowledge that this method is not designed to work efficiently for ages younger than 10.

Kirkby and Currie (2010) extended the approach of Currie et al. (2004) to take into account period effects. The method is based on a Poisson model of deaths using a GLM for estimation, with the period effects estimated in a multi-step procedure. Being an extension of the Currie et al. (2004) approach, this method also lacks the ability to estimate mortality rates efficiently for young ages.

Camarda et al. (2010) proposed to use special bases for P-splines to fit logarithms of mortality rates. This helps to overcome problems related to the steep slope at early ages (which is the obstacle for the methods described in Currie et al. (2004) and Kirkby and Currie (2010)). Unfortunately, the authors have not made available any implementation of their method.

To solve the problem of abrupt changes in the data, we prefer the L_1 norm in place of the L_2 norm. We use it for regularization as well as a measure of the closeness of the approximation to the data. The L_1 norm is often used because it is robust when the data contain outliers (see, for example, Schuette, 1978; Portnoy, 1997). While this is a useful feature, the main reason we adopt the L_1 norm here is different.

When the L_2 norm is applied to second derivatives in order to regularize the estimated function, abrupt changes in the data will be over-smoothed. This occurs because the L_2 norm penalizes large and abrupt changes much more heavily than smaller and smoother changes. Consequently abrupt features such as ridges are distorted while attempting to reduce noise.

In contrast, the L_1 norm does not possess such a feature: the "cost" of a single big change is exactly the same as sum of the "costs" of smaller changes with the same sign and combined magnitude equal to the magnitude of the big change. Therefore, there is no tendency to over-smooth abrupt features in the surface.

In the next five sections, we describe five smoothing algorithms, the last of which is our preferred procedure:

1. The Hyndman and Ullah (2007) algorithm (Section 2), is implemented in the demography R package (Hyndman, 2014), and smooths mortality rates only in the age dimension. This algorithm is presented for comparison only.
2. The Camarda (2012) algorithm (Section 3), is implemented in the MortalitySmooth package, and smooths mortality rates in both dimensions, although it is only designed to work for ages greater than 10. This algorithm is also presented for comparison only.
3. Section 4 describes a new algorithm that uses two-dimensional thin plate splines and therefore uses both dimensions — time and age — for smoothing.
4. Section 5 describes another new algorithm that uses Lasso-type regularization and also uses both dimensions for smoothing. This algorithm copies thin plate splines in many ways, but it uses the L_1 norm instead of the L_2 norm.
5. The last algorithm (Section 6) also uses Lasso-type regularization and both dimensions for smoothing, but incorporates ridges to account for cohort and period effects. This improves the performance and also provides greater insight to the structure of the mortality data.

The minimisation problem of the last two algorithms can be reduced to quantile regression minimisation problems (see Section 5), and therefore we use quantile regression software (Koenker, 2015) to implement these algorithms.

These five algorithms are compared in Section 7 using a cross-validation procedure. Finally, we provide some discussion and conclusions in Section 8.

2 Hyndman-Ullah (2007) method

Hyndman and Ullah (2007) proposed a method for smoothing mortality rates across ages in each year. The method is intentionally one-dimensional to allow for a forecasting procedure, applied after smoothing, that takes into account variation in the time dimension. An implementation of the method is provided in the demography package for R (Hyndman, 2014).

This smoothing method uses constrained weighted penalized spline regression applied independently for each year. Weighted penalized spline regression involves calculating a vector β of length k which minimizes the expression

$$\|w(y - \Xi\beta)\|^2 + \lambda^2 \beta^T D\beta,$$

where y is a vector of observations of length n, Ξ is an $n \times k$ matrix representing k linear spline bases, $D = \text{diag}(0,0,1,1,...,1)$ is an $k \times k$ diagonal matrix, w is a vector of weights of length n and λ is a scalar parameter (see, for example, Ruppert et al., 2003).

In the case of smoothing mortality rates, observations in some arbitrary year t are given by $y_i = m_{x_i,t}$ for age group x_i years old ($i \in [1,\ldots,n]$). The weights w_i are taken as the inverse of the estimated variances of y_i. Assuming deaths follow a Poisson distribution, and using a Taylor series approach, Hyndman and Booth (2008) estimate the variance of y_i as $\sigma_i^2 \approx (E_{x_i,t} M_{x_i,t})^{-1}$, where $E_{x_i,t}$ is the mid-year population of people aged x_i years in year t.

Moreover such splines are constrained to ensure that the resulting function $f(x)$ is monotonically increasing for $x \geq c$ for some c (for example 50 years). Hyndman and Ullah (2007) use a modified version of the method described in Wood (1994) to implement this constraint.

The result of this approach is a surface which is smooth in the age dimension but still “wiggly” in the time dimension (Figure 2).

The residuals (Figure 3) show some serial correlation for early ages as well as diagonal ridges which are cohort effects (effects related to people born in the same year). For example Figure 4 reveals some serial correlation of the residuals for ages 1 and 2. However, it is clear that the residuals do not show any horizontal patterns due to period effects. This is expected, because separate smoothing has been done independently for each year.

3 Camarda (2012) method

Camarda (2012) implements a two-dimensional method using P-splines for smoothing mortality rates in the MortalitySmooth package for R. For ages 0 to 10, the result of smoothing is notably biased (Figures 5 and 6). As we can see in Figure 6 the residuals are serially correlated for early ages. Also diagonal ridges due to cohort effects and horizontal ridges due to period effects are visible.

Camarda (2012) acknowledge that the method was not designed for smoothing of the youngest ages. Nevertheless, we still use it as the comparisons clearly show what problems we faced and have overcome.

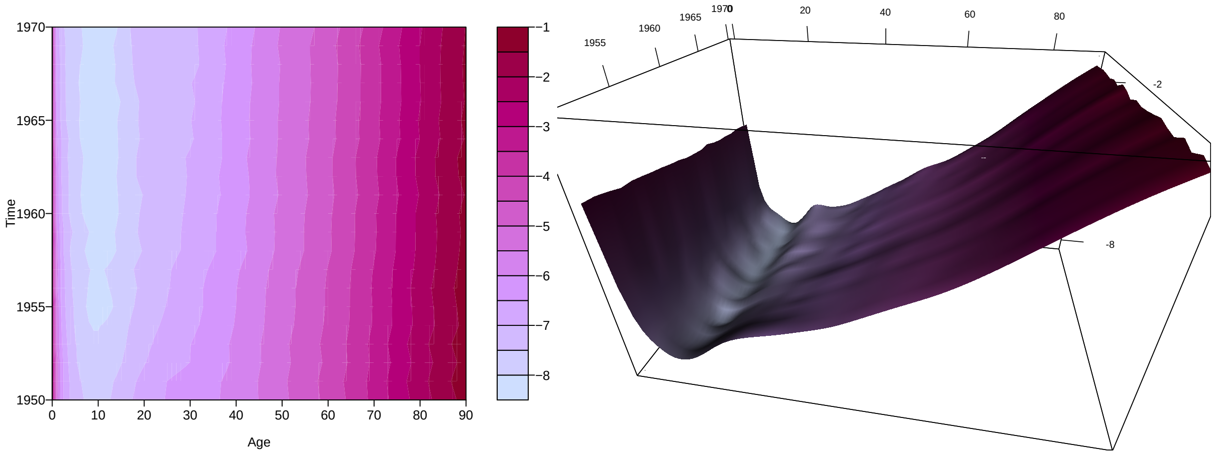


Figure 2: *French female mortality rates smoothed by Hyndman and Ullah (2007) method.*

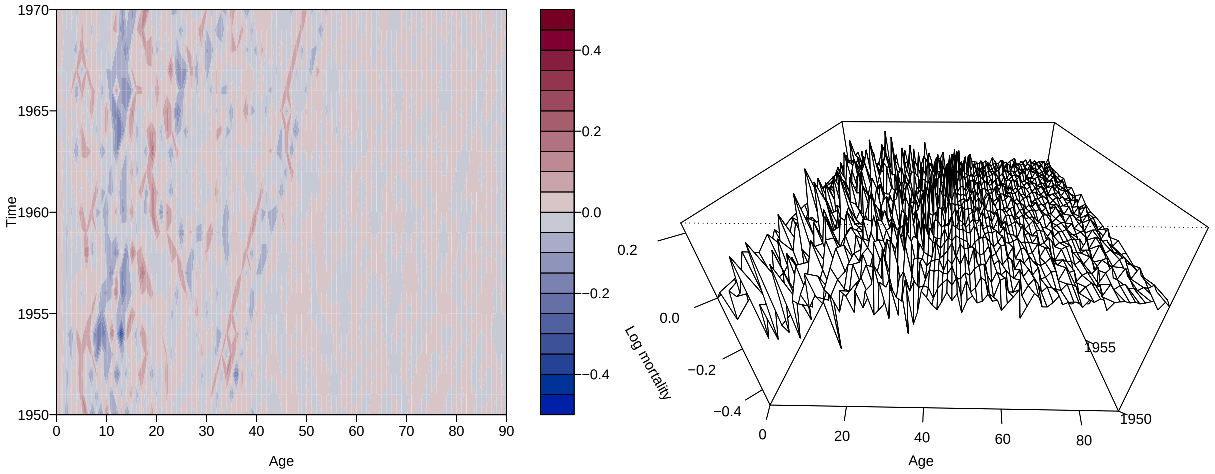


Figure 3: *French female mortality rates residuals after smoothing by Hyndman and Ullah (2007) method.*

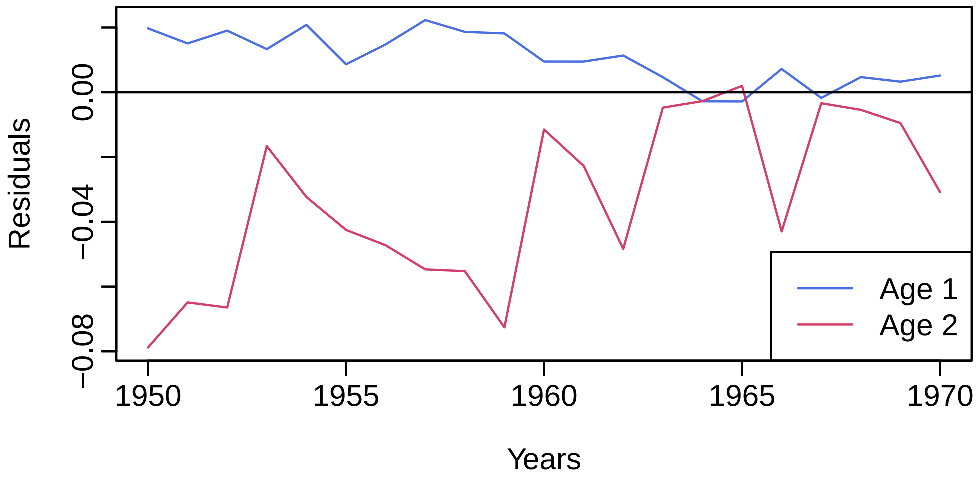


Figure 4: *French female mortality rates residuals for ages 1 and 2 after smoothing in age dimension.*

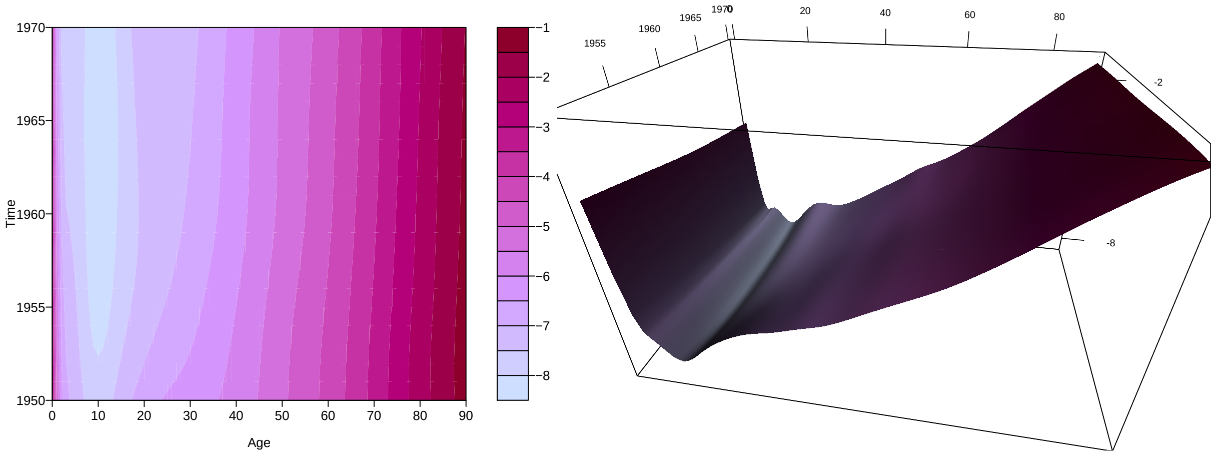


Figure 5: *French female mortality rates smoothed by Camarda (2012) method.*

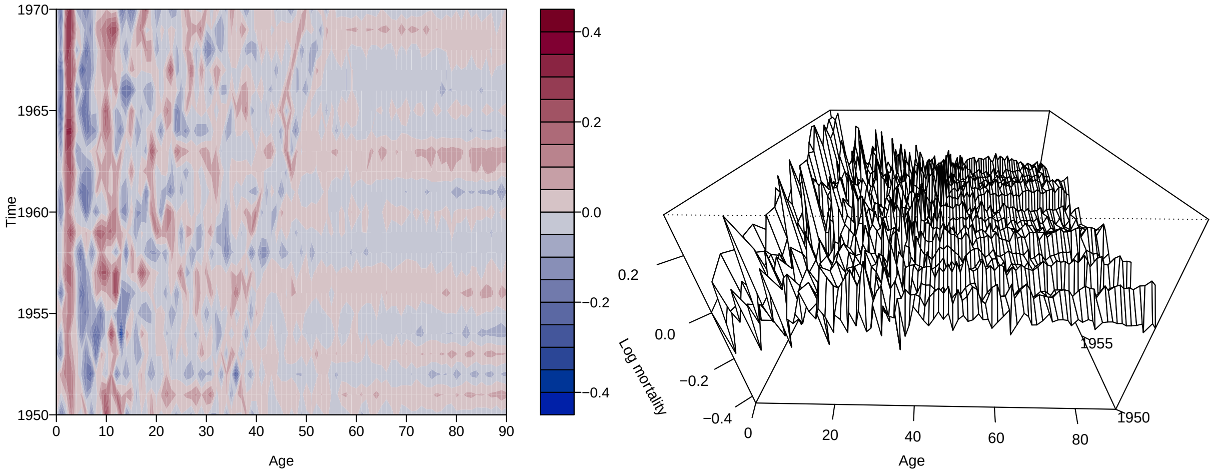


Figure 6: *French female mortality rates residuals after smoothing by Camarda (2012) method.*

4 Boosted thin plate splines

Thin plate splines work well in many cases (Wood, 2006). However, in case of mortality data, direct application of thin plate splines (or adaptive thin plate splines where flexibility varies) to the logs of mortality rates does not lead to precise and unbiased results, especially for early ages. This is the same problem that the Camarda (2012) method experiences for early ages.

We speculate that the reasons for the problems are twofold:

(a) The log mortality rate surface is very steep at early ages, and twisted along the time dimension due to more rapid decrease in mortality for younger ages compare to older ages; and

(b) thin plate splines penalize big errors much heavier than small errors.

This leads to the situation when abrupt jumps in the mortality data generate errors in the proximity of the jumps, causing unsatisfactory performance of thin plate splines over the abrupt surface.

Consequently, improved performance is possibly by first "flattening" the surface and then applying adaptive thin plates splines to the flattened surface before reversing the flattening procedure. As a result the final model becomes a thin plate spline model (in terms of Camarda (2012)) applied to a smooth static surface.

Thus the observations $m_{x,t}$ can be presented as a sum of three components:

$$m_{x,t} = s(x,t) + r(x,t) + \epsilon_{x,t},$$

where $s(x,t)$ is a bivariate surface which is linear in time dimension and smooth in age dimension, $r(x,t)$ is a bivariate surface presented by thin plate splines and $\epsilon_{x,t}$ is "noise". We use a "flattening procedure" described below to obtain estimate of surface $s(x,t)$ and then apply a thin plate spline model to the residuals to estimate $r(x,t)$ and eventually $\epsilon_{x,t}$.

The inputs for the flattening procedure are the log mortality rates (the "data") from year t_0 to year t_1 and for age groups from 1 to q. We denote these inputs as $m_{x,t}$, where x is the age group and t is the time ($1 \le x \le q,\ \ t_0 \le t \le t_1$). We split the data into two halves with the earliest years of observation in the first set of data, and the later years in the second set of data. We then estimate a very crude surface based on the median log mortality by age for these two sets of data.

This crude surface is subtracted from the data, and the resulting differences are then smoothed using thin plate splines.

This has some similarities to the "twicing" procedure proposed in Tukey (1977). It can also be considered as rudimentary two-step "boosting" (see, for example, Bühlmann and Yu, 2003).

We now outline the procedure.

1. The mid-point year is at time $t_{0.5} = (t_0 + t_1)/2$. Data where $t \le t_{0.5}$ will belong to set one, and the remaining data will belong to set two. The age-specific medians for the two halves of the data are given by

$$m^*_{x,0} = \underset{t_0 \le t \le t_{0.5}}{\text{Median}}(m_{x,t}) \qquad \text{and} \qquad m^*_{x,1} = \underset{t_{0.5} < t \le t_1}{\text{Median}}(m_{x,t}).$$

2. Each of the two sets of medians — for older and recent years — is smoothed using a standard smoothing method. Let us denote the resulting curves by $m_0(x)$ and $m_1(x)$.

3. These two curves are used to create an approximate, but very smooth surface for years from t_0 to year t_1, by connecting points corresponding to same age in the smoothed curves by straight lines such that the resulting curve is a linear function in time dimension for any particular age and functions $m_0(x)$ and $m_1(x)$ for years $t_0 + \frac{t_1 - t_0}{4}$ and $t_1 - \frac{t_1 - t_0}{4}$ correspondingly:

$$s(x,t) = \frac{(3t_1 + t_0 - 4t)m_0(x) + (4t - 3t_0 - t_1)m_1(x)}{2(t_1 - t_0)}.$$

4. The smooth surface $s(x,t)$ is subtracted from the original data to give $r_{x,t} = m_{x,t} - s(x,t)$. These flattened values lie on a surface that is not abrupt at early ages and is not twisted.

5. The flattened values $r_{x,t}$ are smoothed using adaptive thin plate splines to give $r(x,t)$.

6. The last step is to "un-flatten" the result obtained in the previous step: $f(x,t) = r(x,t) + s(x,t)$.

The resulting smooth surface is shown in Figure 7. The residuals, $m_{x,t} - f(x,t)$, are shown in Figure 8. They now look more uncorrelated than in Figures 3 and 6. Remaining diagonal ridges due to cohort effects, and remaining horizontal ridges due to period effects, are still clearly visible.

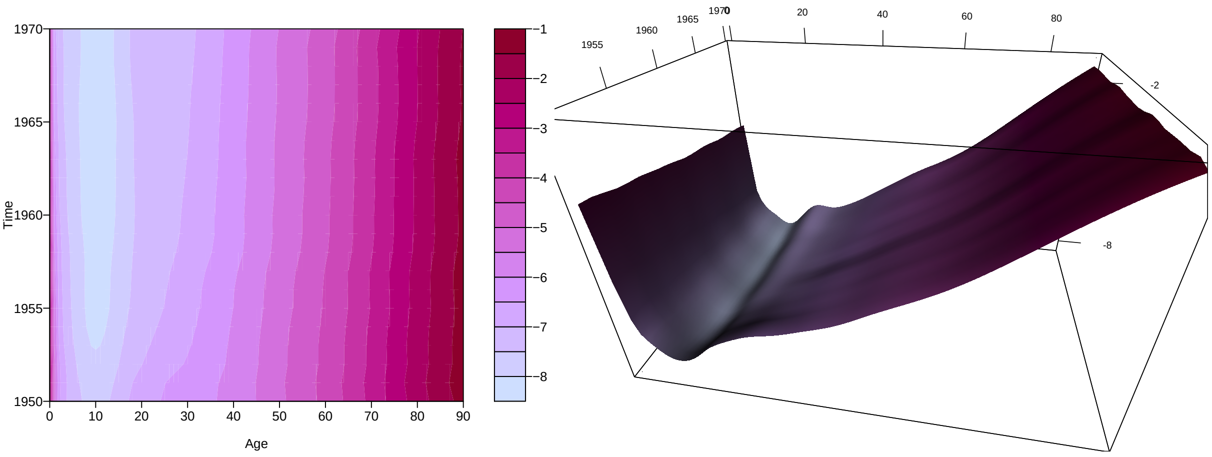


Figure 7: *French female mortality rates smoothed with thin plate splines.*

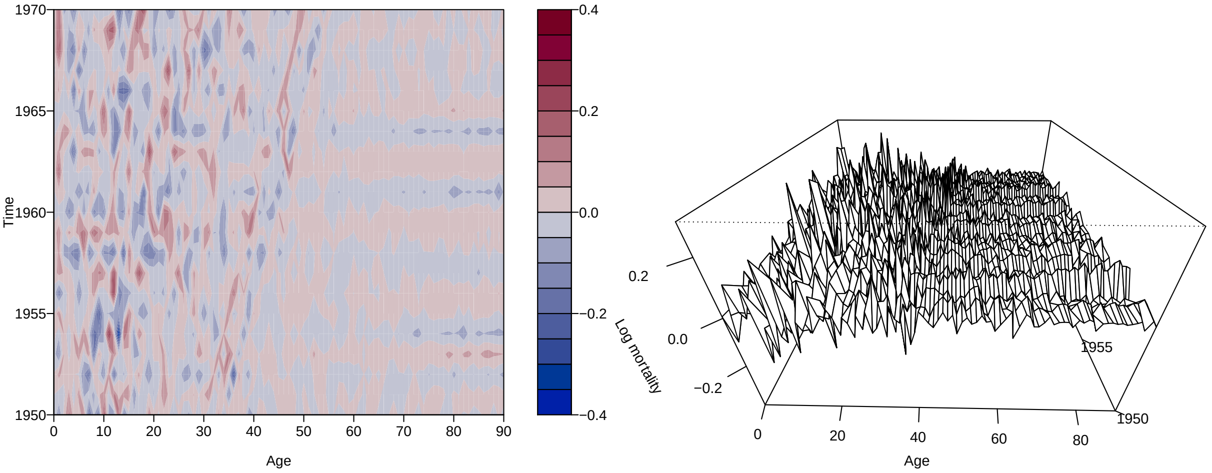


Figure 8: *French female mortality rates residuals after smoothing with thin plate splines.*

5 Quantile Lasso smoothing

A two dimensional thin plate spline is defined as the function $f(x,t)$ which minimises

$$J\left(\{y_i\}_{i=1}^n, f\right) = \sum_{i=1}^n (y_i - f(x_i,t_i))^2 + \lambda \int \left[\left(\frac{\partial^2 f}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial t}\right)^2 + \left(\frac{\partial^2 f}{\partial t^2}\right)^2 \right] \mathrm{d}x\,\mathrm{d}t$$

for some smoothing parameter $\lambda > 0$, knots $\{(x_i,t_i)\}_{i=1}^n$ and values $\{y_i\}_{i=1}^n$ (see for example Wood, 2006).

