Stringing High Dimensional Data for Functional Analysis

Kun Chen, Kehui Chen, Hans-Georg Müller and Jane-Ling Wang
Department of Statistics, University of California, Davis
Davis, CA 95616 USA

November 2010

ABSTRACT

We propose Stringing, a class of methods where one views high dimensional observations as functional data. Stringing takes advantage of the high dimension by representing such data as discretized and noisy observations that originate from a hidden smooth stochastic process. Assuming that the observations result from scrambling the original ordering of the observations of the process, Stringing reorders the components of the high-dimensional vectors, then transforming the high-dimensional vector observations into functional data. Established techniques from Functional Data Analysis can be applied for further statistical analysis once an underlying stochastic process and the corresponding random trajectory for each subject have been identified. Stringing of high dimensional data is implemented with distance-based metric Multidimensional Scaling, mapping high-dimensional data to locations on a real interval, such that predictors that are close in a suitable sample metric also are located close to each other on the interval. We provide some theoretical support, showing that under certain assumptions, an underlying stochastic process can be constructed asymptotically, as the number of data p tends to infinity. Stringing is illustrated for the analysis of tree ring data and for the prediction of survival time from high-dimensional gene expression data and is shown to lead to new insights. In regression applications involving high-dimensional predictors, Stringing compares favorably with existing methods.

KEY WORDS: Functional Cox Model, Functional Data Analysis, Multidimensional Scaling, Regression with High-Dimensional Predictors, Tree Rings, Weak Convergence.

We are grateful to the Editor, an Associate Editor and two referees for constructive critiques which led to many improvements in the paper and to the Editor for pointing out connections with Andrew's plots. This research was supported by NSF grants DMS03-54448, DMS04-06430, DMS05-05537 and DMS08-06199.

1. INTRODUCTION

Modeling and prediction for very high-dimensional data is a challenging problem. For the so-called large n small p problem, dimension reduction is essential. Existing regression methods involving high-dimensional predictors operate under strong assumptions such as sparsity constraints, which often are justified by not much more than plausibility considerations and mathematical convenience. In this paper, we introduce *Stringing* to complement the prevailing concept of sparsity of high-dimensional data, harnessing methods from Functional Data Analysis by transforming very high dimensional data into functional data. Rather than viewing it as a nuisance, this approach takes advantage of the high dimensionality of the predictors. The components of the predictor vector are treated as order-perturbed. After applying Stringing to order these components, the high-dimensional data are mapped to realizations of a smooth stochastic process. Such a generating process does not actually need to exist physically; it suffices to view it as a merely theoretical construct in order to reap the benefits of Stringing.

Our actual assumptions are confined to correlation or neighborhood relationships among predictors, as is further explained in the discussion in Section 5. As we show, under such assumptions, a smooth stochastic process may be constructed from the data. An alternative description is that the predictors possess an order in which their values correspond to smooth functions, but that we observe the predictors in a randomly permuted order, where the permutation is unknown. Implementation of this general concept requires judicious construction of imputed positions for the predictors on the real line, derived from observed sample distances or other measures of proximity between the individual components of the high-dimensional data vectors. Stringing then maps high-dimensional predictor vectors to infinite-dimensional function space. Once this mapping has been constructed, one can take advantage of functional data analysis (FDA) methodology to effectively analyze the resulting infinite-dimensional smooth random functions.

Starting with a proximity measure for the components of the given high-dimensional data vectors, we use Multidimensional Scaling (MDS) to implement Stringing. MDS projects data into a low-dimensional target space, where the configuration in the target space aims to reproduce the proximity relations in the original space, by minimizing a cost function. In our implementation of Stringing we use Euclidean distance as well as transformed Pearson correlation as proximity measures in the original high-dimensional predictor space (Cox and Cox 2001). The configuration obtained by MDS projection into one dimension provides an ordering of the predictors and assigns a location to each predictor, aligning the predictors within a one-dimensional interval like pearls on a string. Predictors with high proximity will tend to be positioned closely together after MDS projection, enabling the construction of smooth trajectories in function space.

Once the data have been converted into a smooth stochastic process by Stringing, functional principal components, a main tool in FDA (Rice and Silverman 1991; Ramsay and Silverman 2005; Yao et al. 2005), can be used to summarize and further analyze the high-dimensional data. Stringing is also of interest to provide a graphical representation of high-dimensional data by transforming each high-dimensional data vector into a function. In a way, this extends the visualization of multivariate data by converting them to functions that was pioneered in Andrew's Plots (Andrews 1972; Embrechts and Herzberg 1991; Garcia-Osorio and Fyfe 2005) to the high-dimensional case.

We illustrate Stringing as a tool to create an ordering of observation years and subsequent functional data analysis for tree ring data, which play an important role in climate research, and in the context of microarray gene expression data, for a situation where gene expression levels are of interest as predictors for survival. Functional embedding, an algorithm that is related to Stringing, has been demonstrated in an applied setting for the classification of high-dimensional gene expression data in Wu and Müller (2010), and other previous approaches to the problem of predicting survival from high-dimensional

gene expression profiles include the work of Rosenwald et al. (2002); Nguyen and Rocke (2002); Li and Li (2004); Bair and Tibshirani (2004); Bair et al. (2006). We address this prediction problem by coupling Stringing with a novel functional Cox regression model.

The paper is organized as follows. In Section 2 we describe Stringing. Its practical performance is studied in Section 3 through Monte Carlo simulations. The application of Stringing to tree ring width data and the prediction of survival for lymphoma patients from high-dimensional gene expression data by combining Stringing with a functional Cox model is the topic of Section 4, followed by a discussion in Section 5. Theoretical justifications and proofs can be found in the online Supplemental Material.

2. STRINGING VIA MULTIDIMENSIONAL SCALING

Multidimensional Scaling (MDS) maps p objects to points $s_1, \ldots, s_p$, situated in a low-dimensional space $\mathbb{R}^m$, given distances (or proximities) D_{jk} between any pair of objects j and $k, 1 \leq j, k \leq p$. The configuration of the low-dimensional points is determined by minimizing a cost function, which measures how well a particular configuration in the low-dimensional space approximates the original distances. MDS can be categorized into different types, depending on whether the distance data is matched qualitatively (non-metric MDS), in which case only the order of the distances matters, or quantitatively (metric MDS), where distances are matched in terms of numerical values; further according to whether the distances in the target space are directly fitted (distance scaling), or are approximated by preserving the inter-point inner products (classical scaling). For further details, we refer to the discussion paper Ramsay (1982) and the textbook Cox and Cox (2001). In preliminary studies, we found that metric distance scaling with the Stress criterion as cost function, in the form of unidimensional scaling (UDS) with $m = 1$, aiming to map predictors to locations in a one-dimensional interval, works best for Stringing; accordingly, this version of MDS is used in our implementation.

In Stringing, the ensemble of predictor values is thought of as being generated by a hidden smooth stochastic process $\{Z(s),\, s \in [0,1]\}$, where each element of a grid of support points $s_j \in [0,1]$ indexes one possible predictor, s_j being the "position" of the corresponding predictor and $Z(s_j)$ its value. Stringing infers the unknown predictor positions from the data; the distance between predictor positions is interpreted as a measure of the relatedness of the predictors. A key feature is that predictor values recorded at nearby predictor positions are close in terms of a suitably selected sample distance measure.

Once the predictor locations have been imputed, the predictor values are viewed as values assumed by smooth functions at these locations. In this way, each high-dimensional predictor vector is converted into a random function, so that methodology from FDA may be applied. Our goal is thus to construct coordinates s_j for each of p predictors on the real line, resulting in an embedding of the high-dimensional data vectors into an infinite-dimensional space of smooth functions.