If the knots form a fine regular grid, then the integral can be approximated by a sum and so $J(\{y_i\}_{i=1}^n, f)$ can be approximated as

$$J\left(\{y_i\}_{i=1}^n, f\right) \approx \sum_{i=1}^n (y_i - f(x_i,t_i))^2 + \frac{\lambda}{n}\sum_{i=1}^n \left[\left(\frac{\partial^2 f}{\partial x^2}(x_i,t_i)\right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y}(x_i,t_i)\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}(x_i,t_i)\right)^2\right].$$

Also if the knots form a fine regular grid, then the second partial derivatives at knots can be approximated as linear combinations of function values at nearby knots.

Denoting $\{y_i\}_{i=1}^n$ as vector y and $\{f(x_i,y_i)\}_{i=1}^n$ as vector z, then $J(\{y_i\}_{i=1}^n, f)$ can be approximated as

$$J(y,z) \approx \|y-z\|_{L_2}^2 + \frac{\lambda}{n}\left(\|D_{xx}z\|_{L_2}^2 + 2\|D_{xt}z\|_{L_2}^2 + \|D_{tt}z\|_{L_2}^2\right)$$

where D_{xx}, D_{xt} and D_{tt} are linear operators (matrices) which calculate approximations of vectors $\left\{\frac{\partial^2 f}{\partial x^2}(x_i,t_i)\right\}_{i=1}^n$, $\left\{\frac{\partial^2 f}{\partial x \partial t}(x_i,t_i)\right\}_{i=1}^n$ and $\left\{\frac{\partial^2 f}{\partial t^2}(x_i,t_i)\right\}_{i=1}^n$.

Using the above expression, we can approximate a thin plate spline computed at its knots as

$$S(y) = \arg\min_z\left(\|y-z\|_{L_2}^2 + \frac{\lambda}{n}\left(\|D_{xx}z\|_{L_2}^2 + 2\|D_{xt}z\|_{L_2}^2 + \|D_{tt}z\|_{L_2}^2\right)\right).$$

In the case of smoothing mortality rates, y becomes the data vector containing log mortality rates (two-dimensional data packed as vector). The order of packing affects only the representation of matrices D_{xx}, D_{xt} and D_{tt}.

Following Schuette (1978), we now replace the L_2 norm with the L_1 norm to give smoothing with the Quantile Lasso. In addition, we use three different λ coefficients before every derivative to separately adjust the influence of each derivative on the smoothing. Therefore in this method we define smoothing as $Q(y) = \arg\min_z(K(y,z))$ where

$$K(y,z) = \|y - Mz\|_{L_1} + \lambda_{xx}\|D_{xx}z\|_{L_1} + \lambda_{xt}\|D_{xt}z\|_{L_1} + \lambda_{tt}\|D_{tt}z\|_{L_1}$$

and y, D_{xx}, D_{xt}, D_{tt} are as described above. Since we use the same number of knots and data points, each positioned at the same places, our matrix M becomes an identity matrix.

For this approach, we do not use the "flattening" procedure described in Section 4. While it would be possible to use it, the effect is negligible, and so for simplicity we have not included it.

Minimization of $K(y,z)$ appears to be difficult, but due to a well known procedure (described for example in Wood, 2006) the problem can be reduced to a quantile regression problem, which then can be solved with existing software (Koenker, 2015). In this study we adopt the following reduction procedure:

- Matrices M, $\lambda_{xx}D_{xx}$, $\lambda_{xt}D_{xt}$, and $\lambda_{tt}D_{tt}$ are stacked on top of each other to give $R = [M', \lambda_{xx}D'_{xx}, \lambda_{xt}D'_{xt}, \lambda_{tt}D'_{tt}]'$.
- Vector y is extended by zeros until its length is equal to the number of rows in R: $y_{ext} = [y', 0']'$.
- $K(y,z) = \|y - Mz\|_{L_1} + \lambda_{xx}\|D_{xx}z\|_{L_1} + \lambda_{xt}\|D_{xt}z\|_{L_1} + \lambda_{tt}\|D_{tt}z\|_{L_1}$ is replaced with the equivalent expression $K(y,z) = \|y_{ext} - Rz\|_{L_1}$.

Then finding $Q(y) = \arg\min_{z}(K(y,z))$ is a quantile regression problem.

The smoothing method described above is defined for some fixed parameters λ_{xx}, λ_{xt} and λ_{tt}, which need to be optimised to get maximum performance. As a measure of performance we use the predictive ability of the procedure, estimated using the mean absolute error based on five-fold cross validation (see Hastie et al. (2008) for an example of using five-fold cross validation). The function measuring performance depends on parameters λ_{xx}, λ_{xt} and λ_{tt}. This function may have many local minima which makes the process of finding optimal parameters difficult. We optimize parameters λ_{xx}, λ_{xt} and λ_{tt} using the optimization procedure "malschains" (Bergmeir et al., 2012) which tends to avoid local minima and therefore has a greater chance to find a global optimal solution than standard gradient descent algorithms.

Every subset of data has about 20% of missing values and they each have the same (but shifted) pattern (Figure 9). The points are missed in a regular pattern to ensure the distance between them is as large as possible. Assuming that distant points affect smoothed value less than close points, the result is a fair compromise between closeness to leave-one-out cross validation and good computational time. Clearly such pseudo leave-one-out cross validation requires about five times more resources (processor time) comparing to a single smoothing task over the whole data set.

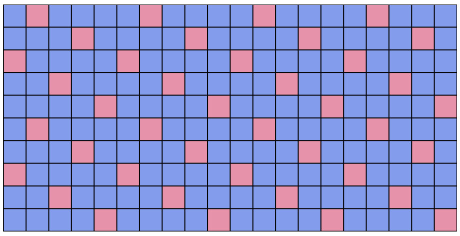


Figure 9: *Missed data pattern for pseudo leave-one-out cross validation.*

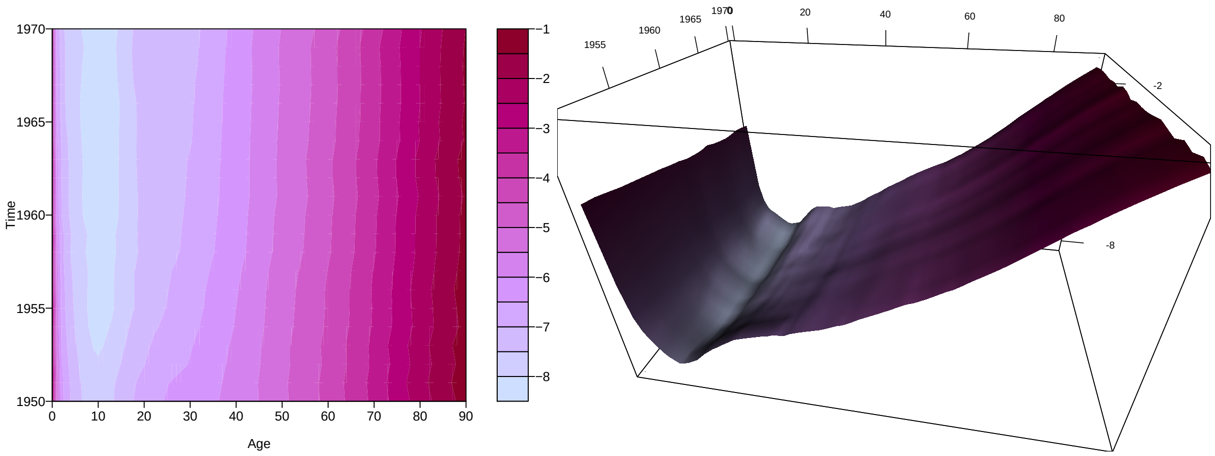


Figure 10: *French female mortality rates smoothed with the Quantile Lasso.*

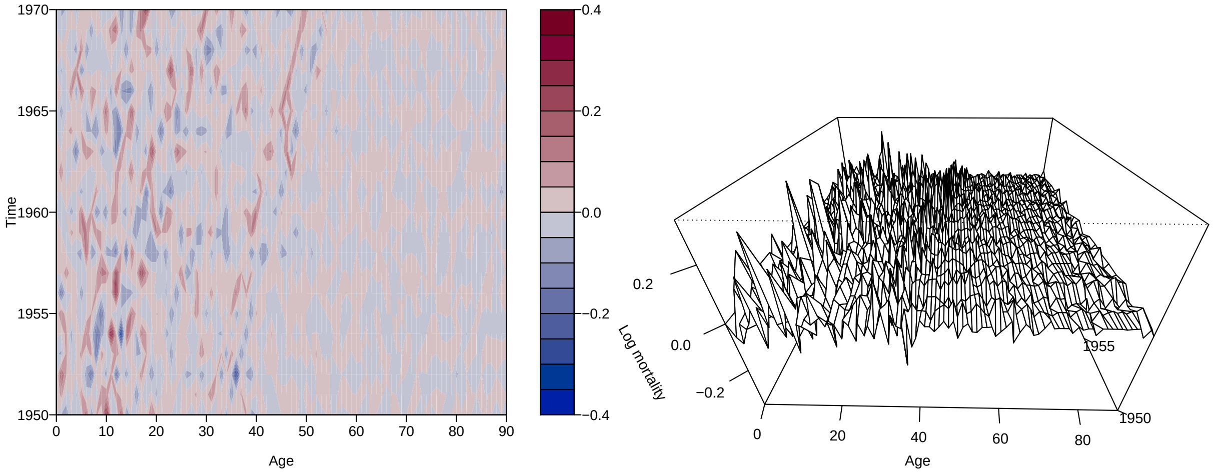


Figure 11: *French female mortality rates residuals after smoothing with the Quantile Lasso.*

The result of smoothing is shown in Figure 10. It is less "smooth" than the results of the previous two methods. Nevertheless we will see in Section 7 that this smoothing method reflects "features" of the data more precisely than two previous smoothing methods.

The residuals are shown in Figure 11. Visually it is difficult to find any serial correlation in the errors. However, cohort and period effects are clearly visible.

6 Quantile Lasso smoothing with cohort and period effects

The cohort and period effects seen in Figure 11 suggest that the smoothing model can be improved by incorporating these features explicitly. This leads us to our new and final smoothing method in which we first apply the Quantile Lasso algorithm of the previous section, and then identify and incorporate significant period and cohort effects. We call this the SMILE method: Smooth Mortality Involving Lasso and period and cohort Effects.

To identify the period and cohort effects, we compute the residuals from the Quantile Lasso smoothing algorithm described in the previous section. Then we split the matrix of the residuals into a set of vectors (of different length) representing diagonals, and carry out the following tests over each diagonal.

1. Perform two-sided t-tests of the residuals over all diagonals to find diagonals with residual mean values significantly different from zero.
2. Perform one-sided sample correlation tests for residual diagonals to find diagonals with positively correlated errors (every diagonal is tested for serial correlation with lag one).

To identify the period effects, we perform the same procedure over every column (representing the same year) of the residuals:

1. Perform two-sided t-tests of the residuals over all years to find years with residual mean values significantly different from zero.
2. Perform one-sided sample correlation tests to find years with positively correlated errors (residuals for every particular year are tested for serial correlation with lag one along age dimension).

Since we run multiple tests, with high probability some of them will give false positive results. We accept such behaviour since the vast majority of the tests will test data correctly and they will improve the performance more than the minority of false positive tests spoil it. Therefore overall we expect greater performance after such a procedure. Moreover the procedure reduces the sizes of the matrices in computations, which makes computations faster.

The new smoothing model involves summing four components: smooth mortality rates, effects being non zero only along the diagonals identified in tests 1 and 2 above, period effects being non zero only along years identified in tests 3 and 4 above, and the noise. These four components are estimated using the following model:

$$Q(y) = \operatorname*{arg\,min}_{z_{sm}, z_{coh}, z_{long}} \left(K(y, z_{sm}, z_{coh}, z_{long})\right),$$

where

$$K(y, z_{sm}, z_{coh}, z_{long}) = ||y - (z_{sm} + z_{coh} + z_{long})||_{L_1} + \lambda_{xx}||D_{xx}z_{sm}||_{L_1} + \lambda_{xt}||D_{xt}z_{sm}||_{L_1} + \lambda_{tt}||D_{tt}z_{sm}||_{L_1} \\ + \lambda_{coh}||D_{coh}z_{coh}||_{L_1} + \theta_{coh}||z_{coh}||_{L_1} + \lambda_{long}||D_{long}z_{long}||_{L_1} + \theta_{long}||z_{long}||_{L_1};$$

- y, D_{xx}, D_{xt} and D_{tt} are as described above;
- z_{sm}, z_{coh} and z_{long} are estimated components representing respectively smooth mortality surface, cohort effects restricted to some diagonals and period effects restricted to some years;
- D_{coh} is a linear differentiation operator representing a discrete version of the second directional derivative in the direction of vector $(1,1)$;
- $D_{long} = D_{tt}$ is a linear differentiation operator representing a discrete version of the second derivative along the years axis;
- $\lambda_{xx}, \lambda_{xt}$, and λ_{tt} are parameters responsible for the smoothness of the mortality surface;
- λ_{coh} and λ_{long} are parameters responsible for the smoothness of the cohort effects and the period effects respectively;
- θ_{coh} and θ_{long} are parameters responsible for shrinking (respectively) the cohort effects and the period effects towards zero.

It may appear that components z_{sm}, z_{coh} and z_{long} duplicate each other. However, this is not the case because λ_{coh} and λ_{long} are restricted (by setting constraints in the optimisation procedure) to values much greater than values of parameters λ_{xx}, λ_{xt} and λ_{tt}. Such restrictions make it difficult for z_{sm}, z_{coh} and z_{long} to compete for the same features in the data. High values of λ_{coh} and λ_{long} make z_{coh} and z_{long} tend to reflect trends along the diagonals and years. This occurs because “features” which have greater effect on z_{coh} and z_{long} are “one dimensional” (narrow and long along diagonals and ages correspondingly), although “features” which have greater effect on z_{sm} are “two dimensional” (they make a shape which is not narrow in any direction, for example a circle).

It is also worth mentioning that the above tests for cohort and period effects are done only for the purpose of reducing computational complexity of the minimization problem. The multiple testing that is carried out means that the selected cohort and period effects are not necessarily statistically significant overall. Some or all of these cohort and period effects will be dropped in the subsequent minimization.

As in the previous section, to minimize $K(y, z_{sm}, z_{coh}, z_{long})$, we use the corresponding quantile regression problem in which:

- vectors z_{sm}, z_{coh} and z_{long} are stacked on top of each other as a single vector, $z_{ext} = [z'_{sm},\ z'_{coh},\ z'_{long}]'$;
- matrices I, $\lambda_{xx}D_{xx}$, $\lambda_{xt}D_{xt}$, $\lambda_{tt}D_{tt}$, $\lambda_{coh}D_{coh}$, $\lambda_{long}D_{long}$, $\theta_{coh}I$, and $\theta_{long}I$ are combined in one matrix,
$$R = \begin{bmatrix} I & I & I \\ \lambda_{xx}D_{xx} & 0 & 0 \\ \lambda_{xt}D_{xt} & 0 & 0 \\ \lambda_{tt}D_{tt} & 0 & 0 \\ 0 & \lambda_{coh}D_{coh} & 0 \\ 0 & \theta_{coh}I & 0 \\ 0 & 0 & \lambda_{long}D_{long} \\ 0 & 0 & \theta_{long}I \end{bmatrix};$$
- vector y is extended by zeros to have its length equal to the number of rows in R: $y_{ext} = [y',\ 0']'$;
- and
$$K(y, z_{sm}, z_{coh}, z_{long}) = ||y-(z_{sm}+z_{coh}+z_{long})||_{L_1} + \lambda_{xx}||D_{xx}z_{sm}||_{L_1} + \lambda_{xt}||D_{xt}z_{sm}||_{L_1} + \lambda_{tt}||D_{tt}z_{sm}||_{L_1} + \lambda_{coh}||D_{coh}z_{coh}||_{L_1} + \theta_{coh}||z_{coh}||_{L_1} + \lambda_{long}||D_{long}z_{long}||_{L_1} + \theta_{long}||z_{long}||_{L_1}$$
is replaced with the equivalent expression $K(y, z_{ext}) = ||y_{ext} - Rz_{ext}||_{L_1}$.

Then $Q(y) = \underset{z_{ext}}{\arg\min}(K(y, z_{ext}))$ is a quantile regression problem.

The parameters $\lambda_{xx}, \lambda_{xt}$, λ_{tt}, λ_{coh}, λ_{long}, θ_{coh} and θ_{long} (7 parameters) are estimated using cross validation. It is relatively few parameters compare to the number of observations (usually thousands) and they can be chosen reasonably well.

The resulting smoothed surface z_{sm} is shown in Figure 12. The cohort effects have been completely removed, but there are the shadows of some period effects remaining.

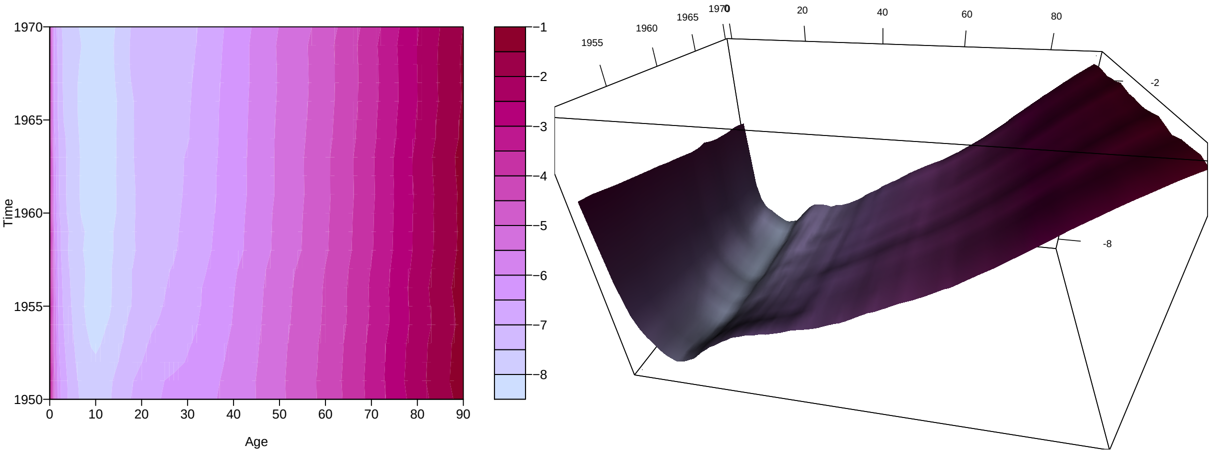


Figure 12: *Logarithms of French female mortality rates smoothed with the SMILE method.*

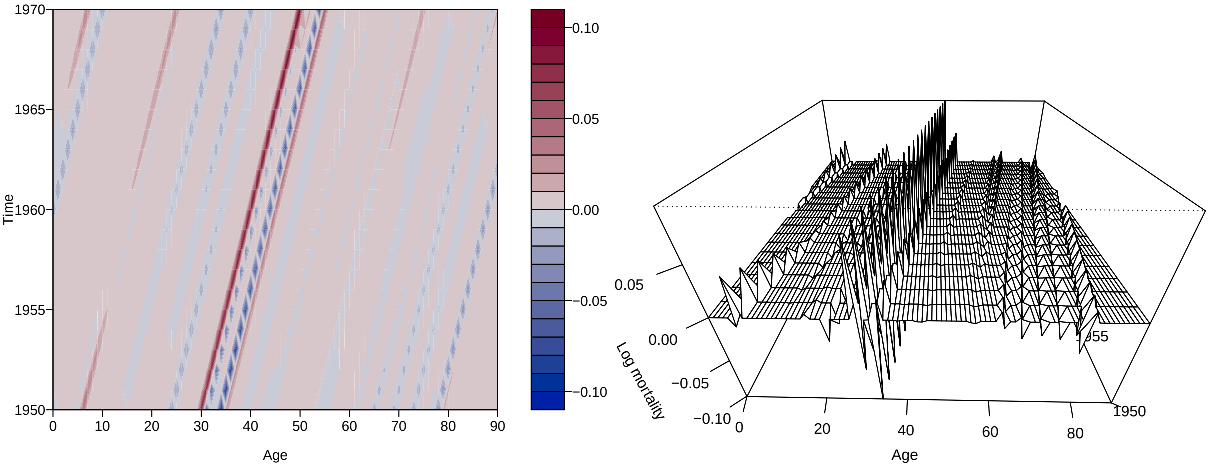


Figure 13: *Cohort effects of logarithms of French female mortality rates.*

The estimated cohort effects z_{coh} are shown in Figure 13. The strongest effects shown are starting at ages and years: (0, 1960), (6, 1950), (30, 1950), (31, 1950), (34, 1950), (35, 1950), (78, 1950).

We speculate that the most visible cohort effects, starting at year 1950 for people aged 30–35, are due to the Spanish flu pandemic and World War I affecting the cohort of people born in 1915–1920. It might also be due to incorrect number of births/deaths registered (records could be lost or births could be registered in adjacent year) during those years which uniformly affected mortality figures for these cohorts.