The available high-dimensional data consist of n independent data vectors, each containing measurements for p predictors, where typically $p \gg n$. In the data matrix $\mathbf{X} = (x_{ij})_{1\le i\le n, 1\le j\le p}$ each element x_{ij} represents the measurements for the jth predictor of the ith subject, and $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. In prediction problems, one also has associated responses Y_i. In our application to gene expression data, this is survival time, which may be censored. It is well known that one needs to implement dimension reduction for such very high-dimensional prediction problems. In some applications, such as tree ring data analysis, with data of somewhat lower dimension, the emphasis is on identifying trajectories of an underlying smooth stochastic process, which in this case has a physical interpretation through an association with historical climatic trends, while in many other applications, there will be no physical interpretation for this process, and its significance derives from the fact that each random function corresponds to one of the high-dimensional vectors. Throughout, we refer to the high-dimensional observed vectors

as vectors of predictors, irrespective of whether the data actually include scalar responses.

As sample distance measures between various predictors, we use empirical Euclidean distances $\hat{D}_{jk} = [\frac{1}{n}\sum_i (x_{ij} - x_{ik})^2]^{1/2}$ as well as distances derived from proximities such as empirical Pearson correlation, $\hat{D}_{jk} = (2 - 2\hat{\rho}_{jk})^2$, where $\hat{\rho}_{jk} = \frac{1}{n-1}\sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/\{\hat{\sigma}_j \hat{\sigma}_k\}$, with $\bar{x}_j = \frac{1}{n}\sum_i x_{ij}$, $\hat{\sigma}_j = [\frac{1}{n-1}\sum_i (x_{ij} - \bar{x}_j)^2]^{1/2}$. Once a distance matrix $\mathbf{D}$ has been determined, the predictors are stringed into the real line by minimizing the stress function $S_{\mathbf{D}}(s_1, \ldots, s_p) = \sum_{j<k} (d(s_j, s_k) - D_{jk})^2$, where $s_j \in \mathbb{R}$ is the coordinate of the projected location of the jth predictor in the one dimensional projection space, and the metric $d(\cdot, \cdot)$ is Euclidean distance. To implement UDS, we observe $d(s_j, s_k) = |s_j - s_k| = (s_j - s_k)\text{sign}(s_j - s_k)$ and adopt a classical method (Kruskal 1964) to minimize the stress function. It is well-known that in UDS there is a high chance that various minimization algorithms terminate in local minima, which however is of less concern when only the rank order of the proximities matters (Borg and Groenen 2005; Hubert and Arabie 1986; Hubert et al. 1997; Pliner 1996), as is the case in our application of UDS. This is confirmed by simulation results, where the UDS solutions (even where they might correspond to local minima) prove to have excellent properties under various settings.

After applying UDS, the resulting one-dimensional configuration $\mathbf{s}$, reflecting pairwise predictor distances, provides support points for constructing a trajectory of predictor levels for each subject and implies a natural order of the predictors. This order is characterized by a permutation ψ_p such that $\hat{s}_{\psi_p(1)} < \ldots < \hat{s}_{\psi_p(p)}$, from which we define the regularized position for the j-th predictor with rank order $\psi_p(j)$ as $\tilde{s}_{\psi_p(j)} = \frac{j-1}{p-1} = s_{jp}$. Here, the domain of the stringed data is normalized to $[0, 1]$. We refer to the permutation ψ_p, which defines Stringing for given data of dimension p, as the *Stringing function*.

We conclude this section by noting that there is some theoretical support for Stringing. Specifically, invoking assumptions regarding the performance of UDS and the proximity (correlation) structure of the predictors, one can show that Stringing recovers the under-

lying ordering of the predictors (Theorem 1 in the online Supplemental Material) and that the stringed predictor series converges weakly to an underlying smooth stochastic process (Theorem 2 in the online Supplemental Material).

3. SIMULATION STUDIES

The performance of Stringing under various settings is demonstrated with three simulation studies. The first two relatively small simulations are designed to explore the efficiency of the UDS projection for ordering scrambled predictors for situations where measurements are taken without or with additional measurement errors and for varying numbers of predictors p. The third, more comprehensive simulation compares the performance of Stringing with that of the lasso for various linear and generalized linear regression settings with high-dimensional predictors.

To assess the proficiency of Stringing for reordering, in the first two simulations we study whether an underlying true but unknown order of predictors can be recovered from data where this order has been randomly permuted, and for a situation mimicking a real tree ring data example, that will be described in the following section. Using a Karhunen-Loève representation (Ash and Gardner 1975) for zero mean processes Z,

$$Z(s) = \sum_{j=1}^{\infty} \xi_j \phi_j(s), \quad s \in [0,1], \tag{1}$$

these simulations are based on potentially noisy measurements, generated according to

$$x_{ij} = \sum_{k=1}^{K} \xi_{ik}\phi_k(t_j) + \sigma z_{ij},\ z_{ij} \text{ i.i.d. } \sim N(0,1),\ t_j = \frac{j-1}{p-1},\ j = 1,\ldots,p,\ i = 1,\ldots,n, \tag{2}$$

for base functions ϕ_k, where all ξ_{ik} are independent, $\xi_{ik} \sim N(0, \lambda_k)$. The number of included components K, noise error variance σ^2 and eigenvalues λ_k are as specified below.

To simulate the situation with unknown underlying time order of the observations, we randomly permute the recording times t_j for each simulation run and then apply Stringing, aiming to recover the true underlying order. The Stringing step is based on Euclidean

or correlation-based sample distances between pairs of predictors and on UDS, mapping the p predictors into a real interval by assigning a location within the interval to each predictor. The order of the predictors in this UDS configuration is then compared to their true order, as determined by the (original and hidden) t_j, $j = 1, \ldots, p$.

When comparing mean order errors, defined as suitable difference between the order obtained by Stringing and the true order, averaged over p predictors, one observes that the order obtained by Stringing might be the exact reverse of the true order, since first and last elements are not identifiable. As Stringing is invariant under complete order reversal, we select the order associated with the smaller value of $\sum_{j=1}^{p} |o_j^{\mathrm{S}} - o_j|$, where o_j is the true rank of predictor j and o_j^{S} is the rank of predictor j induced by the one-dimensional UDS configuration. For a random permutation of the data, the expected order error is $E_R = E\{\sum_{j=1}^{p} |o_j^{\mathrm{R}} - o_j|\}$, where o_j^{R} denotes the order of the j-th element under random permutation. A simple calculation shows that $E_R(p) = (p-1)(p+1)/3$, which serves as normalization factor to obtain the relative order error

$$ROE = \sum_{j=1}^{p} |o_j^{\mathrm{S}} - o_j| / E_R(p). \tag{3}$$

In a first simulation study (*Simulation 1*), we investigated the effect of various signal-to noise-ratios and studied the following choices in the Karhunen-Loève expansion (eq. (1)): number of components $K = 4$; generating functional principal components ξ_1, ξ_2, ξ_3, ξ_4 as independent normal random variables with mean zero and variances 4, 2, 1, 0.5, respectively; and four orthonormal basis functions ϕ_k from the Fourier base, $\phi_1(t) = -\sqrt{\frac{1}{5}} cos(\frac{1}{5}\pi t)$, $\phi_2(t) = \sqrt{\frac{1}{5}} \sin(\frac{1}{5}\pi t)$, $\phi_3(t) = -\sqrt{\frac{1}{5}} \cos(\frac{2}{5}\pi t)$, $\phi_4(t) = \sqrt{\frac{1}{5}} \sin(\frac{2}{5}\pi t)$; sample size $n = 50$; and predictor dimension $p = 50$. The distance matrix was obtained as one minus the pairwise correlation and 400 simulation runs were generated for various values of the signal-to-noise ratio (SNR), defined as $mean(|x|)/\sigma$. The resulting relative order errors (ROE) in Table 1 indicate that ROE, not surprisingly, increases with decreasing

SNR, but Stringing still leads to substantial gains for situations with noisy data under low SNR levels. A typical result of Stringing for one simulation with $\sigma = 0$ is shown in Figure 1, indicating perfect order identification for this case.