Barry (2004) reports that in thirteen studies of hospitalised pregnant women during the Spanish flu pandemic the death rate ranged from 23% to 71%. 26% of pregnant women who survived childbirth lost their child. Therefore it is quite possible that such severe death rates could affect

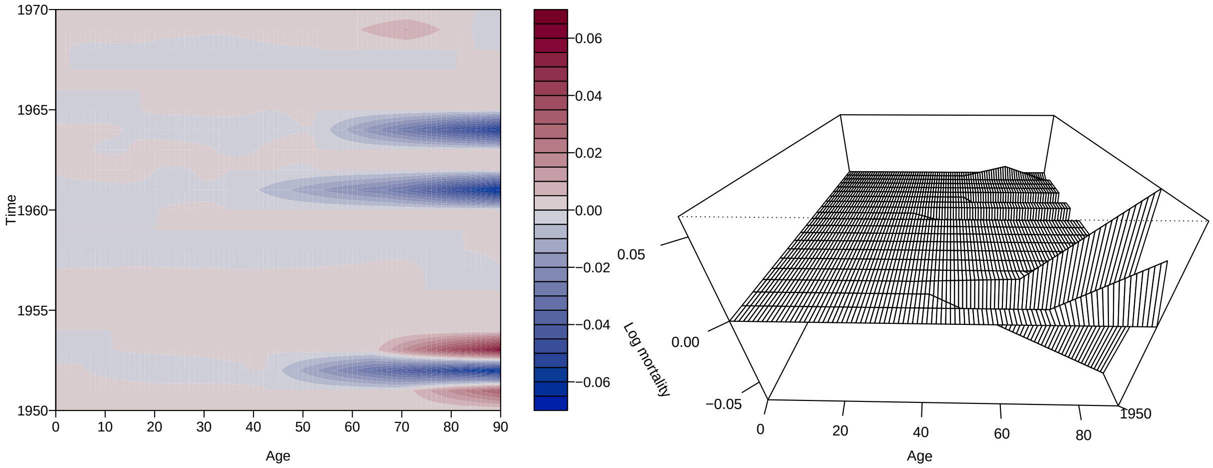


Figure 14: *Period effects of logarithms of French female mortality rates.*

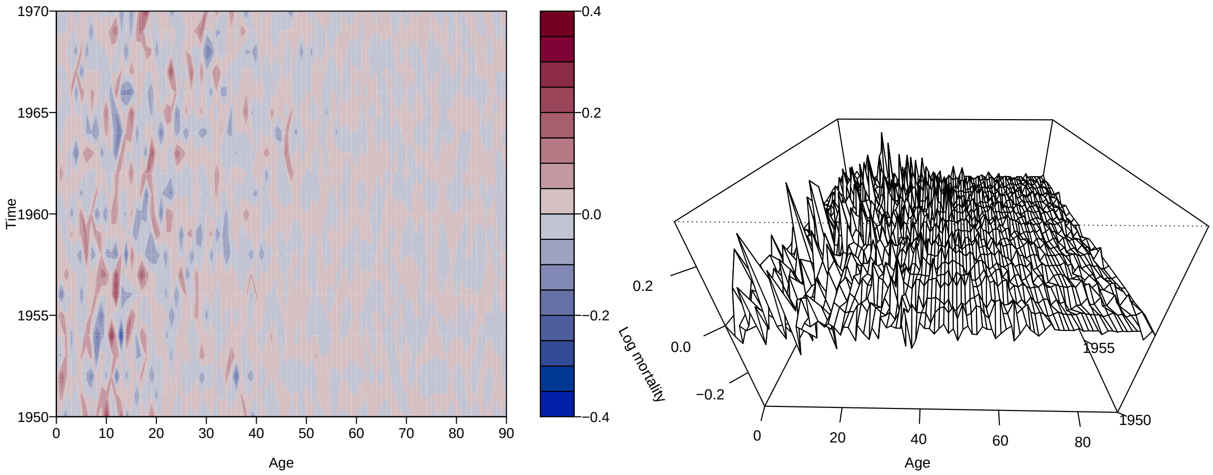


Figure 15: *Residuals of logarithms of French female mortality rates after smoothing with the SMILE method.*

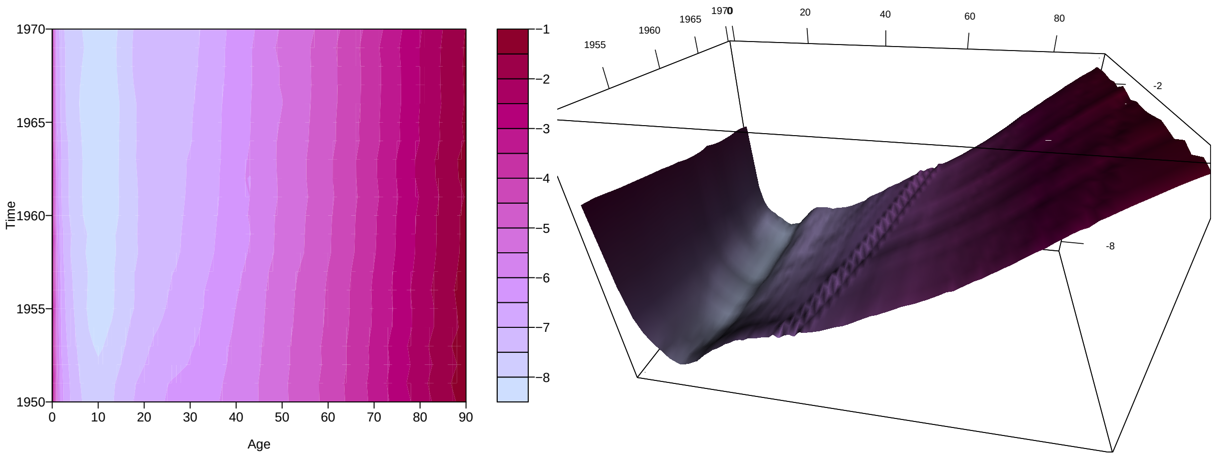


Figure 16: *Logarithms of French female mortality rates smoothed with the SMILE method.*

both health of the cohort born during those years as well as accuracy of the number of registered people in that cohort.

The estimated period effects z_{long} are shown in Figure 14. The most visible period effects can be observed for years 1951, 1952, 1953, 1961 and 1964. They become stronger for older years. Possibly the 1951 and 1953 period effects are due to extreme climate events, or unusual flu epidemics, which tend to affect older people more severely. Years 1952, 1961 and 1964 show some *reduction* in mortality for older ages, the cause of which is unknown.

The residuals from the model are shown in Figure 15. The period and cohort effects are no longer visible.

Figure 16 depicts the complete surface with the cohort and period effects added to the smooth surface. It is the "signal" which we have separated from the "noise" (represented by the residuals).

7 Comparison

We use a cross validation procedure for comparing the different smoothing methods we have discussed. We randomly split all points in the original data into 20 subsets of approximately equal size. Therefore each of these subsets has only 5% of the original data. The resulting cross validation error is the average of errors over those 20 subsets.

We use four subsets of the available French female mortality data for our comparisons.

1. Data for years 1950–1970 and ages 10–60 represent a relatively smooth surface. This comparison is useful to ensure that the most "responsive" algorithms using the L_1 norm do not perform any worse than the more "stable" algorithms based on the L_2 norm. This data set is also important because it is the only comparison satisfying the requirements for Method 2 (Camarda, 2012) which is designed to work for ages starting from 10 years and where there are no outliers.

2. Data for years 1950–1970 and ages 0–60 represent a period when no outliers happened — there were no wars or large pandemics. The younger ages from 0–9 have abrupt changes in mortality rates which are more challenging to smooth.

3. Data for years 1935–1955 and ages 10–60 represent a period including "outliers" due to World War II. It is important to mention that such "outlier" should be considered as an

outlier only along time dimension and it is a smooth data curve along ages dimension. Therefore in our case, when only uncorrelated errors are considered as noise, such one-dimensional outlier should be preserved during smoothing as signal. These data are important for testing the smoothing abilities of the algorithms in the presence of one-dimensional outliers.

4. Data for years 1935–1955 and ages 0–60 represents the most complex dataset containing the one-dimensional outliers (WWII) and also a period of abrupt mortality changes from ages 0–9.

All calculations are done using R (R Core Team, 2015).

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.02	0.50	0.61	0.46	0.41	5.88	5.09	5.42	4.92	4.53
1950-1970	0-60	1.72	0.82	0.52	0.49	0.45	6.60	6.31	5.26	5.18	4.88
1935-1955	10-60	0.79	4.32	1.00	0.32	0.28	4.39	14.26	6.89	3.97	3.64
1935-1955	0-60	3.03	3.99	0.77	0.37	0.34	6.42	13.88	6.41	4.24	3.99

Table 1: *Cross validation performance of different smoothing methods against French female mortality data (SE1 is MSE for cross validation of Method 1 mutiplied by 100, AE1 is MAE for crossvalidation of Method 1 mutiplied by 100, ... , AE5 is MAE for crossvalidation of Method 5 mutiplied by 100)*

Table 1 demonstrates that the SMILE method shows better or similar performance (in terms of MSE and MAE) compare to the other methods in all tests. Amongst the other methods, the Quantile Lasso performs better than Hyndman-Ulah, Camarda and Boosted thin plate splines methods in most cases. Boosted thin plate splines work well except for the cases when there are outliers, for which the Hyndman-Ullah method does better. Overall, the SMILE method is the best performing method amongst those tested.

In the Appendix we provide additional cross validation performance tests for the above methods for both sexes in eleven other countries, showing that the conclusions drawn here based on French females are supported when tested against other data sets.

8 Conclusion

In this paper we have considered three new methods to smooth mortality data in two dimensions. We have also introduced a comparison technique that is computationally feasible. Using this technique we compared our new methods between each other and also with existing one- and two-dimensional methods. We found that our proposed SMILE method gave the best results.

This method also provides us with some insights into the mortality data including the existence of cohort and period effects that might otherwise be overlooked.

We conclude that use of two-dimensional data and thin plate splines for smoothing mortality data can lead to improvements compared to a one dimensional approach (except in terms of MAE on abrupt data). On the other hand, using the L_1 norm instead of L_2 can lead to further performance improvements for abrupt data. Moreover, the methods which use the L_1 norm do not require overcomplicated preprocessing of data, which was necessary for methods based on L_2. Further performance improvements can be achieved by building into the model abilities to project the cohort and period effects.

Additional improvements in these proposed methods are possible. Our boosted thin plate spline method used adaptive splines, while only partially adaptive splines were used for the Quantile Lasso methods. A fully adaptive approach applied to the Quantile Lasso methods requires further investigation and may provide further improvements.

The λ coefficients used in the Quantile Lasso methods were estimated using lengthy numerical methods. A simpler procedure, similar to what is used by Camarda (2012), would improve their practical usefulness. We leave this to a later paper.

References

Barry, J. M. (2004). *The Great Influenza: The Epic Story of the 1918 Pandemic*. Viking.

Bergmeir, C., Molina, D., and Benítez, J. M. (2012). *Rmalschains: Continuous Optimization using Memetic Algorithms with Local Search Chains (MA-LS-Chains) in R*. http://cran.r-project.org/package=Rmalschains.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42:693–734.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Camarda, C. G. (2012). MortalitySmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, 50(1):1–24.

Camarda, C. G., Eilers, P. H. C., and Gampe, J. (2010). Additive decomposition of vital rates from grouped data. In *Proceedings of the 25th International Workshop on Statistical Modelling*, pages 113–118.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Forfar, D. O., McCutcheon, J. J., and Wilkie, A. D. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, 115(1):1–149.

Hastie, T. J., Tibshirani, R., and Friedman, J. H. (2008). *The elements of statistical learning*. Springer-Verlag, New York, 2nd edition.

He, X., Ng, P., and Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B*, 60(3):537–550.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries (1886-1994)*, 107(1):49–80.

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. (2014). *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.16. With contributions from Heather Booth and Leonie Tickle and John Maindonald. http://cran.r-project.org/package=demography.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3):323–342.

Hyndman, R. J. and Ullah, S. M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Kirkby, J. and Currie, I. (2010). Smooth models of mortality with period shocks. *Statistical Modelling*, 10(2):177–196.

Koenker, R. (2015). *quantreg: Quantile Regression*. R package version 5.11. http://cran.r-project.org/package=quantreg.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Portnoy, E. (1997). Regression-quantile graduation of Australian life tables, 1946–1992. *Insurance: Mathematics and Economics*, 21(2):163–172.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Sangalli, L. M. (2014). Statistical and numerical techniques for spatial functional data analysis. In Bongiorno, E. G., Salinelli, E., Goia, A., and Vieu, P., editors, *Contributions in infinite-dimensional statistics and related topics*. Società Editrice Esculapio.

Schuette, D. R. (1978). A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30:407–431.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15(5):1126–1133.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. CRC Press.

Appendix

In this appendix, we repeat the cross-validation performance tests for the French male data, and eleven other countries (male and female). All data is taken from Human Mortality Database (2008). The countries selected are all those for which the Human Mortality Database has data available for the period 1935–1970. Overall 96 tests were performed.

The test results provided in Tables 2 to 24 show that the Quantile Lasso and SMILE methods (Methods 4 and 5 in the tables below) show similar or better performance for ages from 10 to 60 and outperform other tested methods in the majority of tests for ages from 0 to 60. In general the SMILE (Method 5) is the best performing method among those tested.

The most difficult data sets for Methods 4 and 5 are the data where log mortality rates are almost flat (for example for ages 10–60), where there are few features to extract, and which contain much noise (for example when country population and mortality are low at the same time). The combination of such features leads to modest results of the new methods. Although disappointing, this is to be expected since for completely flat and very noisy data, it is very difficult to outperform a simple linear regression.

Similarly to Table 1, Tables 2 to 24 contain figures for cross validation performance of tested smoothing methods against mortality data for a particular country and gender. SE1 column contains MSE for cross validation of Method 1 mutiplied by 100, AE1 column contains MAE for cross validation of Method 1 mutiplied by 100, . . . , AE5 column contains MAE for crossvalidation of Method 5 mutiplied by 100.

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	2.72	1.76	1.69	1.67	1.64	11.23	9.99	9.78	9.74	9.57
1950-1970	0-60	2.69	2.17	1.70	1.65	1.64	11.53	11.09	9.89	9.71	9.56
1935-1955	10-60	1.79	1.42	1.32	1.31	1.28	9.73	9.13	8.94	8.84	8.53
1935-1955	0-60	1.82	1.68	1.37	1.37	1.34	10.07	9.88	9.02	9.02	8.78

Table 2: *Australian female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	2.64	0.99	0.98	0.98	0.92	9.46	7.57	7.45	7.52	7.27
1950-1970	0-60	2.85	1.71	1.00	1.01	0.99	10.00	9.42	7.65	7.66	7.59
1935-1955	10-60	2.18	1.40	1.19	1.12	1.08	9.06	8.64	7.97	7.93	7.76
1935-1955	0-60	1.82	1.64	1.07	1.05	1.05	9.06	9.39	7.79	7.66	7.64

Table 3: *Australian male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.83	1.08	1.02	1.06	1.04	8.89	7.69	7.56	7.67	7.46
1950-1970	0-60	2.97	1.77	0.98	1.02	1.02	9.68	9.58	7.40	7.55	7.48
1935-1955	10-60	1.38	0.83	0.91	0.83	0.79	7.76	6.89	7.09	6.85	6.64
1935-1955	0-60	1.33	1.28	0.84	0.83	0.80	7.98	8.28	7.02	6.97	6.80

Table 4: *Canadian female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.27	0.60	0.58	0.55	0.53	7.01	5.84	5.70	5.58	5.43
1950-1970	0-60	3.19	1.48	0.59	0.57	0.56	8.23	8.17	5.82	5.75	5.69
1935-1955	10-60	0.97	0.68	0.62	0.60	0.58	6.91	6.37	6.14	5.93	5.84
1935-1955	0-60	1.32	1.24	0.62	0.60	0.56	7.34	8.07	6.13	5.96	5.77

Table 5: *Canadian male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	6.50	4.68	4.56	4.67	4.70	16.69	15.16	15.24	15.26	15.28
1950-1970	0-60	8.50	5.87	5.06	5.02	5.11	18.82	16.95	16.03	15.77	15.99
1935-1955	10-60	4.14	2.73	3.08	2.57	2.56	12.82	12.04	12.30	11.76	11.75
1935-1955	0-60	4.92	3.55	3.00	2.98	2.99	14.23	13.68	12.77	12.79	12.79

Table 6: *Danish female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	5.01	2.68	2.76	2.57	2.74	13.81	11.62	11.94	11.56	11.77
1950-1970	0-60	7.92	3.94	2.90	2.83	2.94	15.94	13.85	12.40	12.14	12.33
1935-1955	10-60	3.42	2.49	2.40	2.07	2.11	11.67	11.14	10.97	10.44	10.65
1935-1955	0-60	3.72	3.05	2.37	2.28	2.34	12.64	12.41	11.23	11.07	11.20

Table 7: *Danish male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.05	0.97	0.84	0.64	0.62	6.27	7.14	6.72	5.95	5.82
1950-1970	0-60	0.95	1.20	0.86	0.62	0.62	6.47	7.98	6.94	5.91	5.89
1935-1955	10-60	1.46	2.44	1.38	0.94	0.91	7.66	11.59	8.47	7.21	7.12
1935-1955	0-60	1.48	2.97	1.23	1.01	0.98	7.92	12.63	8.20	7.37	7.31

Table 8: *Dutch female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	0.72	0.63	0.69	0.45	0.46	4.90	5.67	5.78	4.81	4.89
1950-1970	0-60	0.85	1.01	0.61	0.44	0.43	5.36	7.07	5.65	4.85	4.78
1935-1955	10-60	1.56	6.38	2.71	0.91	0.81	6.97	17.68	11.40	6.52	6.39
1935-1955	0-60	1.46	6.30	2.38	0.94	0.91	7.27	17.86	10.79	6.78	6.81

Table 9: *Dutch male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.09	0.64	0.65	0.59	0.54	6.54	5.67	5.68	5.54	5.22
1950-1970	0-60	2.73	1.55	0.67	0.62	0.61	7.83	8.24	5.97	5.78	5.58
1935-1955	10-60	1.00	0.60	0.76	0.35	0.27	5.58	5.88	5.63	4.43	3.87
1935-1955	0-60	1.20	1.15	0.56	0.42	0.34	5.94	7.55	5.56	4.81	4.30

Table 10: *English female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.10	0.51	0.52	0.38	0.33	6.02	5.11	4.90	4.56	4.16
1950-1970	0-60	3.26	1.48	0.45	0.44	0.40	7.57	7.72	4.94	4.95	4.57
1935-1955	10-60	1.01	6.38	2.41	0.44	0.39	5.78	15.10	9.65	4.62	4.40
1935-1955	0-60	1.56	6.00	2.02	0.48	0.43	6.41	15.17	9.04	4.92	4.65

Table 11: *English male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	5.10	3.48	3.62	3.51	3.59	15.09	13.45	13.68	13.67	13.77
1950-1970	0-60	6.27	4.13	3.72	3.73	3.81	15.92	14.68	14.09	14.02	14.20
1935-1955	10-60	3.19	1.65	2.21	1.53	1.51	10.44	9.36	9.98	8.89	8.90
1935-1955	0-60	2.60	2.12	1.77	1.55	1.59	10.28	10.60	9.62	9.18	9.29

Table 12: *Finnish female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	3.41	2.14	1.93	1.98	2.03	11.54	10.48	10.25	10.32	10.29
1950-1970	0-60	4.77	2.59	2.09	2.09	2.11	12.85	11.65	10.57	10.62	10.58
1935-1955	10-60	2.01	7.15	5.04	1.68	1.66	9.41	18.09	14.50	9.10	9.09
1935-1955	0-60	2.54	6.70	4.31	1.69	1.67	9.99	18.08	13.72	9.32	9.40

Table 13: *Finnish male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.41	0.50	0.45	0.38	0.31	6.03	5.19	4.90	4.46	3.97
1950-1970	0-60	1.77	0.98	0.47	0.39	0.34	6.54	6.82	5.08	4.59	4.20
1935-1955	10-60	1.00	9.38	2.53	0.35	0.30	4.52	20.04	10.35	4.02	3.76
1935-1955	0-60	2.85	8.52	2.28	0.41	0.37	6.43	18.92	9.76	4.28	4.05

Table 14: *French male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	0.53	0.40	0.38	0.34	0.30	4.78	4.83	4.71	4.42	4.16
1950-1970	0-60	2.14	0.80	0.39	0.36	0.33	6.14	6.20	4.75	4.54	4.28
1935-1955	10-60	0.81	0.53	0.33	0.23	0.21	3.90	5.65	3.95	3.46	3.29
1935-1955	0-60	8.96	0.70	0.43	0.30	0.27	7.67	6.38	4.48	3.77	3.65

Table 15: *Italian female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	0.83	0.38	0.31	0.29	0.25	4.97	4.76	4.19	4.01	3.74
1950-1970	0-60	1.51	0.84	0.32	0.29	0.24	5.65	6.25	4.26	4.10	3.76
1935-1955	10-60	0.95	4.57	1.09	0.39	0.35	4.31	15.20	7.07	3.92	3.75
1935-1955	0-60	7.23	3.90	1.06	0.32	0.30	7.43	13.89	6.98	3.98	3.78

Table 16: *Italian male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.25	0.99	0.90	0.65	0.57	8.26	7.76	7.37	6.24	5.82
1950-1970	0-60	2.47	1.78	0.88	0.67	0.60	8.83	9.73	7.28	6.40	6.00
1935-1955	10-60	3.20	2.19	2.20	0.73	0.44	12.47	10.77	10.90	6.34	5.00
1935-1955	0-60	3.82	2.04	1.99	0.68	0.44	12.27	10.55	10.20	6.25	5.06

Table 17: *Spanish female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	1.18	0.66	0.53	0.41	0.39	6.98	6.31	5.54	4.75	4.76
1950-1970	0-60	3.10	1.50	0.69	0.46	0.38	8.77	8.51	6.49	5.18	4.90
1935-1955	10-60	1.28	1.41	1.23	0.55	0.51	8.82	9.46	8.48	5.67	5.48
1935-1955	0-60	2.00	1.82	1.45	0.49	0.50	9.54	10.15	8.86	5.41	5.25

Table 18: *Spanish male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	4.15	2.90	2.95	2.72	2.75	13.15	11.92	12.12	11.63	11.66
1950-1970	0-60	7.93	3.70	3.13	3.06	3.09	15.91	13.62	12.62	12.47	12.46
1935-1955	10-60	2.53	1.71	1.91	1.61	1.59	9.87	9.25	9.52	9.01	8.92
1935-1955	0-60	3.65	2.50	1.92	1.76	1.76	11.25	10.97	9.79	9.45	9.41

Table 19: *Swedish female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	3.53	1.58	1.68	1.60	1.59	11.01	9.18	9.24	9.24	9.25
1950-1970	0-60	7.82	2.71	1.82	1.79	1.77	13.61	11.50	9.79	9.76	9.75
1935-1955	10-60	3.03	1.35	1.51	1.26	1.28	9.88	8.43	8.66	8.48	8.49
1935-1955	0-60	3.39	2.01	1.39	1.29	1.29	10.26	10.08	8.69	8.53	8.56

Table 20: *Swedish male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	4.34	3.36	3.28	3.16	3.25	14.32	13.27	13.26	13.03	13.20
1950-1970	0-60	5.93	3.98	3.38	3.31	3.38	15.87	14.54	13.48	13.33	13.52
1935-1955	10-60	3.07	2.36	3.37	2.22	2.22	11.19	10.44	10.96	10.14	10.17
1935-1955	0-60	3.36	2.86	2.34	2.20	2.25	11.87	11.91	10.68	10.39	10.56

Table 21: *Swiss female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	3.11	1.56	1.64	1.48	1.52	11.03	9.24	9.31	9.11	9.24
1950-1970	0-60	5.67	2.61	1.81	1.82	1.83	13.21	11.41	9.78	9.87	9.96
1935-1955	10-60	2.52	1.34	1.44	1.23	1.29	9.63	8.47	8.56	8.11	8.35
1935-1955	0-60	2.92	2.00	1.54	1.45	1.47	10.90	10.20	9.03	8.81	8.87

Table 22: *Swiss male*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	0.29	0.14	0.10	0.10	0.09	2.72	2.83	2.40	2.09	2.00
1950-1970	0-60	1.52	1.06	0.12	0.11	0.10	3.93	5.76	2.61	2.28	2.12
1935-1955	10-60	0.17	0.07	0.05	0.01	0.01	0.81	2.01	1.53	0.49	0.49
1935-1955	0-60	0.60	0.63	0.08	0.03	0.03	1.73	4.25	1.97	0.83	0.83

Table 23: *USA female*

Years	Ages	SE1	SE2	SE3	SE4	SE5	AE1	AE2	AE3	AE4	AE5
1950-1970	10-60	0.39	0.11	0.09	0.07	0.06	2.64	2.45	2.17	1.79	1.69
1950-1970	0-60	2.88	1.22	0.10	0.09	0.08	4.38	5.69	2.29	1.99	1.89
1935-1955	10-60	0.17	0.15	0.06	0.01	0.01	1.07	2.81	1.66	0.54	0.54
1935-1955	0-60	0.93	0.79	0.12	0.03	0.03	2.10	5.05	2.32	0.91	0.90

Table 24: *USA male*

Chapter 3

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

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

Alexander Dokumentov
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

4 January 2016

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:

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

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;
- Here and further $\|.\|$ is the Frobenius norm;
- Here and further $\odot$ 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 slightly different problem but with obvious similarities. The problem is also closely related to Maximum-Margin Matrix Factorisation (Srebro et al., 2005).

On the other hand ROPES differs from smoothing splines and mixed effects methods considered in James (2010). We can note that (1) can be reorganised to represent a smoothing splines problem by fixing matrix V, which becomes a "predefined" matrix of spline bases:

$$U = \arg\min_{U} \left(\left\| W \odot (Y - UV^T) \right\|^2 + \lambda \|KU\|^2 \right),$$

The main difference between ROPES and the above problem is that ROPES estimates both the spline bases and the coefficients, and is able to apply some reasonable complexity/smoothing restrictions on both of them.

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).

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 goal is to interpolate and extrapolate the data, as well as to remove the noise. 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 $n \times m$ matrix Y, where measurements related to each person $i, 1 \le i \le n$, represent a row in the matrix. Each column will contain observations taken at a given moment $t_j, 1 \le j \le m$, where moments t_j are equally spaced. Many cells in the matrix will be missing. We also create a matrix W, which has same dimensions as Y. This matrix

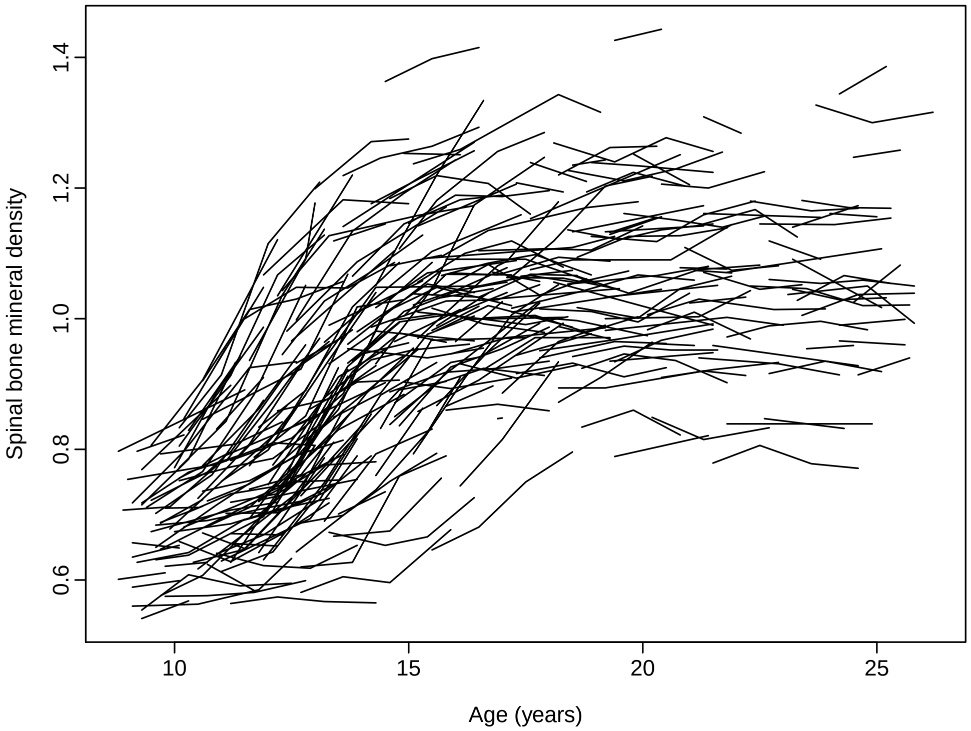


Figure 1: *Measurements of the spinal bone mineral density (g/sm^2) for 280 people.*

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". Thus Z can be described as a "low complexity" approximation of Y. 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:

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

which is the same problem as (1) with the following differences:

- U is more suitable to be called a matrix of "coefficients";
- V can be interpreted as a matrix of "shapes";
- L is a complexity matrix which takes second derivatives of the columns of the matrix V;
- K, which appears in (1) becomes the identity matrix;
- 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.

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 $y_{x,t}$ denote an observed mortality rate for a particular age x and a particular year t. We define $y_{x,t} = d_{x,t}/e_{x,t}$, 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.

In this example we have mortality data available every year and presented as $n \times m$ matrix Y:

$$Y = [y_{j,k}],$$

where $j \in [1 \ldots n]$ and $k \in [1 \ldots m]$, n is the number of age groups and m is the number of years where the data is available.

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).

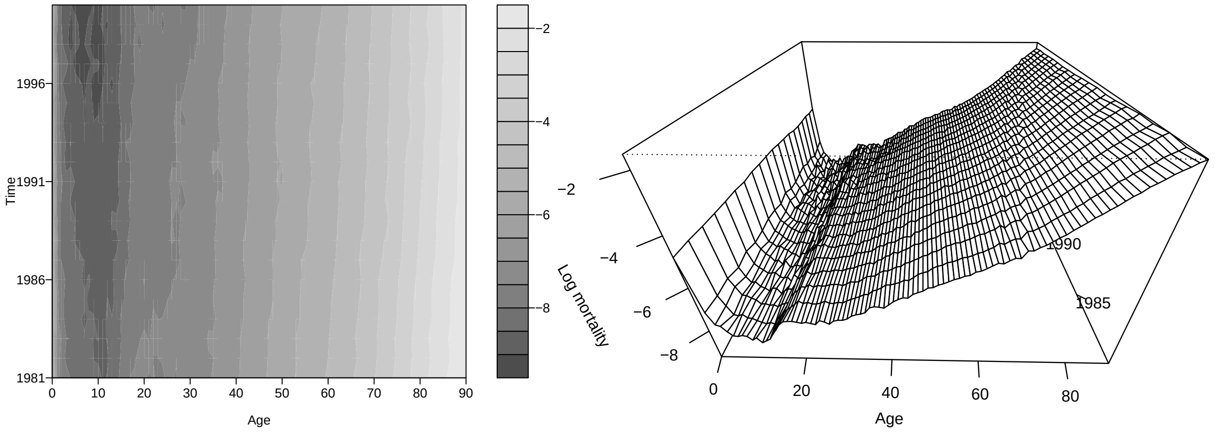


Figure 2: *Natural logarithm of French female mortality rates (ages: 0-90, time: 1981–2000 (in years)).*

Mortality rates are functional data: they are smooth in two dimensions (time and age) although observed with noise. The functional nature of the data allows for the following presentation:

$$y_{x,t} = f(x,t) + \epsilon_{x,t},$$

where $y_{x,t}$ is logarithm of mortality rate for age group x and time t, $f(x,t)$ is a smooth function and $\epsilon_{x,t}$ is "noise". The main goal, when working with such data, is to remove the noise. Another goal, that we set, is to decompose the data into components and use such decomposition for prediction.

To achieve these goals, 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 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). For example for set of observation $\{f_i\}_{i=1}^{n}$, where every observation is a function represented by vector of points $(f_{i1}, f_{i2}, \dots, f_{im})$, matrix Y will be $Y = [f_{ij}]$.

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;

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:

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

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;

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:

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

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;

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:

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

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:

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

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

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

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 $U^T U = V^T V$.*

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:

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

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

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:

$$U^T U = V^T V. \tag{12}$$

□

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

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

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

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

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:

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

Similarly

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

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: $U^TU = V^TV$.

  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, $(U_2,V_2) \in \operatorname*{arg\,min}_{U,V}(J(U,V))$. 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

$$\Psi = S^T Q, \tag{21}$$

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 $S^T Q = \prod_{i=1}^{\ell} \Psi_i$ 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:

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

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

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

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

Let us recall (8):

$$\{Z\} = \underset{Z}{\arg\min}\,(f(Z) + \|Z\|_*).$$

Using one of the definitions of star norm it can be rewritten as

$$\{Z\} = \arg\min_{Z} \left(f(Z) + \sum_{i} |\sigma(Z)_i| \right), \tag{24}$$

where $\{\sigma(Z)_i\}$ is the set of singular values of matrix Z. Equation (24) makes similarity between ROPES and LASSO obvious. Use of L_1 regularisation over singular values instead of L_2 regularization (which in our case corresponds to Frobenius norm of Z: $\|Z\| = \sum_i \left(\sigma(Z)_i^2\right)$) gives us hope that part of the singular values of solution of (8) will be exactly zero. Our experiments confirm such behaviour. All this allows us to consider ROPES solutions as a sum of few rank one matrices corresponding to non-zero singular values.

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

$$Z = \sum_{i=1}^{m} d_i z_i. \tag{25}$$

Let us call the decomposition in (25) 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 $Z_* = KUV^TL^T$.
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:

$$\begin{aligned} J(U,V) &= \left\| W \odot \left(Y - UV^T \right) \right\|^2 + \lambda \|KU\|^2 + \theta \|LV\|^2 \\ &= 2\sigma^2 \left(\left\| \frac{1}{\sqrt{2}\sigma} W \odot \left(Y - UV^T \right) \right\|^2 + \left\| \frac{\sqrt{\lambda}}{\sqrt{2}\sigma} KU \right\|^2 + \left\| \frac{\sqrt{\theta}}{\sqrt{2}\sigma} LV \right\|^2 \right) \\ &= \text{Const} \times \left(\left\| \frac{1}{\sqrt{2}\sigma} W \odot E \right\|^2 + \left\| \frac{\sqrt{\lambda}}{\sqrt{2}\sigma} (I \otimes K)\text{vec}(U) \right\|^2 + \left\| \frac{\sqrt{\theta}}{\sqrt{2}\sigma} (I \otimes L)\text{vec}(V) \right\|^2 \right), \end{aligned} \tag{26}$$

where $E = Y - UV^T$.

Equation (26) can be considered as a minus log likelihood function (with some multiplicative and additive constants) of (27), where the observations Y are partially visible (at the places where W has ones):

$$Y = UV^T + E, \tag{27}$$

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

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 $\delta \in \Delta(i, j, Y) \iff Z_{ij}(Y + \delta) \geq \ell_{ij}(Y)$ & $Z_{ij}(Y + \delta) \leq u_{ij}(Y)$. This definition implies that $\text{Prob}(D \in \Delta(i, j, Y)) \geq p_{\text{conf}}$.

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}:

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

Thus, $\text{Prob}(Y_0 \in Y + \Delta(i, j, Y)) \geq p_{\text{conf}}$, 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 $\Delta = \bigcup\limits_{k=1}^{\ell} \{\Delta_k\}$.
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 $\Delta = \bigcup\limits_{k=1}^{\ell} \{\Delta_k\}$ and $\Upsilon = \bigcup\limits_{k=1}^{\ell} \{\Upsilon_k\}$.
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: $Y_{\Delta\Upsilon} = \bigcup\limits_{k=1}^{\ell} \{Z(Y_\Delta)_k + \Upsilon_k\}$.
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 (27) 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):

$$\{(U,V)\} = \operatorname*{arg\,min}_{U,V}\Big(\|W\odot(Y-UV^T)\|^2+\|U\|^2+\|\mathrm{DIFF}_2(\lambda_2)V\|^2+\|\mathrm{DIFF}_1(\lambda_1)V\|^2+\|\mathrm{DIFF}_0(\lambda_0)V\|^2\Big), \tag{28}$$

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 $D^{(\cdot)}$ are conforming differentiation matrices

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

Problem (28) 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.

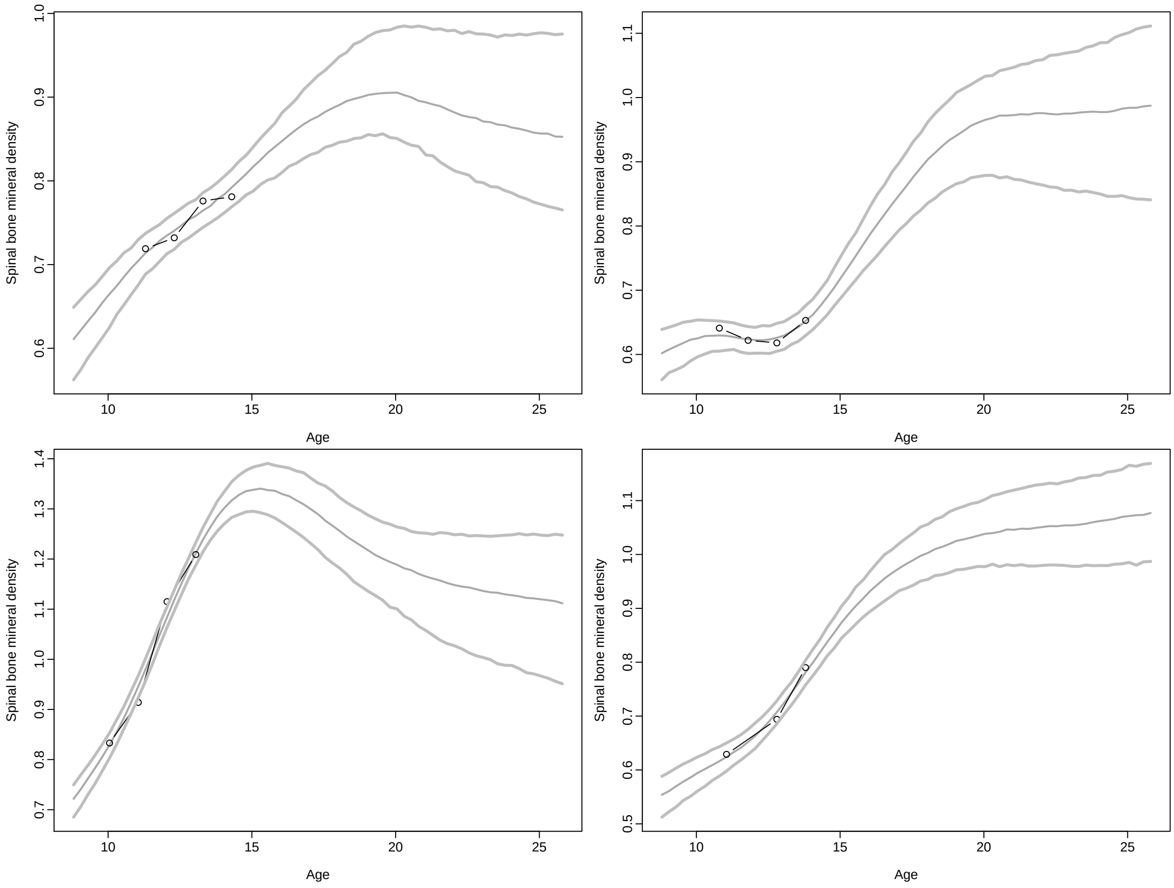


Figure 3: *Interpolation and extrapolation of the spinal bone mineral density (g/sm^2) for cases 1, 3, 4 and 190.*

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

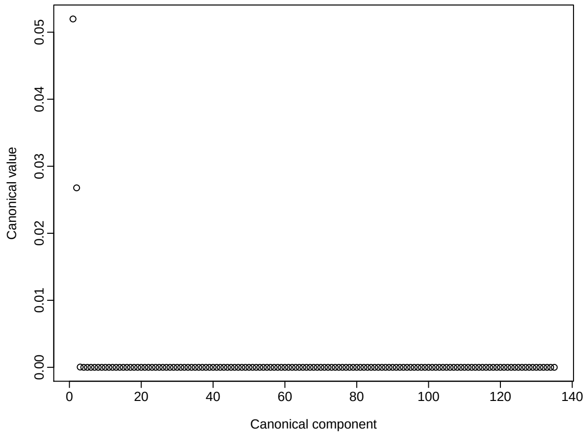


Figure 4: *Canonical values of the spinal bone mineral density dataset.*

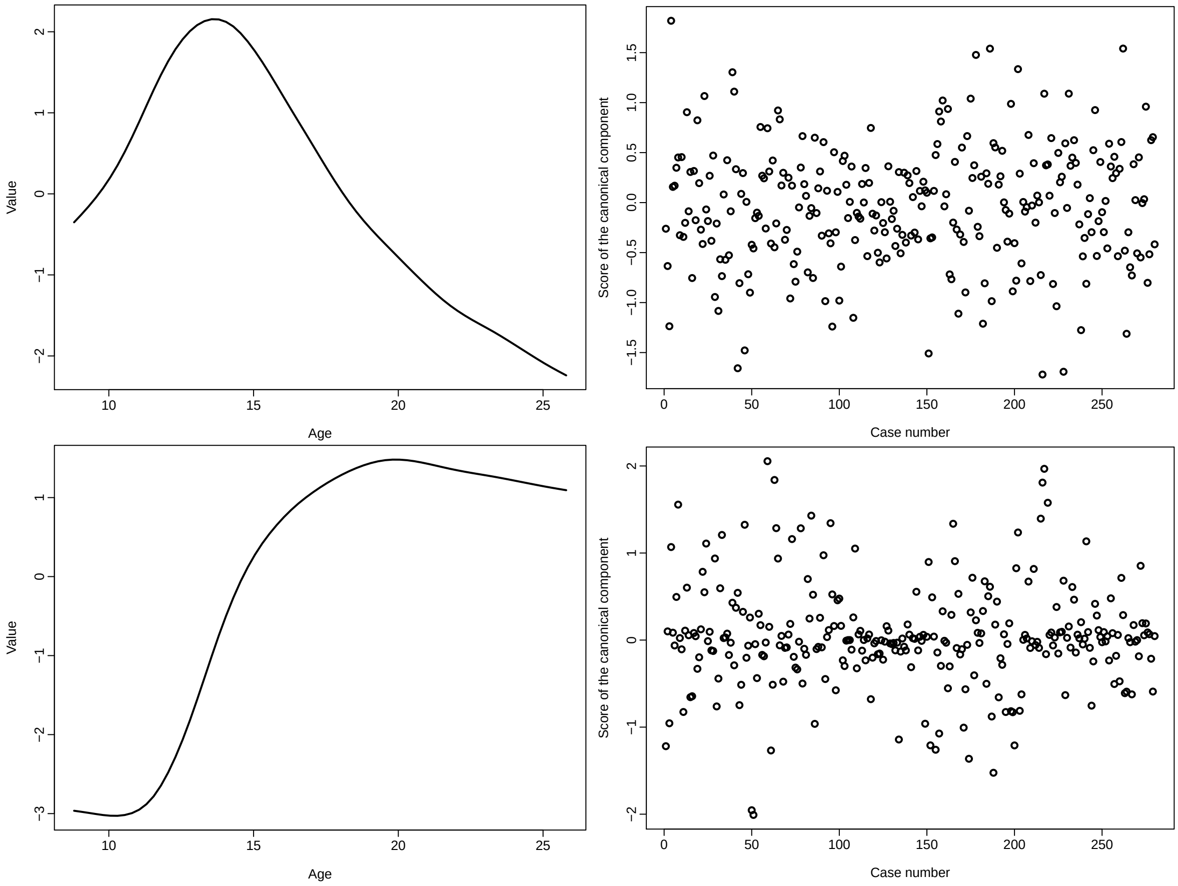


Figure 5: *The first two canonical components and their scores for the spinal bone mineral density dataset.*

Canonical vectors and their scores are presented in Figure 5. It is reasonable to consider the first component in Figure 5 as an element which is responsible for the curvature of the lines around ages 10-20, while the second component appears to be mostly concerned about the steepness of the growth between ages 10 and 20. Figure 7 shows four individual curves with high-high, low-low, high-low and low-high combinations of scores of the first and the second components.

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

The overlay plot of 40 randomly chosen forecasts is presented in Figure 8.

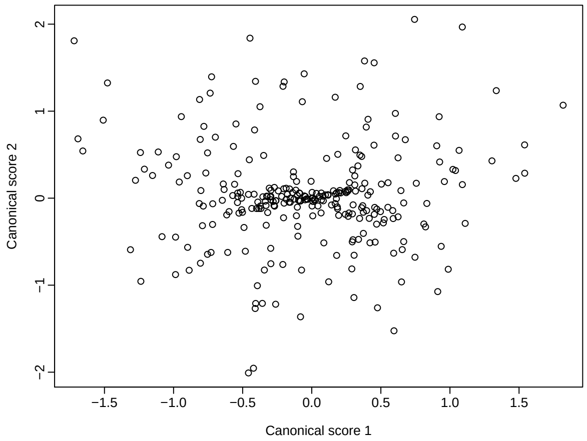


Figure 6: *Scatter plot of the first two canonical scores of the spinal bone mineral density dataset.*

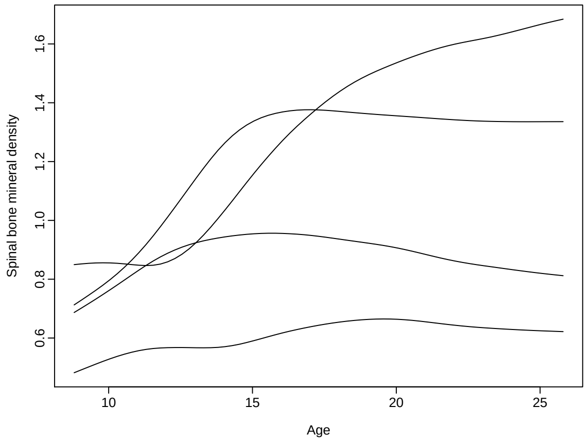


Figure 7: *Four individual curves with high-high, low-low, high-low and low-high combinations of scores of the first and the second components.*

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.

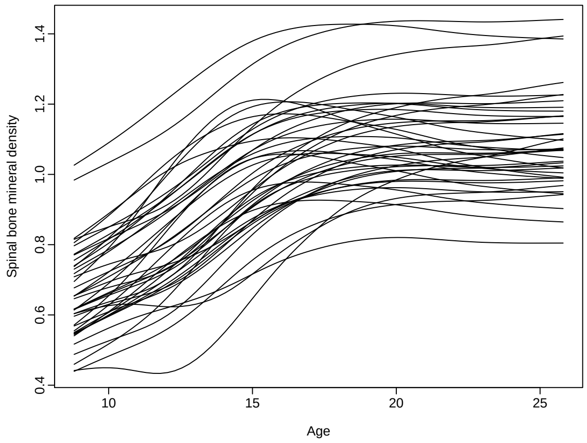


Figure 8: *Overlay plot of 40 randomly chosen forecasts.*

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:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\Big(\big\|W \odot (Y - UV^T)\big\|^2 + \|\mathrm{DIFF}_2(\mu_a)U\|^2 + \|\mathrm{DIFF}_1(\theta_a)U\|^2 + \|\mathrm{DIFF}_0(\lambda_a)U\|^2 + \|\mathrm{DIFF}_2(\mu_y)V\|^2 + \|\mathrm{DIFF}_1(\theta_y)V\|^2 + \|\mathrm{DIFF}_0(\lambda_y)V\|^2 \Big), \tag{29}$$

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 $D^{(\cdot)}$ are conforming differentiation matrices

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

Let us note that problem (29) 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 (29) in the case when all $\theta_{.}$ and $\lambda_{.}$ 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 (29) 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 (29) 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 (29) 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 9.

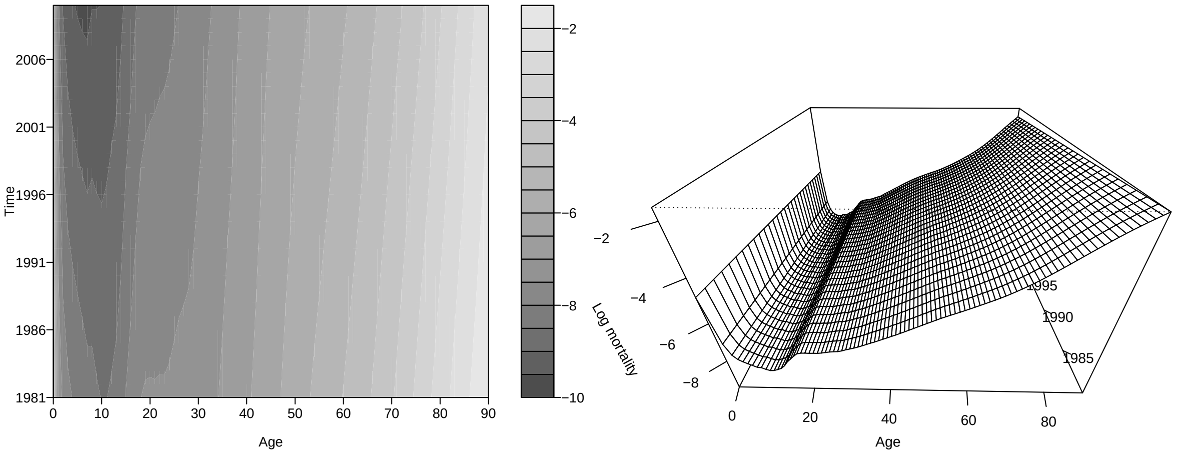


Figure 9: *The fitted (first 20 years) and forecast (last 10 years) surface for French female log mortality rates using LPM.*

The residuals are shown in Figure 10. The results of the 10-year forecasting with 90% prediction intervals are shown in Figure 11.