A second simulation study (*Simulation 2*) was designed to evaluate the behavior of Stringing when sample size n and predictor dimension p are changing. We simulate data that closely resemble the tree ring data, described in the data analysis section (Section 4). Accordingly, we use the three estimated eigenfunctions as base functions ϕ_k, $k = 1, 2, 3$, and their associated eigenvalues λ_k, as determined in the application, to generate simulated data. The error variance was chosen as $\sigma^2 = 0.004^2$, with associated signal-to-noise ratio of $SNR = 5.5$, similar to that observed for the tree ring data. The UDS step was based on L^2 distance. Results for one sample run with sample size $n = 50$ and predictor dimension $p = 50$ are shown in Figure 2. To assess how the behavior of Stringing depends on sample size n and predictor dimension p, the results of 400 simulations for different combinations of p and n, for $\sigma = 0$, are reported in Table 2. They indicate that relative order errors ROE decrease as p and n increase, respectively, as predicted by theory (see online Supplemental Material). Overall, Stringing is seen to identify the true ordering under which the data were generated with very good precision.

To explore the performance of in general regression settings and to compare it with Lasso as an established method for regression with high-dimensional predictors, a comprehensive third simulation study (*Simulation 3*) was conducted. We studied linear and generalized regression settings, $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, 1)$, and $Y \sim Bernoulli(\mu)$, where $\mu = \exp(X\beta)/(1+\exp(X\beta))$, $X \sim N_p(0, \Sigma)$, with the following predictor covariance structures chosen for $\Sigma = \text{cov}(X_i, X_j) = \sigma(i, j)$, $i, j = 1, \ldots, p$: (i) $\sigma(i, j) = 0.5^{\sqrt{|i-j|}}$, with average correlation of 0.07, (ii) $\sigma(i, j) = 0.9^{\sqrt{|i-j|}}$, with average correlation of 0.58, (iii) $\sigma(i, j) = \min(i, j)$, and (iv) $\sigma(i, j) = \sigma(j, i) = \mathcal{U}(0, 1)$, chosen as i.i.d. uniform random numbers, followed by projecting the resulting symmetric matrices onto the space of

non-negative definite matrices via spectral decomposition, as in Hall et al. (2008).

For all simulations, the order of the predictors X_j, $j = 1, \ldots, p$, was scrambled randomly and X was multiplied by a constant to stabilize $\mathrm{var}(X\beta) = 6$. Regression coefficients β_j were generated as $\beta_j \sim \mathcal{U}(0, 1)$, $j = 1, \ldots, p$, with those falling below the αth quantile set to zero, for $\alpha = 0,\ 0.5,\ 0.8,\ 0.9$, i.e., the fraction of nonzero coefficients was controlled at 100%, 50%, 20% and 10%. For all settings, $\sum_{j=1}^{p} \beta_j$ was normalized to be 100, and the order of coefficients β_j was scrambled along with X_j. To investigate the effect of varying p/n ratios, combinations (i) $p = 100,\ n = 30$, (ii) $p = 100,\ n = 60$, and (iii) $p = 50,\ n = 60$ were studied, using a test sample of size 50 for all cases.

Stringing was implemented with UDS to generate a relatively smooth order, followed by applying functional linear regression or functional generalized linear modeling (FGLM), reviewed in Müller (2005), to predict the response. All auxiliary parameters involved were chosen by leave-one-out cross validation. Comparison methods included Lasso (Matlab version posted on the Lasso homepage, with parameters chosen by leave-one-out cross validation) and regular least squares linear fitting and likelihood-based GLM. Lasso and Stringing performed much better than least squares and GLM, and only the comparison between Stringing and Lasso is reported. Results on relative mean squared errors or relative misclassification rates (based on 200 simulation runs) for all settings are reported in Table 3. Boxplots of MSE over 200 simulation runs for Stringing and Lasso can be found in Figure 3 for the continuous response case with covariance structure $\mathcal{U}(0, 1)$. Boxplots for other settings show similar patterns and are shown in the online Supplemental Material.

For continuous responses Y, the table and the figure indicate the expected relative deterioration in the performance of Stringing as the regression parameters β become sparser. What is remarkable, however, is that for small n, large p, such as $p = 100$, $n = 30$, Stringing dominates Lasso even in very sparse cases with low predictor correlation. When $p = 100$, $n = 60$, and $p = 50$, $n = 60$, Stringing still outperforms Lasso in most cases

as long as predictors are not too sparse. For the binary response Y, Stringing performs even better relative to Lasso, which it outperforms for nearly all scenarios. We note that these results are obtained for situations where predictors and regression coefficients are generated in standard Lasso simulation settings, with no prior smoothness, and in addition the ordering of all data was scrambled. The results demonstrate that Stringing is competitive for handling high-dimensional predictors. It works better than Lasso for less sparse and large p situations, even when no physically interpretable smooth underlying process exists. Capitalizing on relatively smooth covariance structures of stringed data, it then proves beneficial to apply functional methodology.

4. DATA ILLUSTRATIONS

4.1 Stringing of Tree Ring Widths

Tree ring data were obtained for 45 blue oak trees located at Mary Ranch, Santa Clara, California, from the International Tree-Ring Data Bank, IGBP PAGES/World Data Center for Paleoclimatology, NOAA/NCDC Paleoclimatology Program (at Boulder, Colorado, USA, file name CA645, contributed by D.W. Stahle and R.D. Griffin, June 29, 2009). They consist of annual tree ring width measurements, obtained for each of the years 1932-1976 for the trees in the sample. Since we are mainly interested in the variation of tree ring widths across years and not across trees, differences in overall growth rates between trees were removed by dividing the tree ring widths by the total width gained from year 1932 to year 1976, for each tree separately.

It is well known that tree ring widths are influenced by climatic factors, and more specifically that limitations to tree growth are due to lack of precipitation, unfavorable soil properties or low temperatures (LaMarche Jr 1978; Cook and Kairiukstis 1990; Oberhuber and Kofler 2000; Bunn et al. 2005). A major factor limiting the growth of trees in warm dry climates is low soil water content, while temperature differences matter less. This

climatic setting applies at Mary Ranch in Santa Clara, and it is then likely that tree ring widths are associated with annual precipitation levels. In addition to random variation, weather cycles such as the multi-year El Niño–Southern Oscillation climate cycles in the Pacific ocean give rise to annually varying precipitation patterns, acting in addition to longer term trends, which may be related to climate change.

When applying Stringing to high-dimensional data, we are aiming to uncover an underlying smooth stochastic process generating the data. Specifically, for the tree ring data, we expect dependence of annual tree ring width on precipitation levels. The permutation defining the Stringing function tends to position years with similar growth and thus precipitation levels close to each other, permitting insights into changes in precipitation during the observation period. We base our analysis on the 45×45 distance matrix of pairwise Euclidian distances of tree ring widths, calculated for each pair of years within the time domain, and then apply Stringing. The tree ring series in the original order by year and in the Stringing induced order are shown in Figure 4. The stringed series clearly shows smoother trends of increasing tree growth, particularly towards the right side, overlaid by noise, while the sharp peaks in the original series are removed.