The canonical decomposition has two significant terms, as can be seen from Figure 12. The first two canonical vectors and their scores are presented in Figure 13. It appears that the first component represents the “main shape” of the data, although the second component applies a small “variation” to the first component.

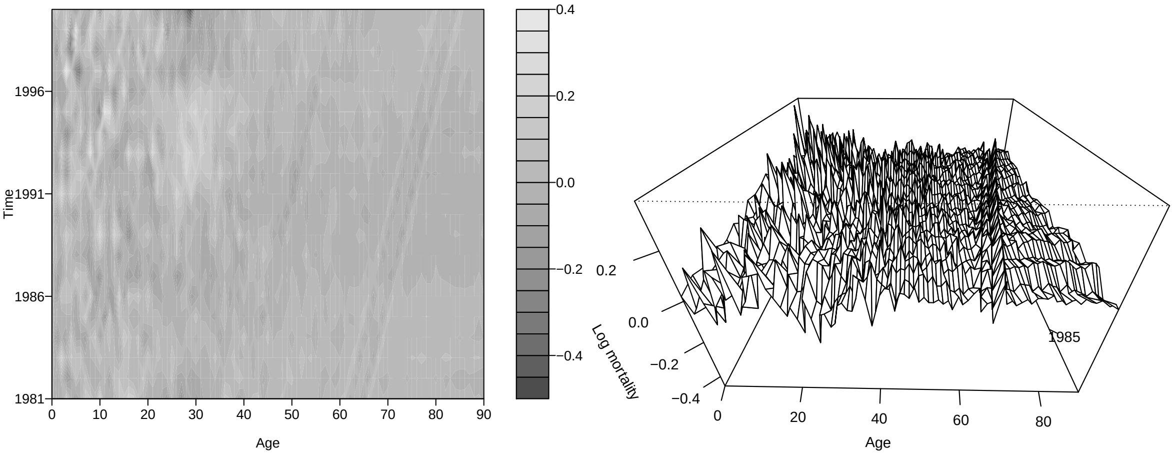


Figure 10: *The residuals for the fitted surface for French female log mortality rates using LPM.*

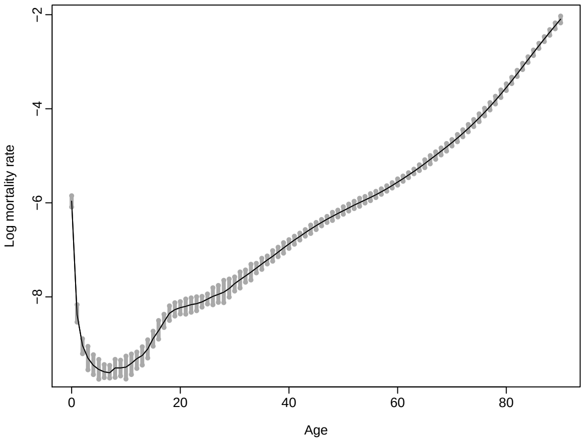


Figure 11: *The forecast of French female log mortality rates for year 2010 (with 90% prediction intervals).*

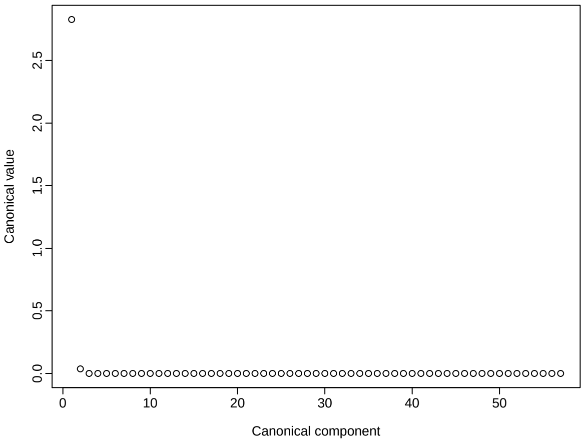


Figure 12: *Canonical values of French female log mortality rates, years 1981–2000.*

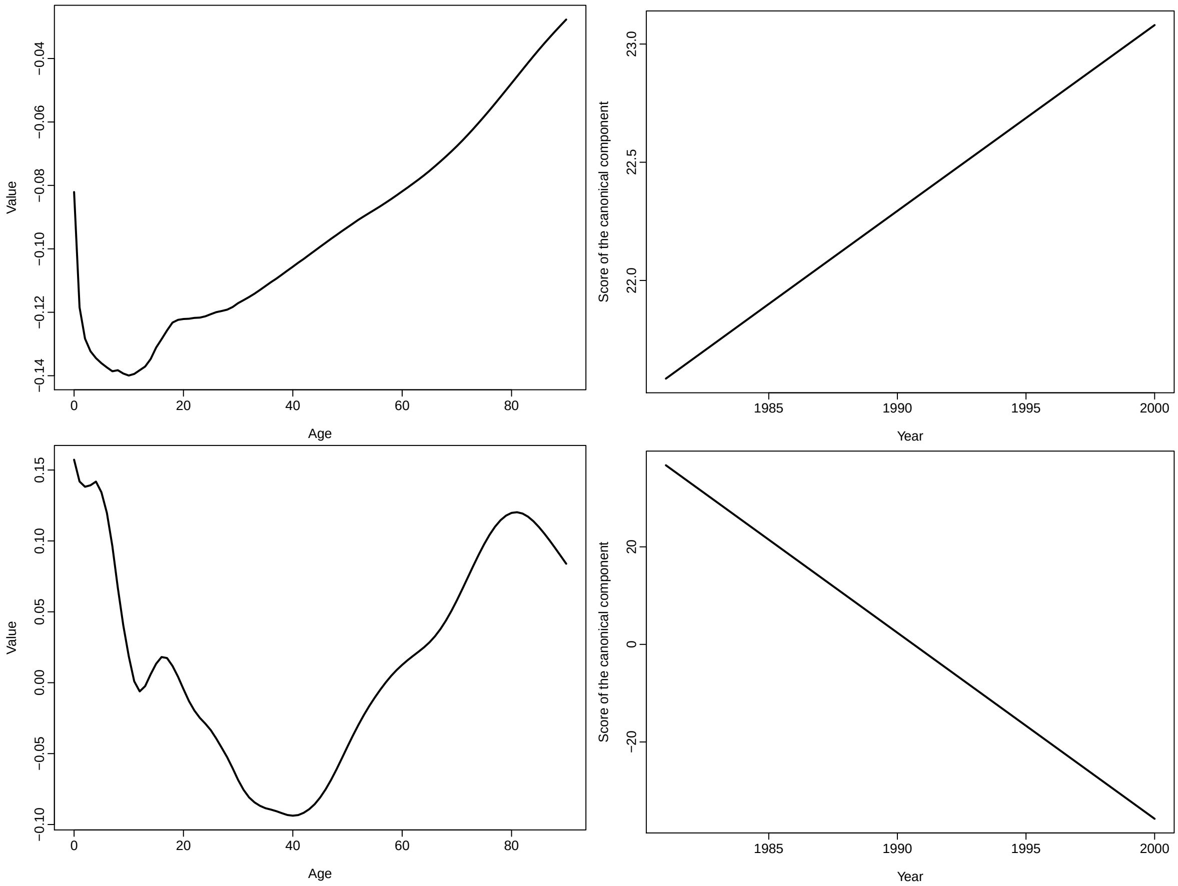


Figure 13: *The first two canonical components and their scores for French female log mortality rates, years 1981–2000.*

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 (30) can also be reduced to a convex optimisation problem:

$$\{(U,V)\} = \underset{U,V}{\arg\min}\Big(\big\|W \odot (Y - UV^T)\big\|_{L_p}^p + \lambda\|KU\|^2 + \theta\|LV\|^2\Big). \tag{30}$$

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.

Chapter 4

Paper “STR: A Seasonal-Trend Decomposition Procedure Based on Regression”

STR: A Seasonal-Trend Decomposition Procedure Based on Regression

Alexander Dokumentov
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
[redacted]

4 January 2016

JEL classification: C10,C14,C22

STR: A Seasonal-Trend Decomposition Procedure Based on Regression

Abstract

We propose new generic methods for decomposing seasonal data: STR (a Seasonal-Trend decomposition procedure based on Regression) and Robust STR. In some ways, STR is similar to Ridge Regression and Robust STR can be related to LASSO. Our new methods are much more general than any alternative time series decomposition methods. They allow for multiple seasonal and cyclic components, and multiple linear regressors with constant, flexible, seasonal and cyclic influence. Seasonal patterns (for both seasonal components and seasonal regressors) can be fractional and flexible over time; moreover they can be either strictly periodic or have a more complex topology. We also provide confidence intervals for the estimated components, and discuss how STR can be used for forecasting.

Keywords: time series decomposition, seasonal data, Tikhonov regularisation, ridge regression, LASSO, STL, TBATS, X-12-ARIMA, BSM

1 Introduction

Time series decomposition is an important task in all national statistics agencies, and in many other contexts in which seasonal variation in time series data is observed. It is the basis for all seasonal adjustment procedures, it is widely used in forecasting, and it underpins the analysis of business cycles.

The first attempts to decompose time series into various components can be dated as early as 1884 when Poynting proposed price averaging as a tool to eliminate trend and seasonal fluctuations (Makridakis et al., 1998). Later his approach was extended by Hooker (1901), Spencer (1904) and Anderson and Nochmals (1914). Further research in that direction included Copeland (1915), who was the first to attempt to extract the seasonal component, until Macauley (1930) proposed a method which became "classical" over time. The work of Macauley led to the Census II method, which became widely used after a computer program developed in 1955 significantly simplified the calculations (Shiskin, 1957). The Census II method has continued to evolve, and various techniques and features have been added such as robustness, calendar effects, regressors, ARIMA extensions, and extensive diagnostics. Widely used versions of this approach are X-11 (Shishkin et al., 1967), X-11-ARIMA (Dagum, 1988), X-12-ARIMA (Findley et al., 1998) and X-13ARIMA-SEATS (Findley, 2005). X-13-ARIMA-SEATS includes a version of the TRAMO/SEATS procedure for seasonal adjustment which was developed at the Bank of Spain (see Monsell and Aston (2003)).

A different approach was followed by Cleveland et al. (1990) who developed STL (Seasonal-Trend decomposition using Loess) which has become widely used outside the national statistics agencies, largely because of its availability in R (R Core Team, 2015). This method uses iterative Loess smoothing to obtain an estimate of the trend and then Loess smoothing again to extract a changing additive seasonal component.

Burman (1980) argued that there were too many seasonal adjustment methods and noted that all but one were ad hoc methods. Since that time, several model-based methods for seasonal decomposition have been developed including the TRAMO/SEATS procedure mentioned above, the BATS and TBATS models of De Livera et al. (2011), and various structural time series model approaches (Harvey, 1990; Commandeur et al., 2011). The big advantage of using a model for such methods is that it allows the easy calculation of confidence and prediction intervals, which are not available in many ad hoc methods.

Despite this long history, and the availability of many time series decomposition algorithms and models, there are many time series characteristics that are not addressed in any of these approaches.

The major deficiencies of the main decomposition methods are as follows:

- Inability to provide a meaningful and simple statistical model (for many methods).
- Inability (or difficulty) to calculate confidence intervals (for many methods).
- Inability to take into account regressors (for some methods).
- Inability to take into account fractional seasonality (for most methods).
- Inability to take into account multiple seasonality (for most methods).
- Inability to take into account complex seasonality and regressors which affect data in a seasonal manner (for all methods).

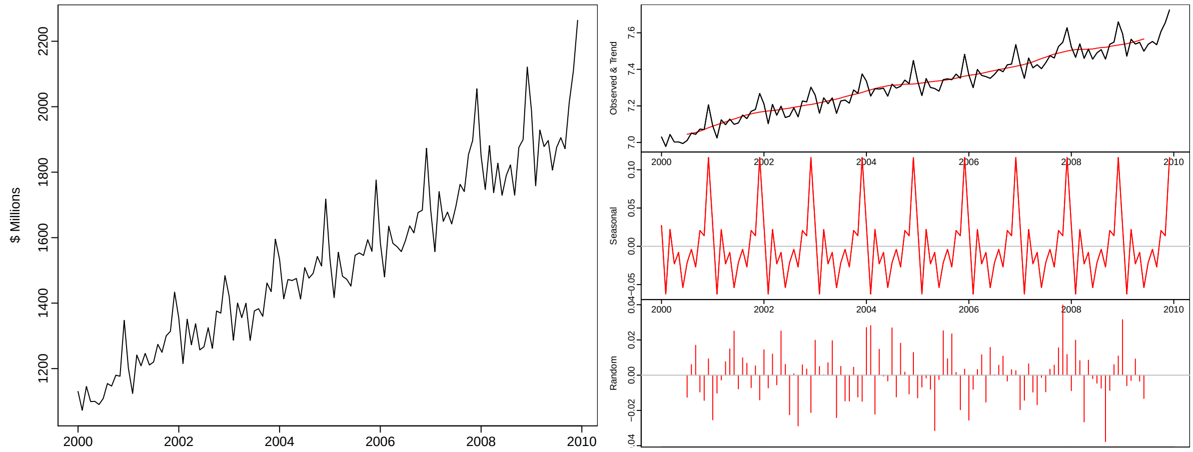
As we can see, currently a variety of methods are available, although few of them have the clarity, simplicity and generality to allow them to handle the many problems which arise with seasonal data decomposition. We aim to fill this gap with our new approach. It is clear, generic, model-based, robust (if required) and is simple — we show that the problem of seasonal decomposition can be re-cast in the framework of ordinary least squares or quantile regression. Moreover our approach allows new features (such as predictors with seasonally varying effects) that have not been developed before. In our opinion, our new STR method is the most generic framework currently available for decomposition of seasonal data.

The structure of this article is as follows: In Section 2 we provide a very simple motivating example which clarifies the main intuitions behind the idea of seasonal decomposition. In Sections 3 and 4 we develop the simplest STR model and show how it can be reduced to ordinary least squares (OLS). In Section 4.2 we show how to efficiently calculate leave-one-out cross validation and propose to use cross validation for estimation of the smoothing parameters. In Section 5 we extend our model to the case when seasonality is considered as a smooth function in two dimensions and defined over a cylinder. In Section 6 we consider cases of multiple seasonality, and we introduce predictors as well as allow them to be “flexible” and “seasonal”. In Section 8 by improving the performance of the method we also solve the problem of fractional seasonality, we introduce the concept of seasonality with complex topology, and we show how to forecast using our model. In Section 9 we introduce RSTR — a robust version of STR. In Section 10 we provide a complex example to highlight the features and capabilities of our approach. Finally in Section 11 we discuss the benefits and disadvantages of this new approach.

We also propose a way to take into account cyclicity in the data and note that it can be handled similarly to seasonality.

2 A simple motivating example

Seasonal time series is often assumed to consist of a few components (see for example Cleveland et al. (1990) or Ladiray and Quenneville (2001)). In this initial example, we consider the simplest case when a time series is assumed to consist of three components: seasonal, trend and remainder. The seasonal component is usually assumed to have a repeating pattern which changes very slowly or stays constant over time. The trend component is usually considered to change faster than the seasonal component. The remainder component is the most quickly changing part. The whole time series is calculated as a function of these three parts. Often they are simply added (additive seasonality) or multiplied (multiplicative seasonality) to each other to recover the original data.



(a) *Original data*

(b) *Log data, classical additive decomposition*

Figure 1: *Supermarket and grocery stores turnover in New South Wales.*

The time series shown in Figure 1a is supermarket and grocery stores turnover in New South Wales from 2000 to 2009 inclusive. It provides a classic example of data with multiplicative seasonality.

There are a few popular methods to deal with such data. One of them is the classical decomposition method for additive seasonality. This method assumes that the seasonal component is additive and is not changing over time. To be able to apply this method to data with multiplicative seasonality, logs are taken as shown in Figure 1b. In this paper, we consider only additive decomposition, applied to either the original or the logged data. See McElroy (2010) for a discussion of the benefits of a direct implementation of multiplicative decomposition.

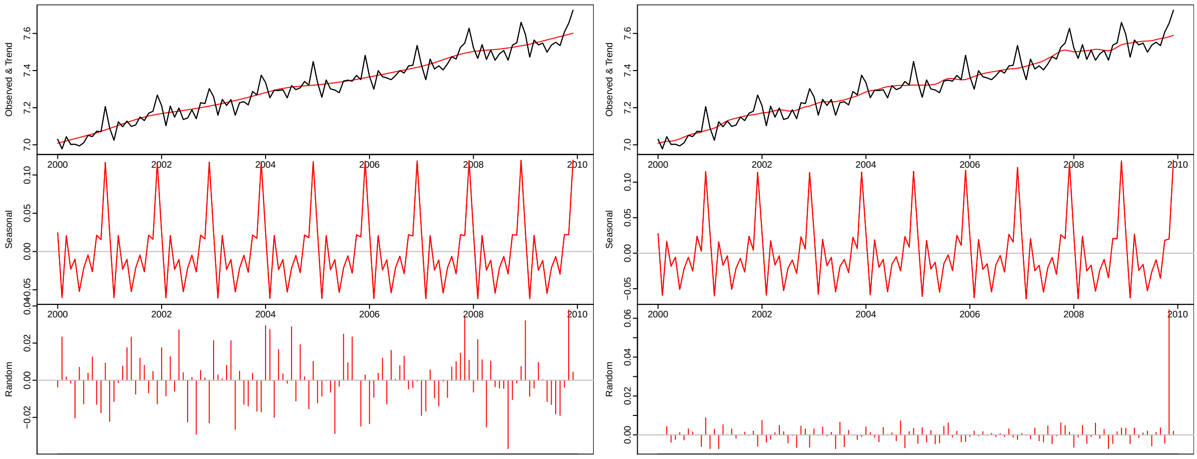


Figure 2: *Supermarket and grocery stores log turnover in New South Wales decomposed with STL and X-13ARIMA-SEATS.*

STL and X-13-ARIMA-SEATS are two other very well known methods; an application of each of them to the data in Figure 1 is presented in Figure 2. Both methods are iterative in nature and rather complex in details; see Cleveland et al. (1990) for STL and Findley et al. (1998) for the capabilities and features of X-12-ARIMA, a predecessor of X-13-ARIMA-SEATS.

Many decomposition methods currently available, including STL and X-13-ARIMA-SEATS, lack an underlying stochastic model. As a consequence confidence intervals are not readily available for the decomposed components. The methods which have models underpinning them (for example BSM, BATS, TBATS; see Harvey (1990) and De Livera et al. (2011)) still do not provide such an option, probably due to the complexity of implementing such a procedure.

In contrast, our new proposed method STR provides a statistical model and we can easily calculate confidence intervals for the decomposed components.

3 A simple STR model

The first and the simplest STR model describes a time series Y_t consisting of three parts:

$$Y_t = T_t + S_t + R_t, \tag{1}$$

where T_t is the trend, S_t is the additive seasonal component, and R_t is the "remainder" component. Time $t \in \{1,\ldots,n\}$ and we have k seasons: $1,\ldots,k$. Let us also denote function

$\mathrm{sn}(t) \in \{1,\ldots,k\}$ which transforms time t into the corresponding season $\mathrm{sn}(t)$. For example when we consider weekly seasonality $\mathrm{sn}(t)$ can be defined as $\mathrm{sn}(t) = t \mod 7$.

We assume that the seasonal component S_t has some stable repeating pattern. At time t we observe (indirectly, through equation (1)) only one element of the seasonal pattern. It is reasonable to ask what the other components of this pattern are at that very moment t. For example with weekly seasonality, if on Wednesday we observe some seasonal value, we consider the question "What is the value of seasonal component corresponding to Friday on this Wednesday?". In other words we propose to define along with S_t, other "invisible" components responsible for seasons other than $\mathrm{sn}(t)$. In this way, we treat the seasonal pattern as two-dimensional, $\langle S_{i,t}\rangle_{i=1}^{k}$, and we assume that $S_t = S_{\mathrm{sn}(t),t}$ (here and further S with one index will represent a vector of seasons from (1) and S with two indexes will represent a matrix of seasonal shapes). Thus (1) can be rewritten as:

$$Y_t = T_t + S_{\mathrm{sn}(t),t} + R_t \tag{2}$$

where $S = [S_{s,t}]$ is $k \times n$ matrix, k is number of seasons and n is length of the time series.

This representation allows us to impose simple constraints on the seasonal patterns represented by matrix $S = [S_{s,t}]$. The whole model can be described as follows:

- The remainder terms R_t are i.i.d. $\mathcal{N}\left(0, \sigma_R^2\right)$;
- The trend terms are smooth such that $\Delta^2 T_t = T_{t+1} - 2T_t + T_{t-1}$ are i.i.d. $\mathcal{N}\left(0, \sigma_T^2\right)$;
- The seasonal terms must sum to zero so that, for any t, they have the property $\sum\limits_{s} S_{s,t} = 0$;
- Each of the seasonal terms are also smoothly changing over time, so that $\forall t$ vectors $\langle \Delta_t^2 S_{s,t}\rangle_{s=1}^{k} = \langle S_{s,t+1} - 2S_{s,t} + S_{s,t-1}\rangle_{s=1}^{k}$ are i.i.d. by t and distributed $\mathcal{N}\left(0, \sigma_S^2 \Sigma_S\right)$ inside the vectors, where Σ_S is a $k \times k$ matrix which can be considered the covariance matrix of k random variables $\xi_s = \eta_s - \frac{1}{k}\sum\limits_{r=1}^{k} \eta_r$ obtained from i.i.d. $\mathcal{N}(0,1)$ random variables $\eta_1, \ldots, \eta_k$;
- The parameters of the model are given by σ_R, σ_T, σ_S, T_0, T_1, $\langle S_{s,0}\rangle_{s=1}^{k}$ and $\langle S_{s,1}\rangle_{s=1}^{k}$ (or $\langle S_{s,n}\rangle_{s=1}^{k}$).

It should be noted that in this, the simplest form of STR, the seasonal component is presented as a two-dimensional array only to ensure that the seasonal components add to zero for any t. Later we will apply additional restrictions on the seasonal component and the two-dimensional structure will become even more important.

Another way to present the model (2) is to write it in the form of a state-space model in the spirit of Harvey (1985) or Harvey and Peters (1990), or as a multivariate ARIMA model. For example, T_t can be considered an ARIMA(0,2,0) process and the seasonal component is a multivariate ARIMA(0,2,0) process with contemporaneously correlated errors (but no autocorrelation).