To explore the association between precipitation and tree ring widths, we obtained data on Monthly Average Temperature (in degrees F) and Monthly Total Precipitation (in inches) for Santa Clara county for the years 1932-1976 from the Western Region Climate Center. Of interest is the connection of the Stringing function (which defines the permutation of the years that uncovers the smooth underlying process) and rainy season precipitation (defined as precipitation from the previous December to April of the current year). Overlaying precipitation and Stringing functions, as in Fig. 5, indeed indicates parallel patterns, supporting the idea that the Stringing function here is directly related to annual precipitation. The 0.99 quantile of the sample of simple Pearson correlations, calculated between the precipitation function and 3000 random permutations

of the Stringing function is found to be 0.33, much below the actual Pearson correlation of the two curves in Fig. 5, which is 0.58, indicating that a significant association exists. This is a scenario where the smooth underlying process that is recovered with Stringing has a physical interpretation, relating growth to precipitation.

To represent tree ring width functions for individual trees, we apply functional principal component analysis for the tree data series with stringed time, using standard methods (Rice and Silverman 1991), implementing a version described in Yao et al. (2005). We aim to estimate the components of the Karhunen-Loève representation (1), i.e., the eigenfunctions ϕ_k of underlying processes Z and the principal components $\xi_{ik} = \int Z_i(t)\phi_k(t)\,dt,\ i = 1,\ldots,n,\ k = 1,2,\ldots.$ Eigenfunction estimates $\hat{\phi}_k$ can be based on smooth covariance surface estimation and estimates $\hat{\xi}_{ik}$ of functional principal components on numerical integration or conditional expectation (Yao et al. 2005; Müller 2005). The first three estimated eigenfunctions are shown in Fig. 6 and the resulting fits for 9 randomly selected trees in Fig. 7, demonstrating that Stringing induces sufficient smoothness to produce reasonably good fits. These fits can alternatively be presented in the original order. The functional principal components obtained by Stringing can then be used to summarize growth patterns of individual trees and for further applications, such as functional regression or clustering for tree ring width series.

4.2 Stringing and Functional Cox Regression for Predicting Survival from High-Dimensional Gene Expression Arrays

In a study of the survival of patients with diffuse large-B-cell lymphoma (DLBCL), Rosenwald et al. (2002) aimed to predict survival from individual high-dimensional microarray gene expression data. DLBCL is the most common type of lymphoma in adults with a cure rate of only 35 to 40 percent. The survival response to treatment varies largely, even for DLBCL patients with similar clinical features, and is thought to be influenced by

genetic differences between subjects. This motivates the study of gene-expression profiles as molecular predictors for survival. The data consist of $n = 240$ patients, for each of whom $p = 7399$ gene expression levels were measured. These measurements form an expression matrix $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, where x_{ij} is the gene expression level of the jth gene for the ith patient. The patients are randomly divided into training (160 subjects) and test (80 subjects) groups; only the training group data is used for model fitting. For the i-th subject, the survival response may be right-censored and accordingly is observed as (Y_i, δ_i), $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$, with survival time T_i, censoring time C_i and censoring indicator $\delta_i = 1_{\{T_i \leq C_i\}}$, where as usual censoring times C_i and survival times T_i are assumed to be independent, given the covariate values.

In addition to Rosenwald et al. (2002), the problem of predicting survival from high-dimensional gene expression data has been addressed by various authors, including Bair and Tibshirani (2004) and Bair et al. (2006), who propose supervised principal component analysis, where principal component analysis is performed using only a subset of those genes which have the strongest correlations with survival time. Reviews of these approaches can be found in Bøvelstad et al. (2007) and Witten and Tibshirani (2010).

For preprocessing the data, we follow the same approach as the above-cited authors and select the genes whose individual Cox scores obtained by fitting univariate Cox regression models satisfy $|z_i| > \vartheta$ for a threshold $\vartheta > 0$, the choice of which is discussed below. Stringing then uncovers an underlying latent smooth stochastic process Z. To model the influence of Z on survival time, we propose a functional Cox regression model, which differs from the well-known Cox model with time-varying covariates in the way covariate information relates to the risk at a particular time. The proposed functional Cox proportional hazards model for the conditional hazard rate $h(t|Z_i)$ is

$$h(t|Z_i) = h_0(t) \exp\left[\int (Z_i(s) - \mu(s))\beta(s)\, ds\right], \tag{4}$$

with baseline hazard function $h_0(t)$. In model (4), the entire covariate trajectory relates to

the hazard function through the coefficient function β. Since the eigenfunctions $\phi_1, \phi_2, \ldots$ of Z (1) form a basis, representing the coefficient function as $\beta(s) = \sum_{l=1}^{\infty} \beta_l \phi_l(s)$ and approximating $\beta(\cdot)$ and $Z_i(\cdot)$ by a finite number of L basis functions, (4) becomes $h(t|Z_i) = h_0(t) \exp(\sum_{l=1}^{L} \xi_{il}\beta_l)$, which has the form of a regular Cox regression model with predictors ξ_{il}, the functional principal components of Z.

The coefficient vector $\beta = (\beta_1, \ldots, \beta_L)'$ is estimated by maximizing the log partial likelihood

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta' \hat{\xi}_i - \log \left(\sum_{j \in R_i} \exp(\beta' \hat{\xi}_j) \right) \right\}, \tag{5}$$

where R_i is the index set of patients at risk at time T_i^- and $\hat{\xi}_i$ is the vector of the first L estimated functional principal components of X_i, obtained as described in Yao et al. (2005). The estimated coefficient function $\hat{\beta}(s) = \sum_{l=1}^{L} \hat{\beta}_l \hat{\phi}_l(s), s \in [0, 1]$ is then obtained by simply plugging in estimates $\hat{\beta}_l$, $\hat{\phi}_l$.

In our application, the threshold ϑ is determined by K-fold cross validation, $CV(\vartheta) = \sum_{k=1}^{K} \{l(\hat{\beta}_{-k}(\vartheta)) - l_{-k}(\hat{\beta}_{-k}(\vartheta))\}$, which leads to $\vartheta = 3.5$ and the selection of 80 genes as input for Stringing and the functional Cox model. Figure 8 displays the coefficient function in model (4) for one random split into training and test data. To obtain 95% confidence intervals for the regression coefficients, we implemented 100 random splits; 21 predictor genes were then found to be significant at the 0.05 level. Evaluation of prediction methods can be based on the deviance criterion

$$DEV = -2\{l^{test}(\hat{\beta}) - l^{test}(0)\}, \tag{6}$$

i.e., partial likelihood (5) computed for test sets, averaging over 50 random splits into training and test sets, as advocated by Bøvelstad et al. (2007). Smaller deviances characterize a preferred method. Comparing with three previously used methods for the DLBCL data, principal component regression, ridge regression, and Lasso, quoting the deviances for these methods from Bøvelstad et al. (2007), Stringing was found to have substantially

smaller deviances (Table 4) and therefore is an attractive alternative.

5. DISCUSSION

Classical regression analysis often fails for high-dimensional predictors. Popular methods to address this problem, such as Lasso, rely on assumptions of sparseness and low correlatedness of predictors. Stringing is a complementary approach, which benefits from increasing dimensionality and increasing predictor correlation, taking advantage of these features by mapping the high-dimensional data into infinite-dimensional functional space. Stringing thus makes high-dimensional data amenable to FDA methodology, e.g., reducing the infinite dimension of the functional data to a manageable number of functional principal components or other basis coefficients and functional regression analysis.

Stringing provides a framework for creating locations and an order for initially unordered components of high-dimensional data. An intriguing possibility for future work is to consider mappings of high-dimensional data into m-dimensional spaces, where $m > 1$, rather than $m = 1$, as considered here. There are similarities with manifold learning (Tenenbaum et al. 2000; Bengio et al. 2006), as closest paths that connect nearby data play an important role in both approaches. Crucial differences are that Stringing attempts to string predictor components onto one of the paths and operates on distances of predictor components, rather than on distances and mappings of predictor vectors for different subjects.

Stringing is attractive for a variety of statistical analyses that pertain to high-dimensional data, including the prediction of a continuous or categorical response, survival, or the graphical or statistical representation of such data. The ordering of the predictors induced by the Stringing function is shown in simulations to work well, also in the presence of additional measurement errors in the data, and may reveal features of interest, as demonstrated for the tree ring data. In the survival regression situation, Stringing is competitive when compared with previous methods. As our extended

simulations show, for non-sparse or correlated predictors with continuous or categorical responses, Stringing performs better than Lasso, often by a wide margin. Stringing can be justified on theoretical grounds under certain regularity assumptions, which are different from and overall equally hard to verify as the currently prevailing sparseness assumptions (see online Supplemental Material).

We note that the postulated underlying smooth stochastic process that generates the data may have a physical interpretation as in the tree ring example, but in typical applications of Stringing to high-dimensional data such an interpretation neither exists nor is it needed. This concept leads to a useful graphical representation of high-dimensional data and provides a useful bridge from high-dimensional to functional data. More specifically, it is supported by two arguments: First, from the derivation of the Karhunen-Loève expansion (1), all that is needed is a smooth covariance surface after stringing the predictors (where smoothness may be taken with a grain of salt). That a smooth covariance can be attained hinges on the structure of the predictor proximities. Then, there exists a stochastic process with smooth trajectories and this particular smooth covariance, irrespective of whether predictors are physically derived from a smooth stochastic process with subsequent reshuffling of locations. Second, as the tree ring example shows, some data actually may be viewed as reshuffled observations of smooth trends, which in this case are climatic factors that vary randomly and non-smoothly from year to year, but overall are smoothly related to growth. This second motivation for a smooth process model is application-specific, while the first is generic and may be invoked for various high-dimensional data, analogously to sparseness for Lasso type methods.

We conclude by noting that key features of Stringing differ from practically all other available approaches, and especially from current multivariate modeling, where large-sample properties are mainly justified by considering increasing numbers n of subjects, while increasing predictor dimension p is a nuisance; in such settings, sparsity is essential.

In contrast, the justification of Stringing hinges on $p \to \infty$, and one takes advantage of non-sparse correlation patterns among predictors. These features define the promise of Stringing for high-dimensional data, for situations where predictors are globally correlated.

REFERENCES

Andrews, D. (1972), "Plots of high-dimensional data," *Biometrics*, 28, 125–136.

Ash, R. B., and Gardner, M. F. (1975), *Topics in Stochastic Processes*, Academic Press, New York.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components," *Journal of the American Statistical Association*, 101, 119–137.

Bair, E., and Tibshirani, R. (2004), "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data," *PLoS Biology*, 2, 511–522.

Bengio, Y., Monperrus, M., and Larochelle, H. (2006), "Nonlocal estimation of manifold structure," *Neural Computation*, 18, 2509–2528.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Borg, I., and Groenen, P. (2005), *Modern Multidimensional Scaling*, Springer, New York.

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. (2007), "Predicting survival from microarray data – a comparative study," *Bioinformatics*, 23, 2080–2087.

Bunn, A. G., Graumlich, L. J., and Urban, D. L. (2005), "Trends in twentieth-century tree growth at high elevations in the Sierra Nevada and White Mountains, USA," *The Holocene*, 15, 481.

Cook, E., and Kairiukstis, L., eds (1990), *Methods of Dendrochronology,* Kluwer Academic Publishers, Dordrecht.

Cox, T. F., and Cox, M. A. A. (2001), *Multidimensional Scaling,* Chapman & Hall Ltd.,

London.

Embrechts, P., and Herzberg, A. (1991), "Variations of Andrews' plots," *International Statistical Review*, 59, 175–194.

Garcia-Osorio, C., and Fyfe, C. (2005), "Visualization of high-dimensional data via orthogonal curves," *Journal of Universal Computer Science*, 11, 1806–1819.

Hall, P., and Hosseini-Nasab, M. (2006), "On properties of functional principal components analysis," *Journal of the Royal Statistical Society: Series B*, 68, 109–126.

Hall, P., Müller, H.-G., and Yao, F. (2008), "Modeling sparse generalized longitudinal observations with latent Gaussian processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 730–723.

Hubert, L., and Arabie, P. (1986), "Unidimensional scaling and combinatorial optimization," in *Multidimensional Data Analysis*, eds. J. De Leeuw, W. Heiser, and J. Meulman, DSWO Press, Leiden, pp. 181–196.

Hubert, L., Arabie, P., and Meulman, J. (1997), "Linear and circular unidimensional scaling for symmetric proximity matrices," *British Journal of Mathematical and Statistical Psychology*, 50, 253–284.

Kruskal, J. (1964), "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, 29, 115–129.

LaMarche Jr, V. C. (1978), "Tree-ring evidence of past climatic variability," *Nature*, 276, 334–338.

Li, L. X., and Li, H. Z. (2004), "Dimension reduction methods for microarrays with application to censored survival data," *Bioinformatics*, 20, 3406–3412.

Müller, H.-G. (2005), "Functional modelling and classification of longitudinal data," *Scandinavian Journal of Statistics*, 32, 223–240.

Nguyen, D., and Rocke, D. M. (2002), "Partial least squares proportional hazard regression for application to DNA Microarray data," *Bioinformatics*, 18, 120–127.

Oberhuber, W., and Kofler, W. (2000), "Topographic influences on radial growth of Scots pine (Pinus sylvestris L.) at small spatial scales.," *Plant Ecology*, 146, 231–240.

Pliner, V. (1996), "Metric unidimensional scaling and global optimization," *Journal of Classification*, 13, 3–18.

Ramsay, J. (1982), "Some statistical approaches to multidimensional scaling data," *Journal of the Royal Statistical Society: Series A*, 145, 285–312.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society: Series B* , 53, 233–243.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E., and Staudt, L. (2002), "The use of molecular profiling to predict survival after themotheropy for diffuse large-B-Cell lymphoma," *New England Journal of Medicine*, 346, 1937–1947.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000), "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290(5500), 2319–2323.

Witten, D., and Tibshirani, R. (2010), "Survival analysis with high-dimensional covariates," *Statistical Methods in Medical Research*, 19, 29–51.

Wu, P., and Müller, H.-G. (2010), "Functional embedding for the classification of gene expression profiles," *Bioinformatics*, 26, 509–517.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

SNR	∞	13.4	6.81	4.42	3.36	2.68
ROE	0	0.027	0.087	0.151	0.213	0.293

Table 1: Simulation results for reordering scrambled data with Stringing (*Simulation 1*): Relative order error (ROE), i.e., order error after stringing the scrambled data, relative to the expected order error of a random permutation, as defined in equation (3), for varying signal-to-noise ratios (SNR).

	$p = 50$	$p = 100$	$p = 200$
$n = 50$	0.0028	0.0015	0.0011
$n = 100$	0.0017	0.0007	0.0006
$n = 200$	0.0012	0.0005	0.0003

Table 2: Simulation results for reordering scrambled data with Stringing (*Simulation 2*): Relative order error (ROE), i.e., order error after stringing the scrambled data, relative to the expected order error of a random permutation, for various combinations of sample size n and predictor dimension p.