4 Estimating an STR model

4.1 Maximum likelihood estimation

We define vector ℓ as $\ell = \langle T_t \rangle_{t=1}^{n}$ and S^- as matrix S without the last row. Since each column of S must sum to 0, we can write $S = P_S S^-$ for some matrix P_S. Let us also define $s = \mathrm{vec}(S^-)$ to be a vector of length $n(k-1)$ representing the seasonal components, and $\Xi = I_{n-2} \otimes \Xi^-$ where Ξ^- is a $(k-1)\times(k-1)$ matrix obtained from Σ_S by removing the last row and the last column. There also exists a matrix P such that

$$\mathrm{vec}(S) = Ps. \tag{3}$$

Thus, the minus log likelihood function for this model is given (up to a constant) by

$$-\log(\mathcal{L}) = \left\| \frac{y - Qs - \ell}{\sigma_R} \right\|_{L_2}^2 + \left\| \frac{\Xi^{-\frac{1}{2}} D_s s}{\sigma_S} \right\|_{L_2}^2 + \left\| \frac{D_\ell \ell}{\sigma_L} \right\|_{L_2}^2, \tag{4}$$

where

- Q is a $n \times n(k-1)$ matrix that computes elements $\langle S_{\mathrm{sn}(t),t} \rangle_{t=1}^{n}$ from vector $s = \mathrm{vec}(S^-)$;
- D_s is a $(n-2)(k-1) \times n(k-1)$ matrix that computes second differences along the time dimension: $\langle \Delta_t^2 S^-_{s,t} \rangle_{s=1}^{k} = \langle S^-_{s,t+1} - 2S^-_{s,t} + S^-_{s,t-1} \rangle_{s=1}^{k-1}$ for $2 \le t \le n-1$;
- D_ℓ is $(n-2)\times n$ matrix that calculates second differences of ℓ: $\Delta^2 T_t = T_{t+1} - 2T_{t-1} + T_{t-1}$ for $2 \le t \le n-1$.

Thus, maximum likelihood estimates are obtained by minimizing

$$\left\| y - Qs - \ell \right\|_{L_2}^2 + \left\| \frac{\sigma_R}{\sigma_S} \Xi^{-\frac{1}{2}} D_s s \right\|_{L_2}^2 + \left\| \frac{\sigma_R}{\sigma_L} D_\ell \ell \right\|_{L_2}^2 \tag{5}$$

over s and ℓ.

We can also note that (4) corresponds to the minus log likelihood function for the following linear model (here we use an approach similar to that described in Dokumentov and Hyndman (2013)):

$$y_{ext} = X\beta + \varepsilon, \tag{6}$$

where $\beta = [s', \ \ell']'$ is a vector of unknown coefficients, s is a vector of seasonal components; ℓ is a vector containing the trend and ε is a vector of i.i.d. errors. Observations are defined by matrix

$$X = \begin{bmatrix} Q & I \\ \lambda_s \Xi^{-\frac{1}{2}} D_s & 0 \\ 0 & \lambda_\ell D_\ell \end{bmatrix}, \tag{7}$$

where $\lambda_s = \frac{\sigma_R}{\sigma_S}$ and $\lambda_\ell = \frac{\sigma_R}{\sigma_L}$ (later λ_s and λ_ℓ will be estimated directly using cross validation without reference to σ_R, σ_S and σ_L). Predictors are defined by vector $y_{ext} = [y', \ 0']'$, which is vector y extended with zeros to make it conform to matrix X defined above. All errors are i.i.d. $\mathcal{N}(0, \sigma^2)$ for some unknown σ (in Section 9 the assumption of normality of the errors will be relaxed).

Since STR model and the linear model (6) have the same likelihood functions, their maximum likelihood solutions will be identical and are given by

$$\hat{\beta} = (X'X)^{-1}X'y_{ext} = \hat{\beta} = (X'X)^{-1}[Q \ \ I]'y. \tag{8}$$

Taking into account that the covariance matrix of $\varepsilon_{ext} = y_{ext} - X\hat{\beta}$ has the following form:

$$\Sigma = \begin{bmatrix} \sigma^2 I_n & 0 \\ 0 & 0 \end{bmatrix},$$

where σ is the standard deviation of the residuals corresponding to the $[Q \ \ I]$ part of matrix X, the covariance matrix of solution $\hat{\beta}$ can be calculated as

$$\Sigma_{\hat{\beta}} = (X'X)^{-1}X'\Sigma X(X'X)^{-1} = \sigma^2(X'X)^{-1}[Q \ \ I]'[Q \ \ I](X'X)^{-1}. \tag{9}$$

We estimate σ using cross-validated residuals (see Section 4.2) instead of residuals of the fit. This ensures that σ is not underestimated if the model is over-fitted.

The trend component $\hat{T}$ and the corresponding confidence intervals can be obtained directly from $\hat{\beta}$ and $\Sigma_{\hat{\beta}}$ ($\hat{T}$ is represented by the ℓ part of β, $\beta = [s', \ \ell']'$, and confidence intervals are calculated in the usual way for a linear regression). In order to obtain the seasonal components $\hat{S}$, $\hat{\beta}$ needs to be linearly transformed with some matrix R (similar to P in (3)) to recalculate every last seasonal component. Using matrix R, a new covariance matrix and confidence intervals for $\hat{S}$ can also be calculated.

Let us consider a variation of model (6) where we allow the errors to be correlated with covariance matrix Σ_y. Then the covariance matrix of $\varepsilon_{ext} = y_{ext} - X\hat{\beta}$ has the form

$$\Sigma = \begin{bmatrix} \Sigma_y & 0 \\ 0 & 0 \end{bmatrix}, \tag{10}$$

and if matrix Σ_y is invertible, the solution is

$$\hat{\beta} = (X'WX)^{-1}[Q \;\; I]'\Sigma_y^{-1}y$$

where

$$W = \begin{bmatrix} \Sigma_y^{-1} & 0 \\ 0 & I \end{bmatrix}.$$

The covariance matrix of the solution will be:

$$\Sigma_{\hat{\beta}} = (X'WX)^{-1}[Q \;\; I]'\Sigma_y^{-1}[Q \;\; I](X'WX)^{-1}.$$

4.2 Smoothing parameter estimation

The model (6) also requires specification or estimation of the parameters λ_s and λ_ℓ. We propose using leave-one-out cross validation to estimate them. Using a model with minimal cross validation will allow choosing the model which "absorbs" information as much as possible and noise (uncorrelated errors) as little as possible.

Since model (6) is a linear model, the leave-one-out cross validation residuals can be calculated using (Seber and Lee, 2003)

$$\mathrm{cv}_i = \frac{y_i - \hat{y}_i}{1 - h_{ii}},$$

where y_i is the ith element of vector y, $\hat{y}_i$ is the ith element of vector $\hat{y} = Hy$ and h_{ii} is the ith diagonal element of the hat matrix $H = X(X'X)^{-1}X'$. Therefore, we can use the well-known formula for cross validation for linear regression (see for example Ruppert et al. (2003)):

$$\mathrm{SSE(cv)} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2 .$$

By minimising SSE(cv), STR finds optimal parameters λ_s and λ_ℓ. The problem of minimising SSE(cv) can be complex, for example with many local minima, and we have no method which guarantees finding the global minima. However, in practice rather simple methods often work well. We use R and the Nelder-Mead method as implemented in the `optim()` function from the "stats" package in R (R Core Team, 2015) for such optimisation.

On the other hand, this approach does not work for model (10) when the errors are correlated, or for very big matrices X due to the difficulties of inverting $X'X$. For such scenarios we use K-fold cross validation, as discussed in Section 9.

5 STR with two-dimensional restrictions on seasonal changes

Optimization problem (5) can be extended to constrain second discrete derivatives $\frac{\partial^2}{\partial s \partial t}$ and $\frac{\partial^2}{\partial s^2}$ of the seasonal component (in our case it is a two-dimensional surface) in addition to $\frac{\partial^2}{\partial t^2}$. This will imply a new model, but let us describe the optimisation problem first:

$$(s,\ell) = \operatorname{arg\,min}\left[\left\| y - Qs - \ell \right\|_{L_2}^2 + \|\lambda_{tt} D_{tt} s\|_{L_2}^2 + \|\lambda_{st} D_{st} s\|_{L_2}^2 + \|\lambda_{ss} D_{ss} s\|_{L_2}^2 + \|\lambda_\ell D_\ell \ell\|_{L_2}^2 \right], \tag{11}$$

where, similar to (5):

- y is a vector of observations of length n;
- s is a vector of two-dimensional seasonal components (an n×(k-1) matrix) presented as a vector of length $n(k-1)$;
- ℓ is a vector of length n representing the trend component;
- D_{tt}, D_{st} and D_{ss} are matrices which compute second differences for the two-dimensional seasonal component along the time, time-season and season dimensions, respectively;
- D_ℓ is a matrix which calculates second differences for trend ℓ;
- λ_{tt}, λ_{st}, λ_{ss} and λ_ℓ are parameters to be selected (they have a similar meaning as in (7)).

It is now evident that the seasonal two-dimensional surface actually has the topology of a tube, where dimension t spreads infinitely into two directions, but dimension s is "circular" (season 1 is connected to season 2, season 2 is connected to season 3, …, season $(k-1)$ is connected to season k and season k is connected back to season 1). It should be noted that matrices D_{st} and D_{ss} take the "tube" topology into account to calculate proper "circular" differences

Similar to the problem discussed in Section 4, this optimisation problem corresponds to the following linear regression problem and the model:

$$y_{ext} = X\beta + \varepsilon, \tag{12}$$

where $y_{ext} = [y',\ 0']'$, $\beta = [s',\ \ell']'$ is a vector of unknown coefficients, s is a vector of seasonal components, ℓ is a vector containing the trend, and

$$X = \begin{bmatrix} Q & I \\ \lambda_{tt} D_{tt} & 0 \\ \lambda_{st} D_{st} & 0 \\ \lambda_{ss} D_{ss} & 0 \\ 0 & \lambda_\ell D_\ell \end{bmatrix}.$$

All errors are i.i.d. $\mathcal{N}\left(0,\sigma^2\right)$ for some unknown σ.

Using (8) and (9) as in Section 4, we find the solution of this new problem and the corresponding confidence intervals. Using cross validation (Sections 4.2 and 9) and an optimisation procedure, we can find good values for λ_{tt}, λ_{st}, λ_{ss} and λ_ℓ.

Let us consider an example of the described decomposition. Figure 3 shows the same data as discussed in Section 1, but now decomposed using STR. As we can see, the result of the decomposition is similar but confidence intervals are now provided for the trend and the seasonal component.

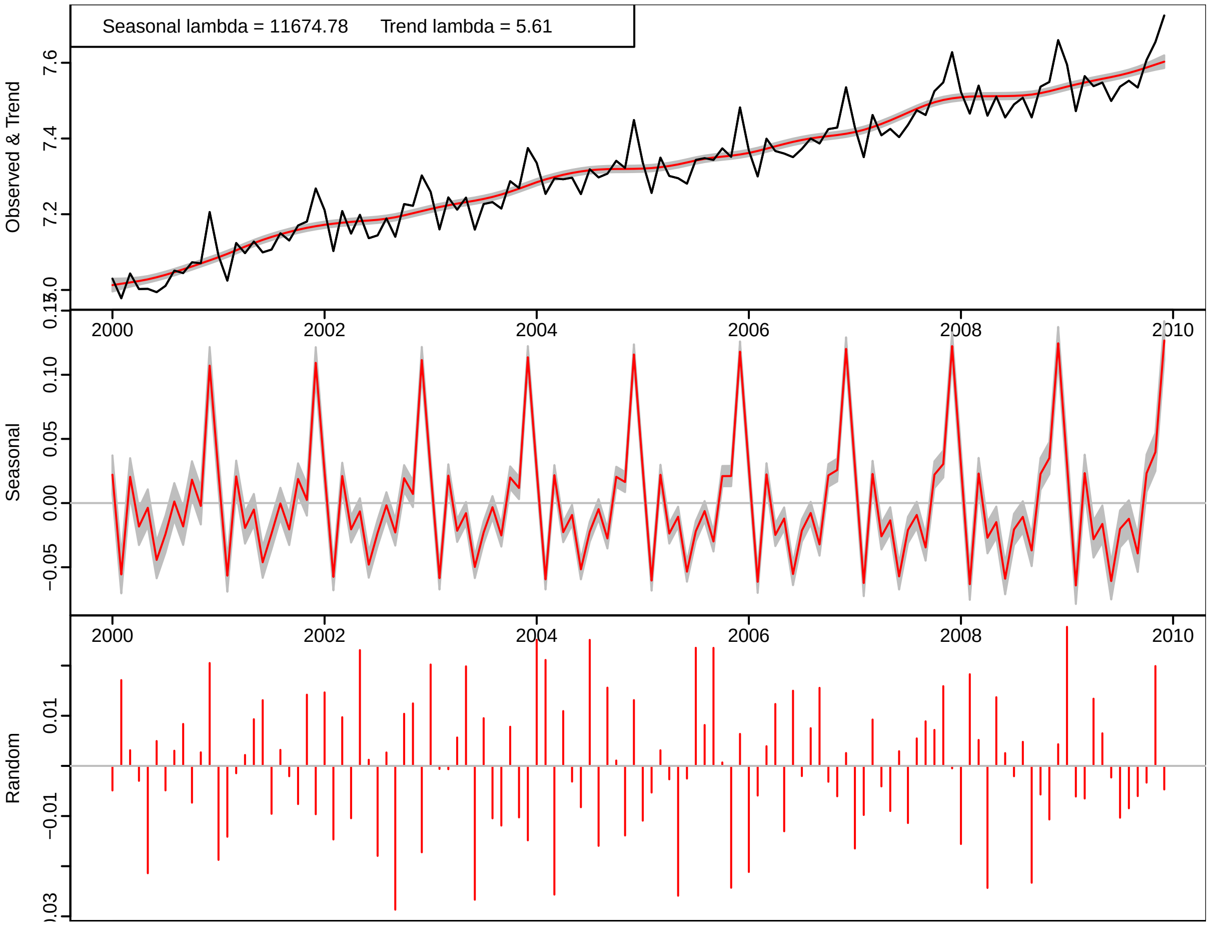


Figure 3: *Supermarket and grocery stores log turnover in New South Wales decomposed with STR. The original data are plotted in black; decomposed trend, seasonal and random components are in red; confidence intervals are in grey.*

6 STR with multiple seasonality and static, flexible and seasonal predictors

Let us revisit the concept of seasonality. Originally we considered it as a single repeating pattern which affects data in an additive manner. Such a pattern is usually a consequence of some scheduled or naturally repeating activity corresponding to the day of the week, hour of the day, etc. (see for example Makridakis et al. (2008)).

Nevertheless, time series representing real world activity are often affected by *several* schedules representing different seasonal periods. For example, electricity demand typically depends on the time of day, day of week, and day of year (Hyndman and Fan, 2010). However, in Turkey, electricity demand is also affected by the Islamic Hijri calendar due to various religious activities such as fasting during the month of Ramadan (De Livera et al., 2011). Another complicating factor is that the periods of these calendars are fractional. More complex examples can include patterns with periods changing in length with time.

Regressors can also be important factors affecting time series beyond seasonal patterns. For example Findley and Soukup (2000), Findley et al. (2009) and Hood and Feldpausch (2011) consider the effect of various moving holidays on human activities. These effects are usually assumed to be linear and constant over time, although Bell and Martin (2004) considers time-varying coefficients. The magnitude of the impact of such regressors on a time series is important in further analysis and decision making in economics or finance.

In this section we consider time series that are affected by multiple seasonal patterns and linearly by regressors of three types. The first type is constant effect regressors where it is assumed the associated coefficients are constant over time. The second type is regressors with coefficients whose magnitudes change smoothly over time but where the coefficients do not exhibit any seasonal pattern. The third type is regressors with coefficients whose magnitudes change smoothly over time in a seasonal pattern. As far as we know, this last type of time-varying coefficient is new in the decomposition of time series.

We consider a time series Y_t consisting of the following parts:

$$Y_t = T_t + \sum_{i=1}^{r} S_{it} + \sum_{i=1}^{m} P_{it} + \sum_{i=1}^{\tilde{m}} \tilde{P}_{it} + \sum_{i=1}^{\tilde{\tilde{m}}} \tilde{\tilde{P}}_{it} + R_t, \tag{13}$$

where:

- T_t is the smoothly changing trend;
- S_{it} are seasonal components with corresponding seasonal periods k_i;
- P_{it} are additive components linearly depending on predictors p_i with constant coefficients;
- $\tilde{P}_{it}$ are additive components linearly depending on predictors $\tilde{p}_i$ with time-varying but non-seasonal coefficients;
- $\tilde{\tilde{P}}_{it}$ are additive components linearly depending on predictors $\tilde{\tilde{p}}_i$ with time-varying coefficients, where the coefficients have seasonal patterns with corresponding seasonal periods $\tilde{\tilde{k}}_i$;
- R_t is the “random” error.

Because this description is rather vague, we define decomposition and the model uniquely through a minimization problem which extends the optimisation problems and the corresponding models of Sections 3 and 5.

In this new optimisation problem defined below, we assume the existence of multiple seasonal periods and dependence on a number of regressors. We also assume that such dependence can change over time for some of the regressors (we call them flexible and seasonal regressors depending on whether the seasonal pattern appears in the changes).

$$
\begin{aligned}
(s,\ell,\beta,\tilde{\beta},\tilde{\tilde{\beta}}) = \arg\min \Bigg\{ & \Bigg\| y - \sum_{i=1}^{r} Q_i s_i - \ell - P\beta - \sum_{i=1}^{\tilde{m}} \tilde{P}_i \tilde{\beta}_i - \sum_{i=1}^{\tilde{\tilde{m}}} \tilde{\tilde{P}}_i \tilde{\tilde{\beta}}_i \Bigg\|_{L_2}^2 \\
& + \sum_{i=1}^{r} \left(\left\| \lambda_{tt_i} D_{tt_i} s_i \right\|_{L_2}^2 + \left\| \lambda_{st_i} D_{st_i} s_i \right\|_{L_2}^2 + \left\| \lambda_{ss_i} D_{ss_i} s_i \right\|_{L_2}^2 \right) + \| \lambda_\ell D_{tt} \ell \|_{L_2}^2 \\
& + \sum_{i=1}^{\tilde{m}} \left\| \tilde{\lambda}_i D_{tt} \tilde{\beta}_i \right\|_{L_2}^2 + \sum_{i=1}^{\tilde{\tilde{m}}} \left(\left\| \tilde{\tilde{\lambda}}_{tt_i} \tilde{\tilde{D}}_{tt_i} \tilde{\tilde{\beta}}_i \right\|_{L_2}^2 + \left\| \tilde{\tilde{\lambda}}_{st_i} \tilde{\tilde{D}}_{st_i} \tilde{\tilde{\beta}}_i \right\|_{L_2}^2 + \left\| \tilde{\tilde{\lambda}}_{ss_i} \tilde{\tilde{D}}_{ss_i} \tilde{\tilde{\beta}}_i \right\|_{L_2}^2 \right) \Bigg\},
\end{aligned}
\tag{14}
$$

where

- y is a vector of observations of length n;
- Q_i are $n \times n(k_i - 1)$ matrices which compute observable seasonal elements from seasonal vectors s_i, each of which represents the corresponding two-dimensional seasonal component;
- P is an $n \times m$ matrix of static predictors, where every predictor occupies a single column;
- β is an m-vector of coefficients of the static regressors;
- $\tilde{P}_i = \text{diag}(\tilde{p}_i)$ for $1 \le i \le \tilde{m}$ is the ith predictor matrix with values of the ith predictor arranged along the diagonal and all other values equal to zero;
- $\tilde{\beta}_i$ is the ith vector of changing coefficients for the ith flexible regressor;
- $\tilde{\tilde{P}}_i = \text{diag}(\tilde{p}_i)$ for $1 \le i \le \tilde{\tilde{m}}$ is the ith predictor matrix with values of the ith predictor arranged along the diagonal and all other values equal to zero;
- $\tilde{\tilde{\beta}}_i$ is the ith vector of changing coefficients for the ith seasonal regressor;
- D_{tt} is a matrix taking second differences of a vector representing the trend or flexible coefficients;

- D_{tt_i}, D_{st_i} and D_{ss_i} are matrices taking second differences of the ith seasonal component in the time, time-season and season dimensions;
- $\tilde{\tilde{D}}_{tt_i}$, $\tilde{\tilde{D}}_{st_i}$ and $\tilde{\tilde{D}}_{ss_i}$ are matrices taking second differences of the ith seasonal component in time, time-season and season dimensions;
- λ_{tt_i}, λ_{st_i}, λ_{ss_i}, λ_ℓ, $\tilde{\lambda}_i$, $\tilde{\tilde{\lambda}}_{tt_i}$, $\tilde{\tilde{\lambda}}_{st_i}$, $\tilde{\tilde{\lambda}}_{ss_i}$ are the parameters.

We can also note that the optimisation problem can easily be adapted in case some (or many) observations y are missing (although the model does not allow the predictors to be missing for observed values of y). The adaptation involves excluding missing values in vector y and the corresponding rows in the matrices which are in front of vectors $s_1,\ldots,s_r$, ℓ, $\bar{\beta}$, $\tilde{\beta}_1,\ldots,\tilde{\beta}_{\tilde{m}}$, $\tilde{\tilde{\beta}}_1,\ldots,\tilde{\tilde{\beta}}_{\tilde{\tilde{m}}}$ (assuming that in front of ℓ there is an identity matrix I_n) in expression (14).

This provides a very natural way to forecast by treating future observations as missing, and then estimating for forecast horizon h. If the covariance matrix of y can be estimated, the confidence intervals also can be found the standard for linear regression way.

Although this is a very generic form of STR, in practice it does not need to be that complex. In most cases, the minimization problem contains only a few of the terms in (14).

The optimisation problem (14) corresponds to the following linear model:

$$y_{ext} = X\beta + \varepsilon, \tag{15}$$

where $y_{ext} = [y',\ 0']'$,

$$\beta = [s'_1,\ldots,s'_r,\ell',\beta',\tilde{\beta}'_1,\ldots,\tilde{\beta}'_{\tilde{m}},\tilde{\tilde{\beta}}'_1,\ldots,\tilde{\tilde{\beta}}'_{\tilde{\tilde{m}}}]' \tag{16}$$

is a vector of unknown coefficients and

$$X = \begin{bmatrix}
Q_1 & \dots & Q_r & I & R & \tilde{R}_1 & \dots & \tilde{R}_{\tilde{m}} & \tilde{\tilde{R}}_1 & \dots & \tilde{\tilde{R}}_{\tilde{\tilde{m}}} \\
\lambda_{tt_1} D_{tt_1} & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
\lambda_{st_1} D_{st_1} & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
\lambda_{ss_1} D_{ss_1} & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \ddots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & \lambda_{tt_r} D_{tt_r} & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & \lambda_{st_r} D_{st_r} & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & \lambda_{ss_r} D_{ss_r} & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & \lambda_\ell D_{tt} & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & \tilde{\lambda}_1 D_{tt} & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & \tilde{\lambda}_m D_{tt} & 0 & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \tilde{\tilde{\lambda}}_{tt_1} \tilde{\tilde{D}}_{tt_1} & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \tilde{\tilde{\lambda}}_{st_1} \tilde{\tilde{D}}_{st_1} & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \tilde{\tilde{\lambda}}_{ss_1} \tilde{\tilde{D}}_{ss_1} & \dots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \ddots & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & \tilde{\tilde{\lambda}}_{tt_{\tilde{\tilde{m}}}} \tilde{\tilde{D}}_{tt_{\tilde{\tilde{m}}}} \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & \tilde{\tilde{\lambda}}_{st_{\tilde{\tilde{m}}}} \tilde{\tilde{D}}_{st_{\tilde{\tilde{m}}}} \\
0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & \tilde{\tilde{\lambda}}_{ss_{\tilde{\tilde{m}}}} \tilde{\tilde{D}}_{ss_{\tilde{\tilde{m}}}}
\end{bmatrix} \tag{17}$$

for some fixed parameters

$$\lambda = \left[\lambda_{tt_1}, \lambda_{st_1}, \lambda_{ss_1}, \dots, \lambda_{tt_r}, \lambda_{st_r}, \lambda_{ss_r}, \lambda_\ell, \tilde{\lambda}_1, \dots, \tilde{\lambda}_{\tilde{m}}, \tilde{\tilde{\lambda}}_{tt_1}, \tilde{\tilde{\lambda}}_{st_1}, \tilde{\tilde{\lambda}}_{ss_1}, \dots, \tilde{\tilde{\lambda}}_{tt_{\tilde{\tilde{m}}}}, \tilde{\tilde{\lambda}}_{st_{\tilde{\tilde{m}}}}, \tilde{\tilde{\lambda}}_{ss_{\tilde{\tilde{m}}}}\right].$$

If some values of vector λ are zeros, the corresponding rows of matrix X can (and should) be removed (as they have no effect and removing them improves computation time). All errors are i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$ for some unknown σ.