		Y Continuous				Y Binary			
	β / cov	100%	50%	20%	10%	100%	50%	20%	10%
$p =$	$0.5^{\sqrt{\lvert i-j \rvert}}$	0.312	0.421	0.680	0.681	0.630	0.656	0.746	0.814
100	$0.9^{\sqrt{\lvert i-j \rvert}}$	0.669	0.670	0.739	0.847	0.724	0.747	0.800	0.758
$n =$	$\min(i, j)$	0.746	0.778	0.820	0.871	0.826	0.769	0.814	0.800
30	$\mathcal{U}(0, 1)$	0.472	0.519	0.634	0.907	0.664	0.650	0.663	0.723
$p =$	$0.5^{\sqrt{\lvert i-j \rvert}}$	0.444	0.566	1.222	1.801	0.620	0.661	0.717	0.779
100	$0.9^{\sqrt{\lvert i-j \rvert}}$	0.749	0.761	0.801	1.719	0.833	0.848	0.904	0.893
$n =$	$\min(i, j)$	0.794	0.823	0.839	0.843	0.904	0.917	0.943	0.973
60	$\mathcal{U}(0, 1)$	0.603	0.653	0.759	1.219	0.707	0.734	0.770	0.779
$p =$	$0.5^{\sqrt{\lvert i-j \rvert}}$	0.593	0.725	1.510	1.573	0.656	0.752	0.779	1.031
50	$0.9^{\sqrt{\lvert i-j \rvert}}$	0.785	0.807	0.840	0.886	0.843	0.860	0.910	0.885
$n =$	$\min(i, j)$	0.830	0.821	0.832	0.850	0.862	0.900	0.902	0.907
60	$\mathcal{U}(0, 1)$	0.667	0.763	1.047	1.001	0.721	0.752	0.768	1.117

Table 3: Simulation results for comparisons between Stringing and Lasso (*Simulation 3*): Relative Mean Squared Errors (for continuous responses Y) and Relative Misclassification Rates (for binary responses Y) for Stringing relative to Lasso, for various covariance structures, sample sizes n, numbers of predictors p and sparseness levels (indicated as percentage of non-zero regression coefficients for each column of the table). Numbers less than one indicate scenarios where Stringing performs better.

methods / quartile	PCR	Ridge	Lasso	Stringing
1st quartile	1	-6	-1.5	-11.5
median	-3	-8.5	-4.5	-18
3nd quartile	-6.5	-11	-7	-21

Table 4: Comparison of quartiles of deviances (DEV, eq. 5) for survival prediction for four methods across test sets, including Stringing and three previously used methods (from the left, principal component regression (PCR); ridge regression (Ridge); Lasso), which are taken from Bøvelstad et al. (2007). Smaller deviance is better.

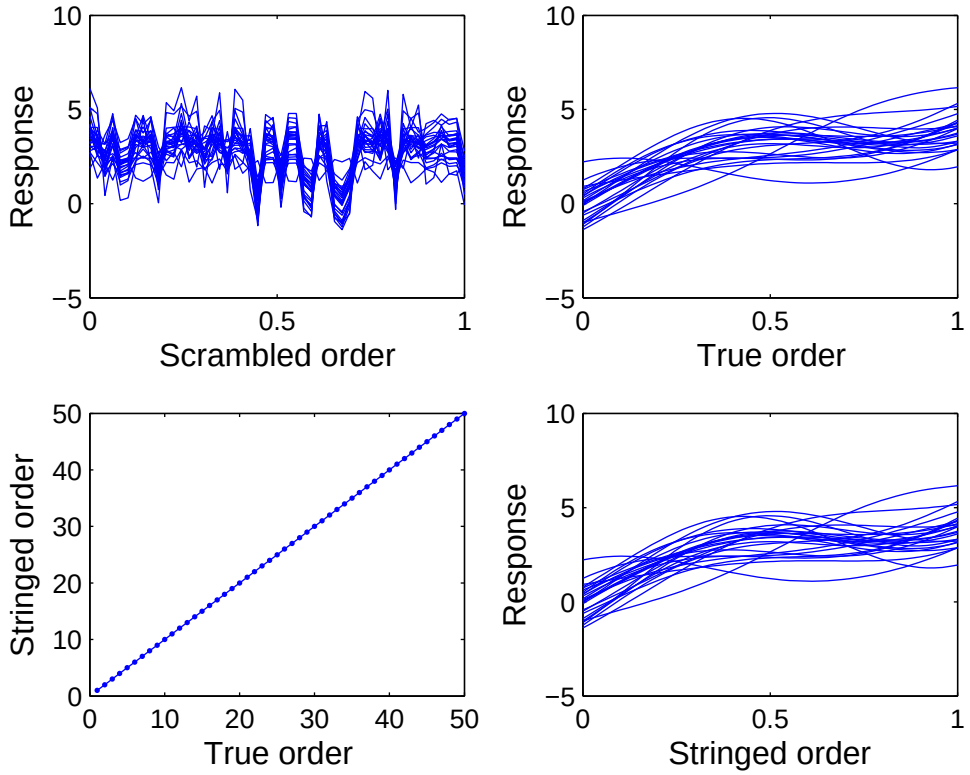

Figure 1: Data from *Simulation 1*. Top left: Responses in observed order (corresponding to a random permutation of the underlying true order). Top right: Responses in true order. Bottom left: Stringed order, obtained data-adaptively through the estimated Stringing function, plotted against true order. Bottom right: Responses from the upper left panel, reordered according to the estimated Stringing function.

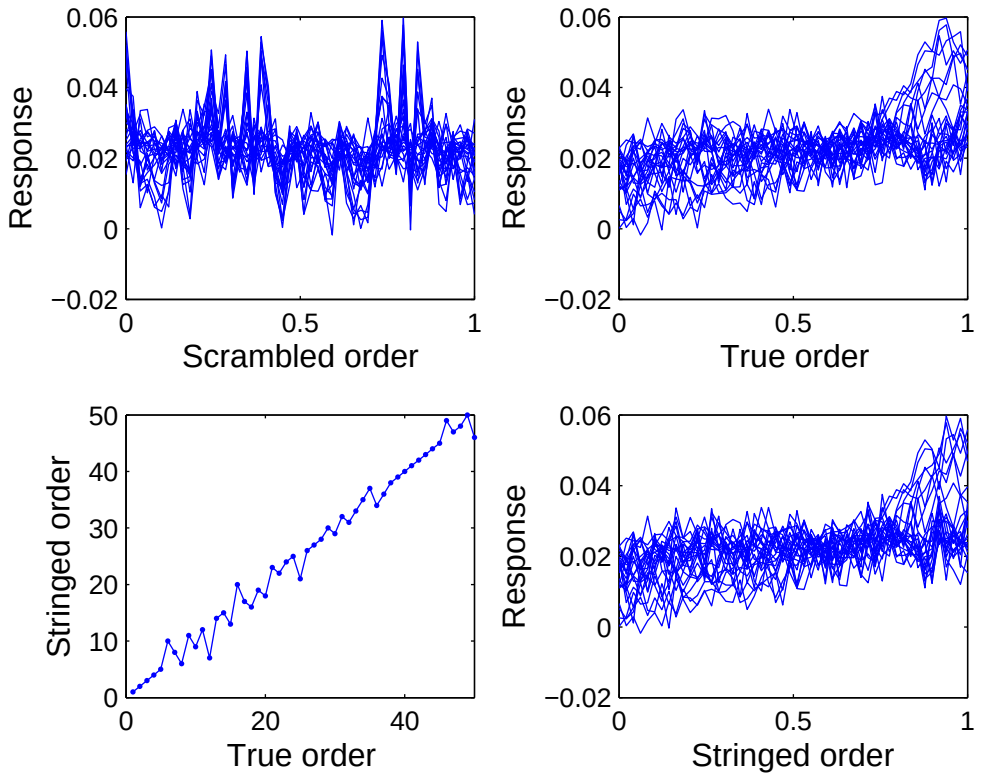

Figure 2: Data from *Simulation 2* with noise contamination. Top left: Observed responses with randomly permuted order. Top right: Responses in true order. Bottom left: Stringed order against true order. Bottom right: Responses ordered according to the Stringing function.