We need to note that the combined number of coefficients (the length of β in equation (16)) that we need to estimate is usually much larger than the number of observations (the length of y). This does not cause problems of estimation since the coefficients are regularised (restricted), and therefore the estimation is performed against observations presented by y_{ext} (regression (15)), where y_{ext} is longer than the number of estimated coefficients.

We can find the solution of this new problem and the corresponding confidence intervals using (8) and (9), similarly to the models discussed in Sections 4 and 6. Using the approach of Section 4.2, we can find good values for λ, although the task becomes more difficult as we need to optimize over a larger parameter space.

An example of decomposition of data with multiple seasonality is presented in Figure 4. The data was originally published in Weinberg et al. (2007). It consists of 10140 observations, although we only use the first 4056 observations (working days from 3 March 2003 to 3 April 2003) to demonstrate our method. The data has two seasonal patterns: the daily pattern has a period of 169 observations and the weekly pattern has a period of $169 \times 5 = 845$ observations.

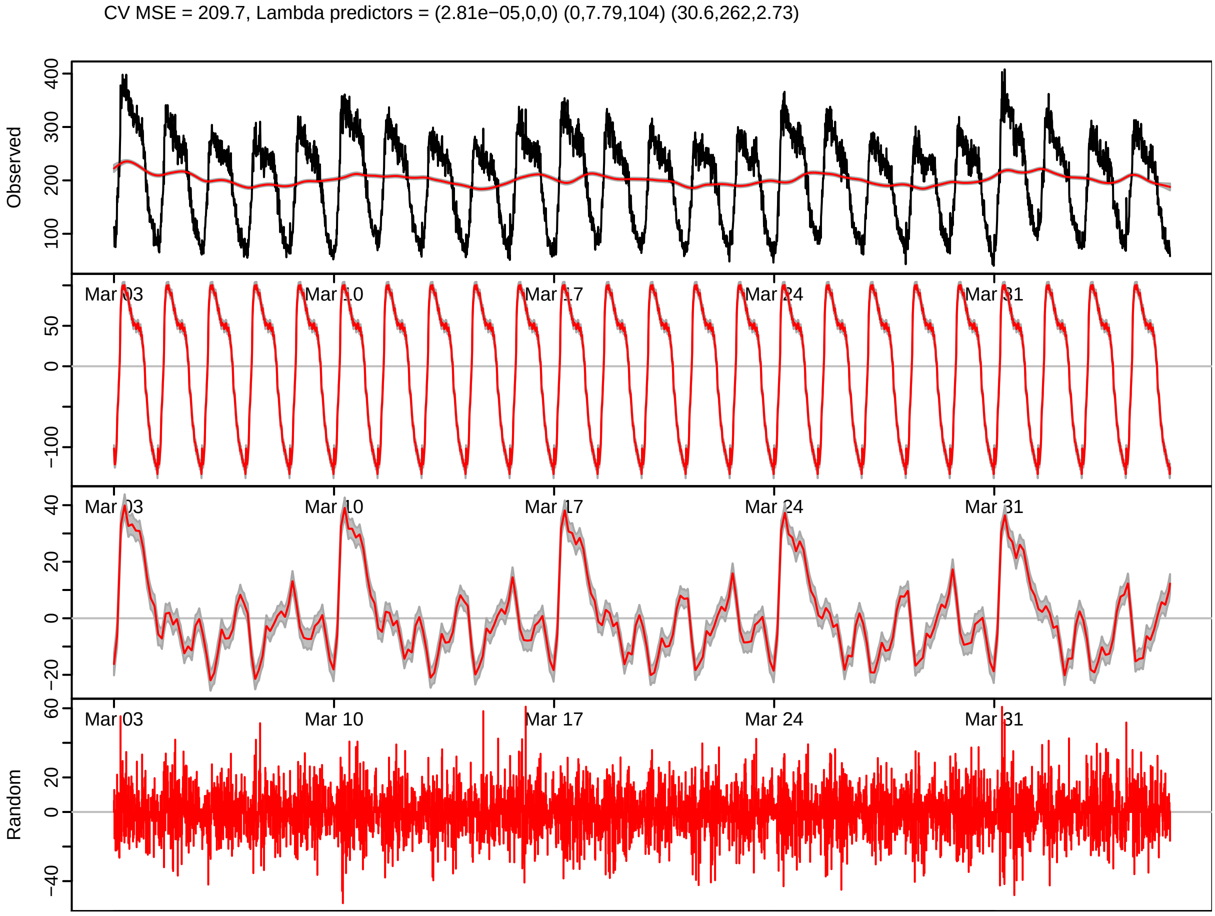


Figure 4: *The number of calls to a big North American retail bank per 5-minute interval (from 7:00 a.m. to 9:05 p.m., weekdays) decomposed with STR. The original data are plotted in black; decomposed trend, seasonal and random components are in red; confidence intervals are in grey.*

Another example, with multiple seasonality and time varying regressors can be found in Section 10.

7 Functional coefficient models

Multiple seasonality and multiple regressors (both flexible and seasonal) bring another problem: the length of β (defined in (16)) and the dimensions of X (defined in (17)) become too large for easy computation. This problem becomes more evident as the model we consider becomes more complex.

To solve this problem, let us consider approximating the varying trend ℓ as a smoothly varying function of time using a linear combination of smooth basis functions. Thus we can write

$$\ell(t) = \sum_{j=1}^{q} \gamma_j \phi_j(t).$$

Then for some matrix Φ, vector ℓ can be presented as $\ell = \Phi\gamma$.

Similarly seasonal functional components $s_i(t)$ can be defined and approximated using a linear combination of smooth basis functions:

$$s_i(t,s) = \sum_{j=1}^{q_i} \gamma_j^{(i)} \phi_j^{(i)}(t,s).$$

Then for some matrices $\Phi^{(i)}$, vectors s_i can be presented as $s_i = \Phi^{(i)}\gamma^{(i)}$.

Time varying and seasonally varying components of formula (13) can also be presented through functional components and functional coefficients in a similar way.

The basis functions here could be splines, Fourier terms (in such cases it is reminiscient of Livera et al., 2010), wavelets etc. In this article we use only piecewise linear regression splines.

Finally, noting that regularisation components can be written as norms of linear transformations of the corresponding gamma coefficients, the minimisation problem (14) can be rewritten in terms of functional coefficients:

$$\begin{aligned}(\gamma^{(\cdot)}, \gamma, \beta, \tilde{\gamma}, \tilde{\tilde{\gamma}}) = \arg\min \Bigg\{ & \Bigg\| y - \sum_{i=1}^{r} \Phi^{(i)}\gamma^{(i)} - \Phi\gamma - P\beta - \sum_{i=1}^{\tilde{m}} \tilde{\Phi}_i \tilde{\gamma}_i - \sum_{i=1}^{\tilde{\tilde{m}}} \tilde{\tilde{\Phi}}_i \tilde{\tilde{\gamma}}_i \Bigg\|_{L_2}^2 \\ & + \sum_{i=1}^{r} \Big(\big\| \lambda_{tt_i} \Psi_{tt_i} \gamma_i \big\|_{L_2}^2 + \big\| \lambda_{st_i} \Psi_{st_i} \gamma_i \big\|_{L_2}^2 + \big\| \lambda_{ss_i} \Psi_{ss_i} \gamma_i \big\|_{L_2}^2 \Big) + \big\| \lambda_\ell \Psi_{tt} \gamma \big\|_{L_2}^2 \\ & + \sum_{i=1}^{\tilde{m}} \big\| \tilde{\lambda}_i \Psi_{tt} \tilde{\gamma}_i \big\|_{L_2}^2 + \sum_{i=1}^{\tilde{\tilde{m}}} \Big(\big\| \tilde{\tilde{\lambda}}_{tt_i} \tilde{\tilde{\Psi}}_{tt_i} \tilde{\tilde{\gamma}}_i \big\|_{L_2}^2 + \big\| \tilde{\tilde{\lambda}}_{st_i} \tilde{\tilde{\Psi}}_{st_i} \tilde{\tilde{\gamma}}_i \big\|_{L_2}^2 + \big\| \tilde{\tilde{\lambda}}_{ss_i} \tilde{\tilde{\Psi}}_{ss_i} \tilde{\tilde{\gamma}}_i \big\|_{L_2}^2 \Big) \Bigg\}, \end{aligned} \quad (18)$$

where

- the various γ vectors are functional coefficients which are used to represent the corresponding components of decomposition defined by formula (13);
- the various Φ_i matrices transform the corresponding functional γ coefficients into corresponding components of representation (13);
- the various Ψ matrices allow the calculation of second derivatives as linear transformations of the corresponding functional coefficients;
- P and β are defined in Section 6.

This approach leads to a reduction in the number of estimated coefficients to 3–4 times the length of the time series (see note in Section 6), and it dramatically improves the computational performance.

Interestingly, this approach directs us, with no additional effort, to a solution of another problem, namely seasonality with a fractional or varying period. As we can note from (18), seasonality is hidden in the Ψ matrices, which take second discrete seasonal derivatives of seasonal components. When seasonal components are functions, such matrices can be written for fractional seasonality, since even in this case second derivatives are linear functions of the corresponding functional coefficients.

8 Seasonality with complex topology

A seasonal two-dimensional surface can have a topology different from the "tube". Again, our functional approach leads naturally to modelling seasonality with more complex topology. Let us again clarify with another example. Suppose we are going to model some social behaviour (electricity demand for instance) during working days and holidays, including weekends. The topology modelling the human behaviour is shown in Figure 5.

The left circle represents a working day, the right circle represents a holiday. They are connected by lines representing transition periods. Points A and C represent hour 0 of a day, points B and D represent hour 12. Every day has 24 hours and the Transition periods take 12 hours.

According to the diagram, a working day can follow a working day. Otherwise a working day can flow until hour 12 when transition period type 1 starts (line B – C), which goes for 12 hours and

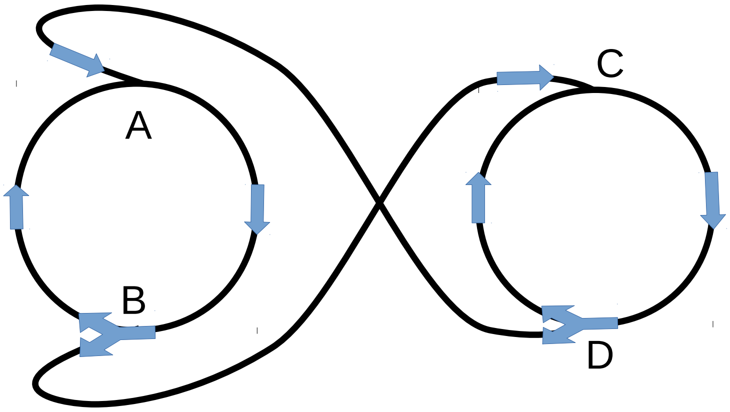


Figure 5: *Topology of complex seasonality modelling human behaviour during working days, holidays and transition periods.*

traverses into a holiday (the right circle). Similarly, a holiday can follow a holiday. Otherwise a holiday flows until hour 12 when transition period type 2 starts (line D – A), which goes for 12 hours and traverses into a working day (the left circle).

Equivalently, the topology shown in Figure 5 can be described as two connected cylinders.

Differencing matrices D_{ss} and D_{st} (or Ψ_{ss} and Psi_{st}) are defined in the standard way to restrict derivatives $\frac{\partial^2}{\partial s^2}$ and $\frac{\partial^2}{\partial s \partial t}$ for all data points except A, B, C and D. At points A, B, C and D, second derivatives can be restricted in various ways. One of them is to regularise derivatives twice: once for each path in the diagram. Matrices D_{ss} and D_{st} (or Ψ_{ss} and Psi_{st}) should be defined accordingly to reflect such an approach.

Finally we should mention that the approaches of this Section and of the previous Section 7 do not change the form of matrix X (defined by equation (17)), although they change slightly the form of the sub-matrices which comprise X.

An example of decomposition with complex topology can be found in Section 10.

9 Robust STR: the model, solution, confidence and forecasting intervals, optimal smoothing parameters

The Robust STR (referred to as RSTR from hereon) uses a linear model which is identical to model (6) and its variations in the subsequent sections, apart from the error distribution; RSTR assumes a double exponential distribution for the residuals, and trend, seasonal and predictor coefficient changes, instead of the normal distribution assumed in the STR model. The double exponential distribution leads to a different minimization problem. In this case, problem (6) is translated into the following (problems (12) and (15) are translated similarly):

$$(s,\ell) = \arg\min\left[\left\|y - Qs - \ell\right\|_{L_1} + \left\|\lambda_s \Xi^{-\frac{1}{2}} D_s s\right\|_{L_1} + \|\lambda_\ell D_\ell \ell\|_{L_1}\right]$$

which (similar to the reduction in Section 4) can be written as a quantile regression:

$$\beta = \arg\min\left\|y_{ext} - X\beta\right\|_{L_1}, \tag{19}$$

where y_{ext}, X and β are defined as before.

There is no known analytical solution for problem (19). We solve it numerically using quantile regression software (Koenker, 2013). The confidence intervals also cannot be expressed analytically and the following procedure is used to calculate them.

We follow the ideas from Dokumentov and Hyndman (2014) and reuse the following Monte-Carlo style algorithm in order to find p-confidence intervals for the coefficients β of problem (19).

1. Take m draws of vectors δ_i of length of ℓ, which have elements i.i.d. $\mathcal{N}\left(0, \sigma_r^2\right)$. We denote a set of m draws by $\delta = \bigcup_{i=1}^{m}\{\delta_i\}$.
2. Create a set of "distorted" observations $y_\delta = y + \delta$, then find a set of solutions $\beta(y_\delta)$ for them.
3. For every $1 \le j \le \text{length}(\ell)$, the $\left(\frac{p}{2}\right)$ and $\left(1 - \frac{p}{2}\right)$ quantiles of the set $\beta_j(y_\delta)$ will be the approximate p-confidence intervals for jth element of the solution.

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

The optimal smoothing parameters λ_s and λ_t are found using m-fold cross validation procedure.

For m-fold cross validation with gaps $g \in \mathbb{N}$, we split the data set into m subsets such that the observation at time $1 \le t$ belongs subset i $(0 \le i < m)$ if and only if

$$(t-1) \bmod (mg) \in [ig, \dots, (i+1)g - 1].$$

With $g = 1$ this rule gives reasonable sparsity of the subsets and we speculate that the result of such m-fold cross validation will not differ much from the result of pure leave-one-out cross validation.

Although, to exclude situations when trend flexibility is exaggerated because of high correlation of nearest observation in the data and because only single observation are missed when $g = 1$, g can be set comparable to number of observations in seasonal patterns. We experimented with g between 1 and 169 in different scenarios.

For optimisation, we use R and the method "optim" (Nelder-Mead) in the "stats" R package (R Core Team, 2015).

To show advantages of Robust STR we provide some examples which compare decomposition using STR and Robust STR on some artificial data. Let us create two datasets by spoiling data of Supermarket and grocery stores with outliers and abruptly changing the trend downwards. We will then decompose both new datasets with STR. The new data sets and the results of decomposition are shown in Figure 6.

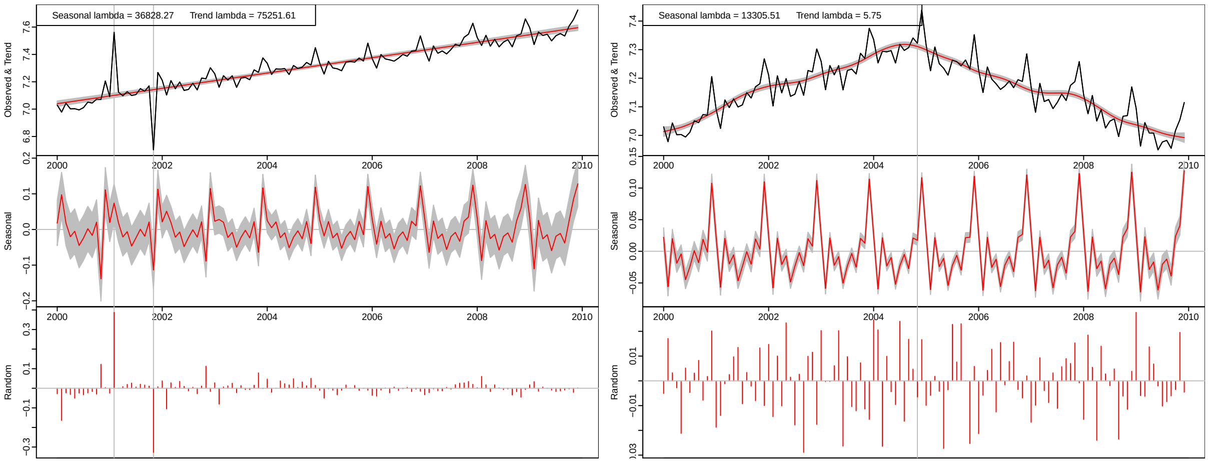


Figure 6: *Supermarket and grocery stores log turnover in New South Wales distorted by outliers and trend change then decomposed with STR. The original data are plotted in black; the decomposed trend, seasonal and random components are in red; confidence intervals are in grey.*

As we can see, outliers dramatically changed the results of the decomposition and the confidence intervals. Very high values of λ_s and λ_ℓ show that STR degraded to a purely linear trend model with seasonal components also estimated linearly. For the second data set, the change in the trend was smoothed and other components were relatively unchanged by STR.

To make STR work better in the presence of outliers we used Robust STR. The results of decomposition using Robust STR for both distorted data sets are shown in Figure 7.

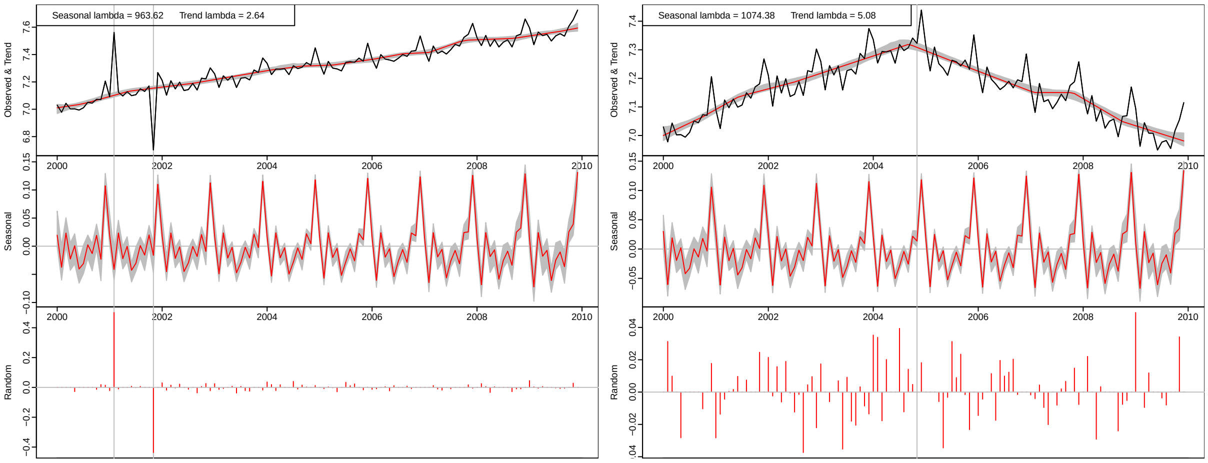


Figure 7: *Supermarket and grocery stores log turnover in New South Wales distorted by outliers and trend change then decomposed with Robust STR. The original data is plotted in black; decomposed trend, seasonal and random components are in red; confidence intervals are in grey.*

As we can see Robust STR works extremely well with outliers. It also finds the break of the trend quite well, although the confidence intervals are slightly wider than for STR. We speculate that this is due to the different distribution of errors assumed by the method.

Let us check performance of Robust STR on some other data. Let us change the data of Supermarket and grocery stores with a quick level change. Such level changes are called shifts. Figure 8 demonstrates that Robust STR deals rather well with abrupt shifts in the trend (See Section 11 for further discussion).

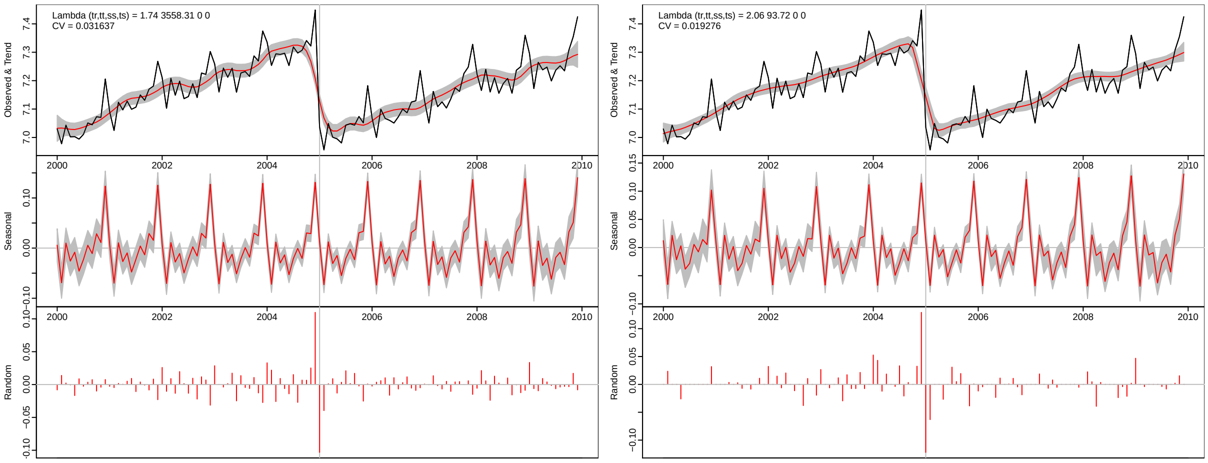


Figure 8: *Supermarket and grocery stores log turnover in New South Wales distorted by an abrupt shift in the trend decomposed with STR and Robust STR. The original data is plotted in black; decomposed trend, seasonal and random components are in red; confidence intervals are in grey.*

10 Example: electricity consumption with temperature predictors

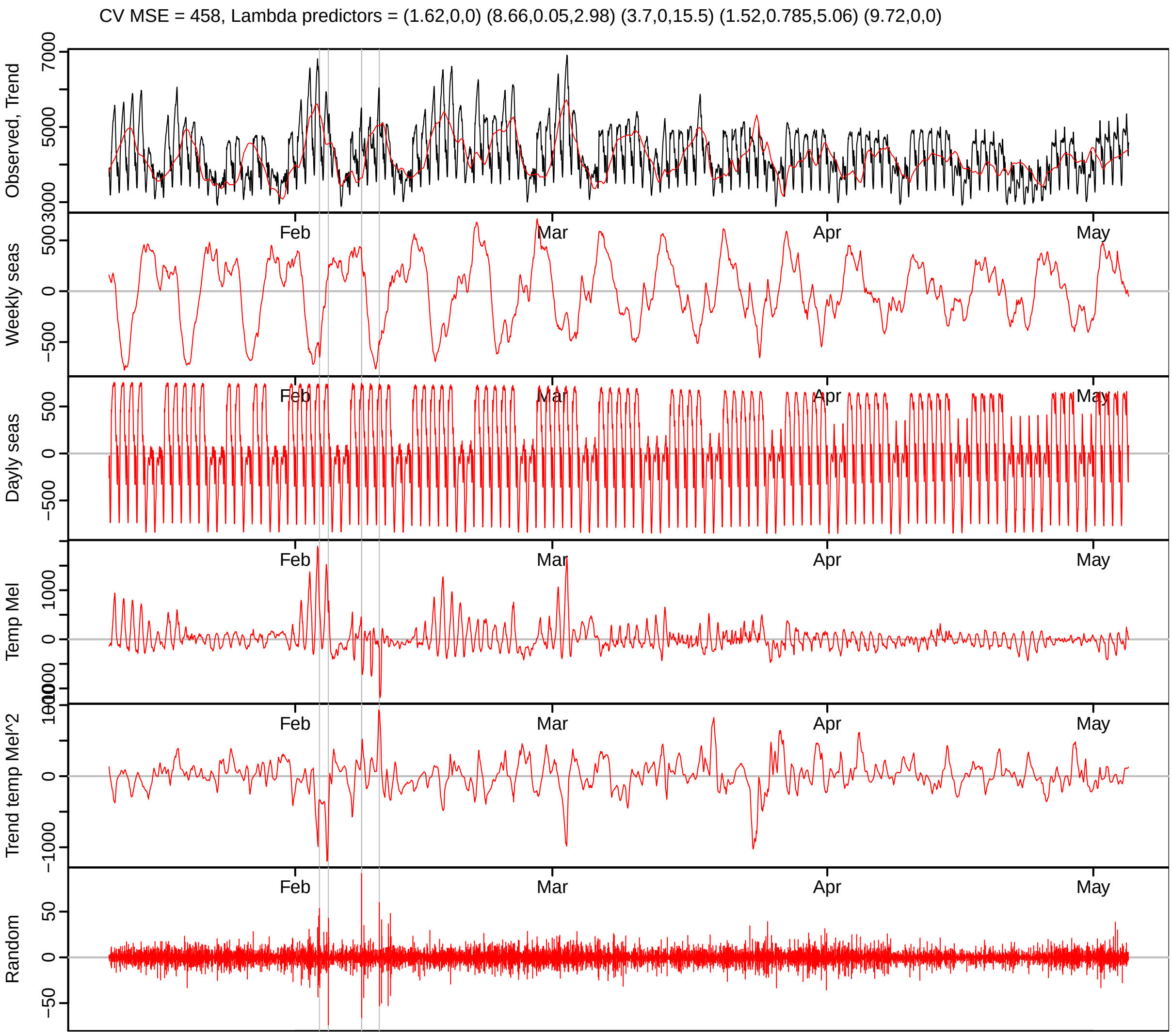


Figure 9: *Peak electricity demand in Victoria decomposed with STR. The original data is plotted in black; the decomposed trend is plotted in red over the data in the first panel; the subsequent windows contain the weekly seasonal pattern, daily seasonal pattern with complex topology, the effect of temperature in Melbourne which has a daily seasonal pattern with complex topology, the effect of squared temperatures in Melbourne which is allowed to vary over time, and the residuals. Five grey vertical lines (the third and fourth lines are very close) represent the residuals with the largest absolute values.*

In this section we present one rather complicated example of time series decomposition. The data is electricity consumption in Victoria during 115 days starting on 10 January 2000. The data set comprises maximum electricity demand in Victoria during 30 minute periods (48 observations per day). In addition, for every 30 minute period, we have the concurrent value of the air temperature at the Melbourne City weather station. We use these temperatures and their squared values as predictors.

In this example we assume that the data set has two seasonal patterns. The first pattern is a weak seasonal pattern representing the specific demand features attributable to a particular day of the week. The second pattern is a daily seasonal pattern with topology of Figure 5. Such a topology is required to be able to distinguish working days and holidays/weekends and to be able to make transitions between them. The pattern reflects the tendency to have more electricity demand during standard working hours and lower demand at nights. It also reflects the tendency to have different demand patterns during working days and holidays/weekends.

Figure 9 shows the graph of the time series decomposed with STR. The λ coefficients were chosen semi-automatically (the starting point for the minimization procedure was chosen according to the previous experiments involving minimization of the same problem with fewer predictors). Five-fold cross validation for the optimal smoothing parameters yielded RMSE = 21.4.

Two seasonal patterns and two regressors are used for the decomposition. Therefore the data is represented as the sum of six components: trend, weekly seasonality, daily seasonality with a complex topology, temperature with a daily seasonality having a complex topology, squared temperature which is time-varying but non-seasonal, and the remainder.

The shape of all components is difficult to interpret since they affect each other in a rather complex manner (for example the trend and the seasonal components also play the role of an intercept for two predictors). Although difficult to interpret, the decomposition is still interesting from at least two points of view.

The first is the point of view of a forecaster, who might not be very interested in interpreting the internal structure, but will use the decomposition "as is" for prediction.

The second is the point of view of a researcher, who is interested in discovering events affecting (or correlated with) electricity consumption. In this case, investigation of the residuals of the decomposition can provide some light. For example, the five biggest residuals in absolute value (see Table 1 and the grey vertical lines in Figure 9) can be investigated.

Date	Period	Time period	Residual
3 February 2000, Thursday	36	17:30 – 18:00	53.6
4 February 2000, Friday	36	17:30 – 18:00	−73.9
8 February 2000, Tuesday	24	11:30 – 12:00	91.6
8 February 2000, Tuesday	25	12:00 – 12:30	−66.0
10 February 2000, Thursday	24	11:30 – 12:00	59.9

Table 1: *Five biggest residuals in absolute value after a STR decomposition.*

Melbourne is famous for its unstable weather. At least three of these five outliers can be explained by unusual weather during those days.

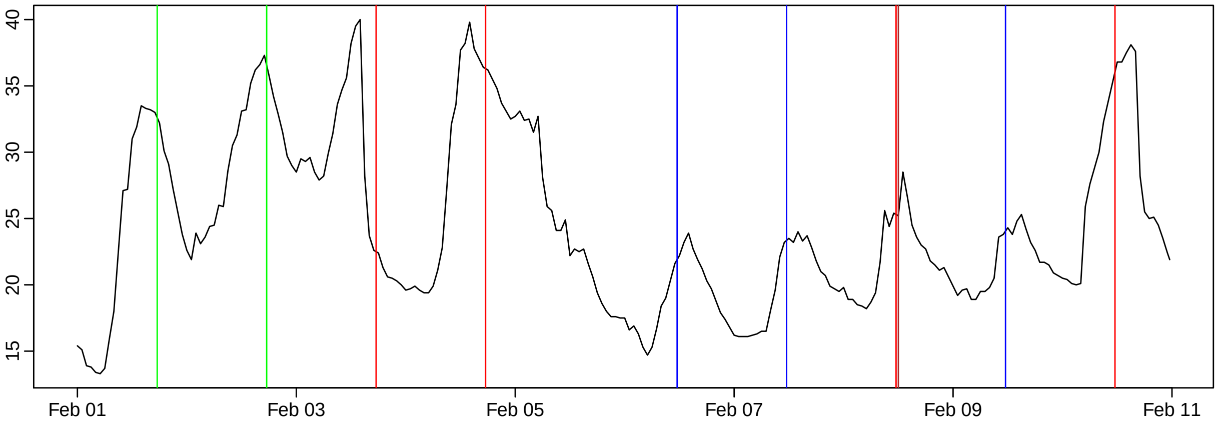


Figure 10: *The temperature in Melbourne starting from 1 February 2000 and during the following ten days. Four red lines and one brown line mark the times of the outliers. The green lines mark 5:30 pm (corresponding to the first two outliers) and the three blue lines mark 11:30 am.*

A positive outlier occurred at 5:30 pm on Thursday 3 February 2000, probably because it was end of one of the hottest days in Melbourne (40°C) followed by a sudden drop in temperature (see the first red line on Figure 10; the two previous green lines mark the same time on the previous two days). We can speculate that although the temperature dropped at 5:30 pm, the buildings stayed hot because they had heated up during the previous three very hot days and two nights. Therefore the electricity consumption was higher than expected by the model.

A negative outlier at 5:30 pm on Friday 4 February 2000 happened near the end of a very hot day in Melbourne (39.8°C) which was then followed by cooler weather (see the second red line on Figure 10). We can speculate that although the temperature did not drop (which made the model predict that electricity consumption will be high) the buildings did not require much cooling since the previous night was rather cold and probably also many people went out of building since it was a Friday night.

A positive outlier at 11:30 am on Thursday 10 February 2000 probably happened because of an early and extreme temperature rise in the beginning of the day. It might have led to a "shock" effect which was followed by a greater use of air conditioners and higher electricity consumption than anticipated by the model.

The outliers at 11:30 am and 12:00 pm (the first one which is positive, immediately followed by a negative one) are more difficult to explain by extreme temperature changes alone, although some sudden temperature jumps are observed on that day (8 February 2000). We could speculate that some other weather or economic events happened that day, although currently we do not have a satisfactory explanation of these two outliers.

11 Concluding remarks and discussion

In this article we introduced a new approach for seasonal-trend decomposition and provided a few examples, including rather complex, for decomposition using this approach. We showed that the new method allows for multiple and complex seasonality, provides confidence intervals, finds smoothing parameters, and allows regressors to be taken into account with coefficients that are possibly time-varying and seasonal. We have also proposed a robust version of the method.

The main disadvantage of our proposed method is its slow speed in cases when many seasonal components or seasonal predictors are used. Although, as was mentioned in Section 6, this problem can be mostly overcome with the assumption that the seasonal components and coefficients for flexible predictors do not change quickly, so that rather sparse knots can be used to achieve good performance without compromising the quality of decomposition.

In Section 9 we demonstrated that the RSTR variation of our approach deals rather well with outliers and shifts in the trend. In spite of the presence of a few very big residuals and some sudden changes in trend direction, the overall separation of the data into trend, seasonal and random components remained good. Here we would like to note that such behaviour is attributable to some good properties of the L_1 norm (and the double exponential distribution associated with it).

In Section 6 we showed that it is relatively easy to take smooth seasonality into account using terms that involve second differences. For example, in (11), we use the second term $\|\lambda_{tt} D_{tt} s\|_{L_2}^2$ to calculate and restrict second derivatives along the time dimension. Interestingly, cyclicity (i.e., smooth aperiodic fluctuations) can also be taken into account using a similar technique. In this case, and using (11), the term responsible for cyclicity will be $\left\|\lambda_{tt}^{(cyc)} \left(\alpha^2 D_{tt} + I\right) \ell_{cyc}\right\|_{L_2}^2$, where D_{tt} is the matrix that takes second differences along the time dimension using nearest observations and α is the coefficient proportional to an average cycle length. We are going to investigate this idea further in future research.

De Livera et al. (2011) show that multiple seasonality can be identified in various ways. In this work we do not consider the identifiability problem, assuming that any seasonal representation which minimises cross validated error will suit.

The L_1 and L_2 norms used in minimization problems can be mixed according to different assumptions on the distributions of the residuals of the model, and assumptions about trend or seasonal component changes. In such cases, the minimization problem can be reduced to the LASSO minimization problem.

A mixture of norms can be useful, for example, in cases when trend and seasonal patterns are smooth and also the noise component has outliers. It can also be useful when the trend changes abruptly, but the seasonal components are smooth, and the noise is distributed normally or, at least, has no outliers. In all such cases the mixture of norms can lead to better performance of the model.

Finally we need to note that our new approach allows us to deal with shifts (rapid changes in the level of the data) separately, treating them as another component along with trend, seasonal, random and cyclic parts. The main difference between, for example, the trend component and this new shift component is how they are regularised in the corresponding minimization problems: while the second differences are regularised for the trend component, for the shift component the first differences are regularised instead (in both cases using L_1 norm). We plan to investigate this approach further in our subsequent research.

References

Anderson, V. O. and Nochmals, U. (1914). The elimination of spurious correlation due to position in time or space. *Biometrika*, pages 269–279.

Bell, W. R. and Martin, D. E. (2004). Modeling time-varying trading-day effects in monthly time series. In *Proceedings of the Joint Statistical Meetings*, pages 8–12.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society. Series A (General)*, pages 321–337.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73.

Commandeur, J. J. F., Koopman, S. J., and Ooms, M. (2011). Statistical software for state space methods. *Journal of Statistical Software*, 41(1):1–18. http://www.jstatsoft.org/v41/i01.

Copeland, M. T. (1915). Statistical indices of business conditions. *The Quarterly Journal of Economics*, pages 522–562.

Dagum, E. B. (1988). *The X11ARIMA/88 Seasonal Adjustment Method: Foundations and User's Manual*. Statistics Canada, Time Series Research and Analysis Division.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527.

Dokumentov, A. and Hyndman, R. J. (2013). Two-dimensional smoothing of mortality rates. Technical report, Monash University. http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2013/wp26-13.pdf.

Dokumentov, A. and Hyndman, R. J. (2014). Low-dimensional decomposition, smoothing and forecasting of sparse functional data. Technical report, Monash University. http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2014/wp16-14.pdf.

Findley, D. and Soukup, R. (2000). Modeling and model selection for moving holidays. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association, Alexandria: American Statistical Association Alexandria*, pages 102–107.

Findley, D. F. (2005). Some recent developments and directions in seasonal adjustment. *Journal of Official Statistics*, 21(2):343.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2):127–152.

Findley, D. F., Monsell, B. C., and Hou, C.-T. (2009). Stock series holiday regressors generated by flow series holiday regressors. *Statistics*, (04).

Harvey, A. and Peters, S. (1990). Estimation procedures for structural time series models. *Journal of Forecasting*, 9(2):89–108.

Harvey, A. C. (1985). Trends and cycles in macroeconomic time series. *Journal of Business & Economic Statistics*, 3(3):216–227.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Hood, C. C. H. and Feldpausch, R. (2011). Experiences with user-defined regressors for seasonal adjustment. Technical report, Catherine Hood Consulting. http://www.catherinechhood.net/papers/chood_asa2011.pdf.

Hooker, R. H. (1901). The suspension of the berlin produce exchange and its effect upon corn prices. *Journal of the Royal Statistical Society*, 64(4):574–613.

Hyndman, R. J. and Fan, S. (2010). Density forecasting for long-term peak electricity demand. *Power Systems, IEEE Transactions on*, 25(2):1142–1153.

Koenker, R. (2013). *quantreg: Quantile Regression*. http://cran.r-project.org/package=quantreg.

Ladiray, D. and Quenneville, B. (2001). Seasonal adjustment with the X-11 method. *Lecture notes in statistics*, 158:220.

Livera, A. M. D., Hyndman, R. J., and Snyder, R. D. (2010). Forecasting time series with complex seasonal patterns using exponential smoothing. (October).

Macauley, F. R. (1930). The smoothing of time series. *National Bureau of Economic Research*, pages 121–136.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

Makridakis, S. G., Wheelwright, S. C., and Hyndman, R. J. (1998). *Forecasting: methods and applications*. John Wiley & Sons, New York, 3rd edition. http://robhyndman.com/forecasting/.

McElroy, T. S. (2010). A nonlinear algorithm for seasonal adjustment in multiplicative component decompositions. *Studies in Nonlinear Dynamics & Econometrics*, 14(4).

Monsell, B. C. and Aston, J. A. (2003). Toward x-13? Technical report, U. S. Census Bureau. http://www.census.gov/ts/papers/jsm2003bcm.pdf.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Seber, G. A. F. and Lee, A. J. (2003). *Linear regression analysis*. John Wiley & Sons, second edition.

Shishkin, J., Young, A. H., and C, M. J. (1967). The x-11 variant of the census ii method seasonal adjustment program. *Bureau of the Census, Technical Paper*, (15).

Shiskin, J. (1957). Electronic computers and business indicators. In *Electronic Computers and Business Indicators*. National Bureau of Economic Research.

Spencer, J. (1904). On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893-97. *Journal of the Institute of Actuaries*, 38(4):334–343.

Weinberg, J., Brown, L. D., and Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198.

Chapter 5

Conclusion

5.1 Summary of the main ideas and contributions

The main contributions of this thesis are contained in the three articles, which are yet to be published. As the articles reflect my own development during my PhD studies, they take slightly different approaches. On the other hand, they have a few features in common, the most important being the method of regularisation. I have shown in this thesis that complexity reduction through regularisation is a useful working technique to achieve good forecasts and decompositions.

In particular, in Chapter 2 some bivariate smoothing methods are introduced to handle smooth or abrupt bivariate surfaces with occasional ridges. The technique was applied to smoothing logarithms of mortality rates. The logarithms of mortality rates are smooth data except for early ages, where they change abruptly with occasional ridges which can be attributed to cohort and period effects. The proposed methods were compared with each other and with some existing techniques. The new SMILES method outperformed all other tested methods.

The main contribution of the first article (Chapter 2), apart from the new method SMILE for decomposition of bivariate data, is the understanding that the distribution of features plays an equally important role as the distribution of errors. This fact, and appreciating that the normal distribution does not fit well the empirical distribution of changes in the data, made possible the introduction of the L_1 norm which corresponds to the double exponential distribution and resulted in this new method.

Another contribution of the first article is the procedure of decomposition which was reduced to a quantile regression. Such a reduction made the process of decomposition rather simple since the problem was reduced to a well-known method, already implemented in an R package.

The main contribution of the second article (Chapter 3) is the new method ROPES which allows for PCA style decomposition for two-dimensional sparse data. The method also allows for smoothing and forecasting two-dimensional sparse data. Using this new approach, a practical method of forecasting mortality rates was proposed, as well as a new method for interpolating and extrapolating sparse longitudinal data.

The main contribution of the third article (Chapter 4) is the new generic methods STR and RSTR for decomposing seasonal data. The new methods allow for multiple seasonal and cyclic components, multiple linear regressors with constant, flexible, seasonal and cyclic influence. Seasonal patterns (for both seasonal components and seasonal regressors) can be fractional and flexible over time, moreover they can be either strictly periodic or have a more complex topology.

Beyond the STR and RSTR methods, the main contribution of the third article (Chapter 4) is the presentation of the seasonal effects as "sparse" observations on surfaces, which are allowed to be of some complex topologies. This can be considered as a continuation of ideas discussed in Marron and Alonso (2014).

5.1.1 Common themes

A very important common theme in all chapters is the idea of "sparsity". This allows an understanding of how an incomplete set of observed data can provide enough information to predict events which are beyond direct observation. I see traces of this idea in many methods starting from Exponential Smoothing methods (see for example Hyndman et al., 2008) where the internal unobserved states are estimated, to such tangled implementations as Neural Networks where hidden layers can also represent "indirect" and "unobserved" knowledge, extracted from the observed data (see for example Mao and Jain, 1995). This idea was particularly important in Chapter 3 where the ROPES method was introduced.

Another common theme is the idea of presenting data as a decomposition with a set of components that correspond in some sense to different "sources". One such method uses Principal Component Analysis (PCA) (Jolliffe, 2002). I extended this idea in three proposed methods (SMILE, ROPES and STR).

Another common idea is the presentation of forecasting and decomposition methods as minimisation problems. The three new methods (SMILE, ROPES and STR, discussed in Chapters 2, 3 and 4) have been introduced as minimisation procedures. Moreover, if the smoothing parameters are fixed, such minimisation procedures appear to be convex.

5.2 Concluding remarks and further research plans

All three articles use cross validation to estimate smoothing parameters. I found this technique very useful, although it can be a relatively slow procedure. Further research may consider replacement of cross validation with a simpler procedure without compromising good estimation of the smoothing parameters. For example we may consider stepwise re-estimation of smoothing parameters. With the current set of smoothing parameters, the "residuals" for problems (1.3.4) or (1.4.1) are calculated. The "residuals" in this case are vectors, used to compute norms of the first and the second terms in equations (1.3.4) or (1.4.1). Assuming that the regularisation technique minimises the overall complexity of data, the "variance" of the residuals in the first and the second terms should depend on the optimal set of the smoothing parameters. This might allow re-estimation of the current set of the smoothing parameters to improve the "residuals". I anticipate that the procedure will converge to a reasonable set of smoothing parameters.

The first article (Chapter 2) proposes a new technique that deals with bivariate data. I plan to extend the approach to three or more dimensions. This can lead to non-trivial computational problems. More research is required to transform this idea into a working technique.

The ROPES method (Chapter 3) is also currently developed to work with two-dimensional data. It can be extended to multi-dimensional data (three or more dimensions), although it is unclear if the corresponding minimisation problems will remain convex. The use of tensors, eigenvectors of tensors and the extension of the star norm to tensors can help to prove convexity. This multidimensional approach will require further investigation, as well as a solution of computational problems that will arise due to higher dimensionality.

Another way to transform the ROPES method is to use a different norm for regularisation. For example, the L_1 norm instead of the L_2 norm can be used. I suspect that this will make the ROPES method more robust and change its decomposition properties significantly. However, a consequence of such a norm change is that the problem will no longer be convex. A computationally efficient algorithm will need to be developed to solve such a problem.

The third article (Chapter 4) proposes a new method STR, which, in my opinion, can be a useful decomposition technique in many applications in economics and finance. I am planning to create an R package to make this technique available to researchers and data scientists. I am also planning to publish an article about this package (in addition to the article in Chapter 4).

The STR method allows an interesting application in the field of signal processing. To be precise, if STR regressors are $\sin(nx)$ and $\cos(nx)$ functions, the resulting method will perform a Fourier decomposition in a “flexible manner”. Such an approach can be more useful than a Fourier transform or Wavelet transform since the former does not allow for changes in intensity of the frequencies, and the latter, while allowing for frequency variability, makes it at discrete points of time. The STR approach to this problem will be investigated further with an aim to publish an additional article.

The above research directions are challenging, but, if successful, can bring a new wave of useful methods to life. I am hoping to continue my research in these directions.

References

Hyndman, R, AB Koehler, JK Ord, and RD Snyder (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Jolliffe, I (2002). *Principal component analysis*. Wiley Online Library.

Mao, J and AK Jain (1995). Artificial neural networks for feature extraction and multivariate data projection. *Neural Networks, IEEE Transactions on* **6**(2), 296–317.

Marron, JS and AM Alonso (2014). Overview of object oriented data analysis. *Biometrical Journal* **56**(5), 732–753.