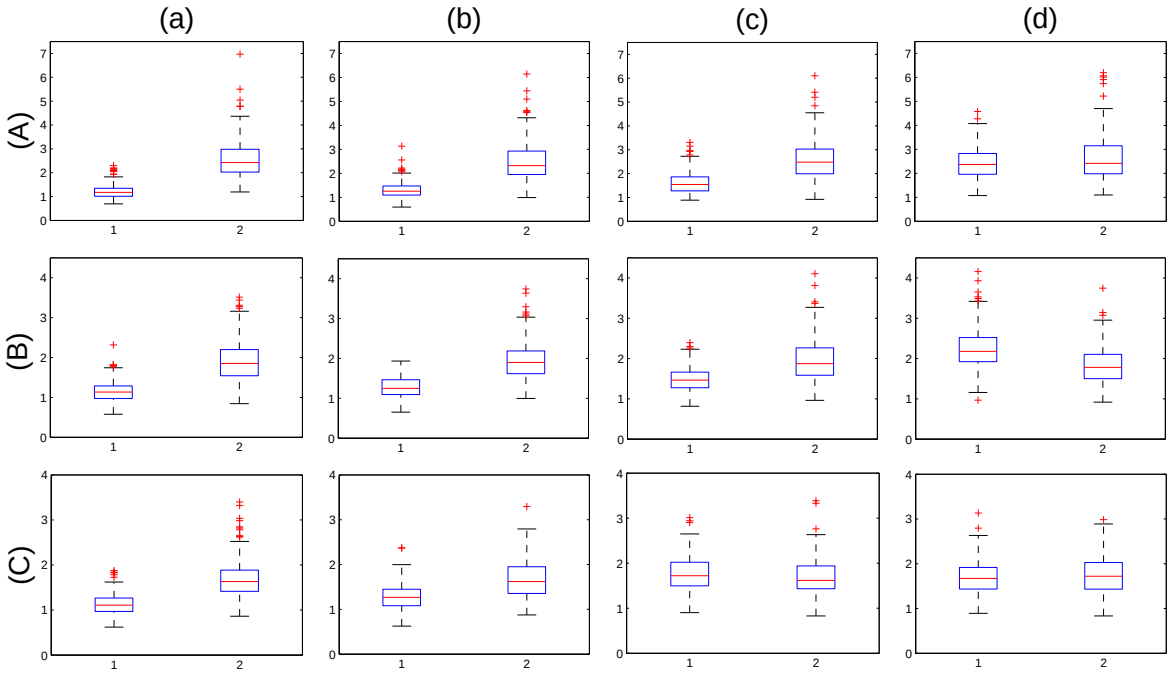

Figure 3: Boxplots of MSEs obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of n high-dimensional predictor p-vectors with continuous responses and predictor covariance structure $\mathcal{U}(0,1)$. Columns indicate level of sparsity (percentage of non-zero β): (a) 100%, (b) 50%, (c) 20%, (d) 10%. Rows indicate p/n ratio, (A) $p = 100,\ n = 30$, (B) $p = 100,\ n = 60$, (C) $p = 50,\ n = 60$. Within each panel, the left boxplot corresponds to Stringing (label 1), the right one to Lasso (label 2).

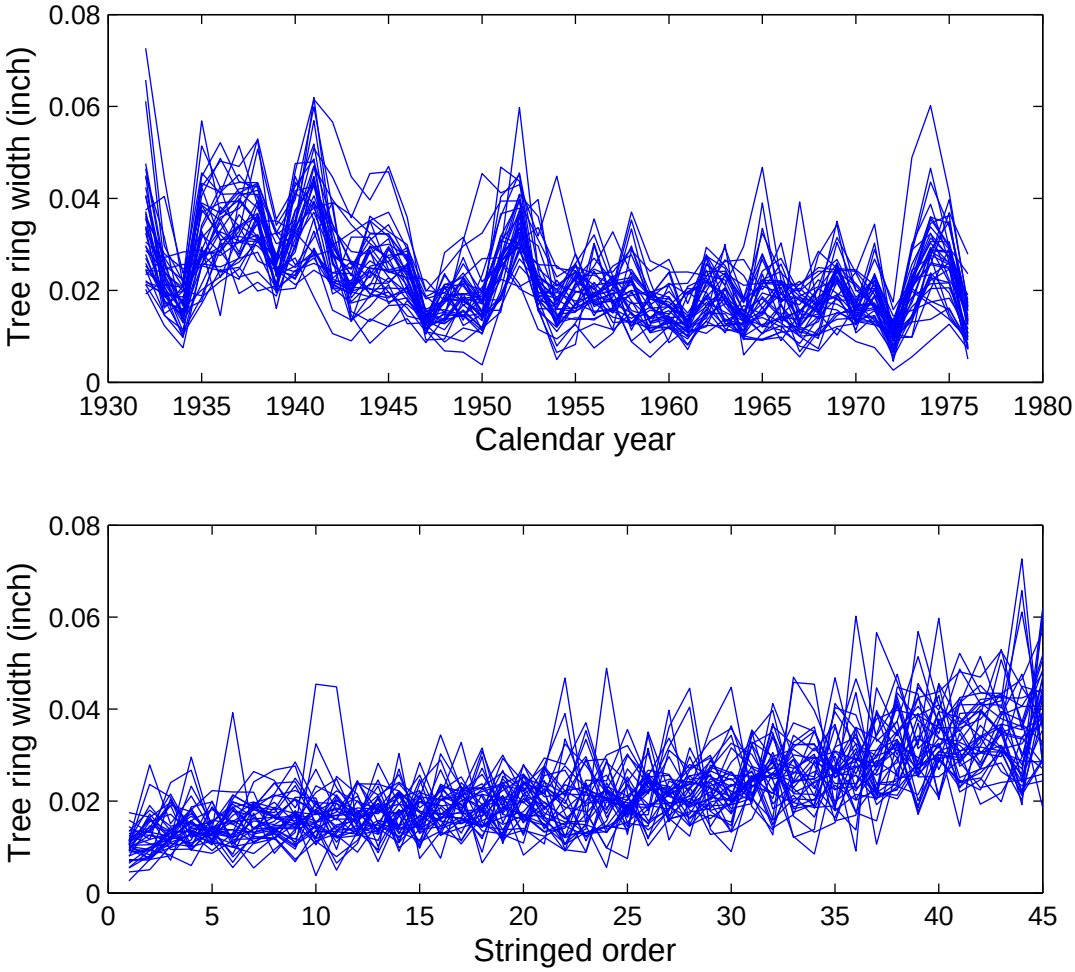

Figure 4: Top: Observed series of tree ring width data. Bottom: Ordered tree ring series obtained by Stringing.

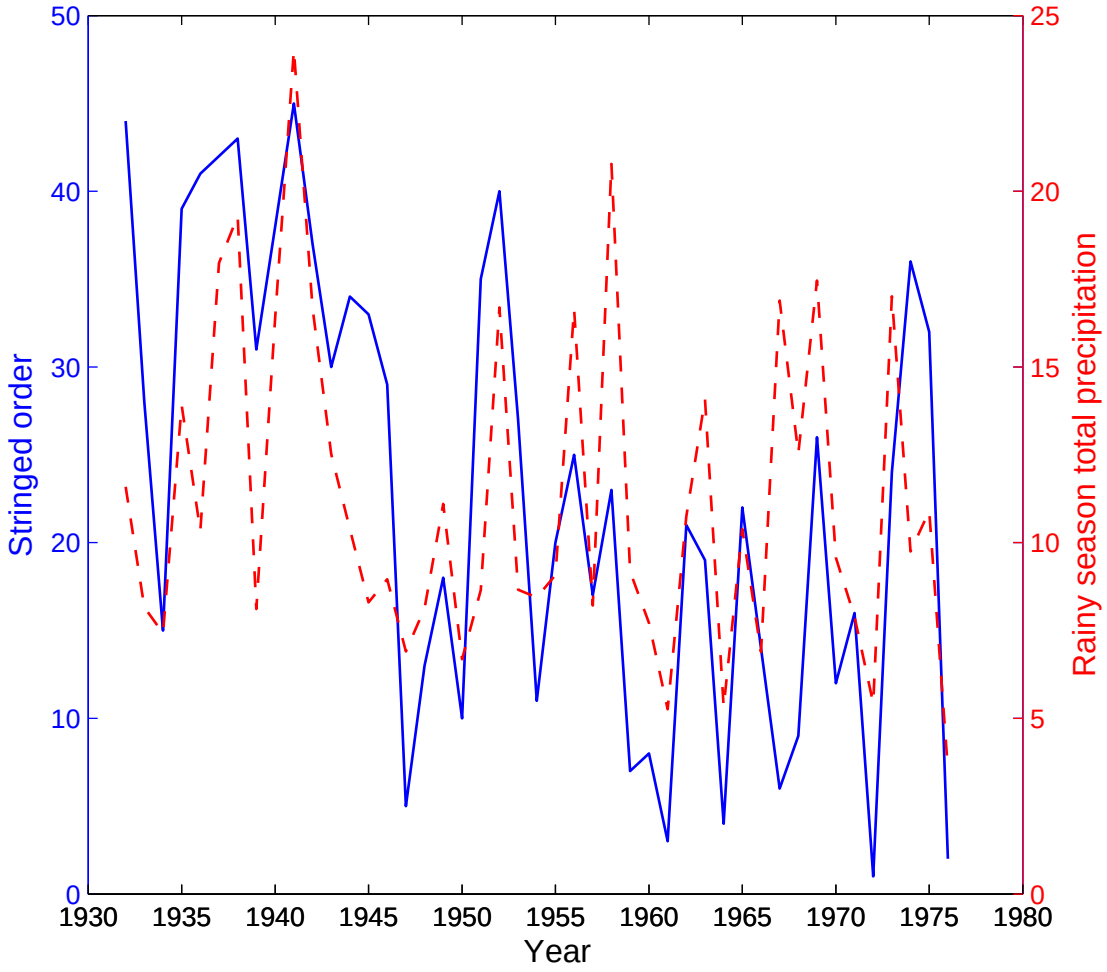

Figure 5: Comparison of the Stringing function (solid) for the tree ring width data with yearly rainy season precipitation (dashed).

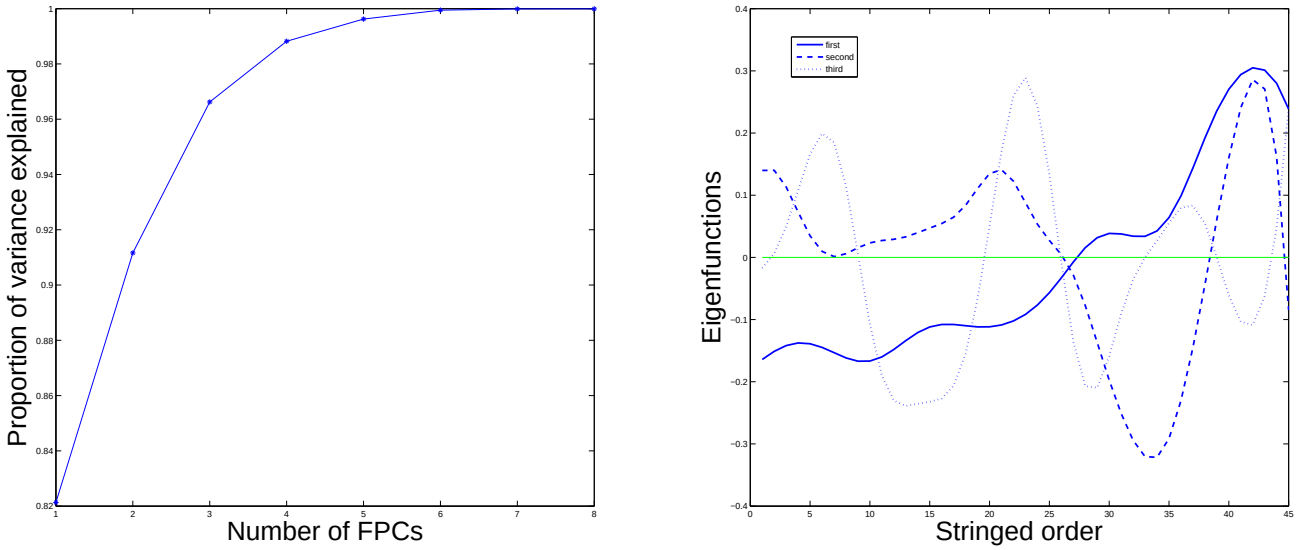

Figure 6: Left: Fraction of variance explained for tree ring widths in dependence on number of included components. The first three components explain more than 95% of the variation in the data. Right: The first three estimated eigenfunctions.

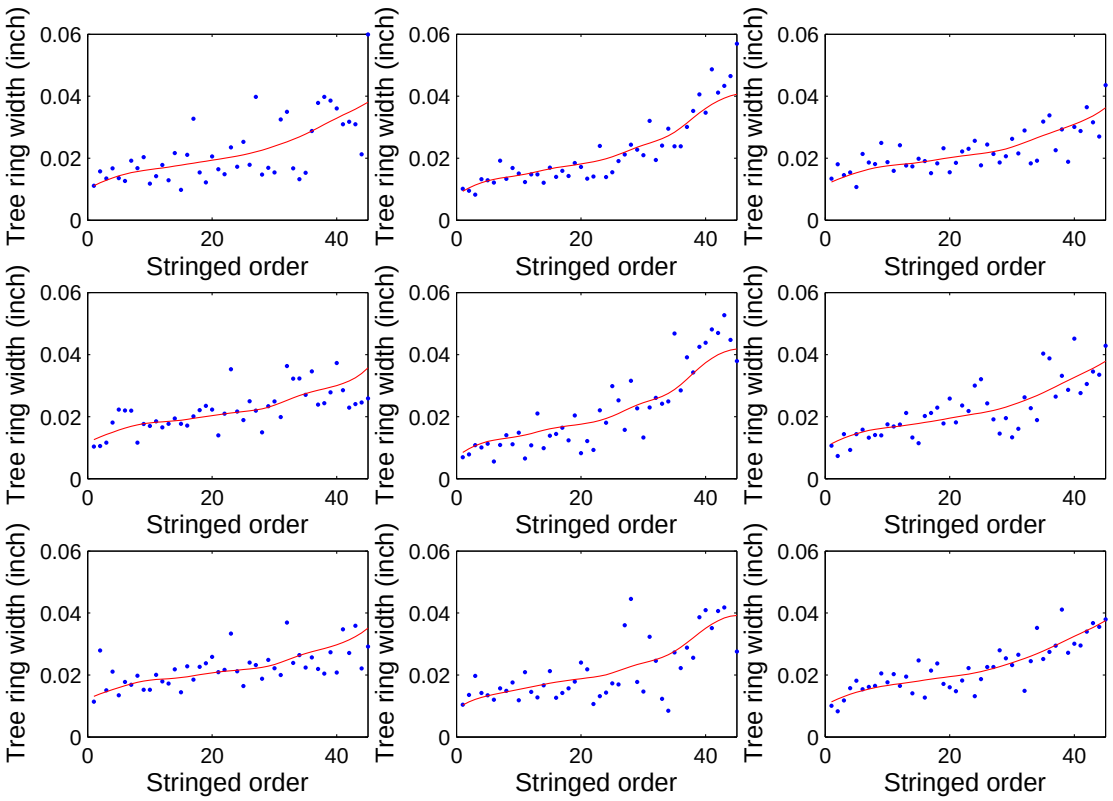

Figure 7: Fitted tree ring width trajectories for nine randomly selected stringed tree ring series, overlaid with stringed measurements.

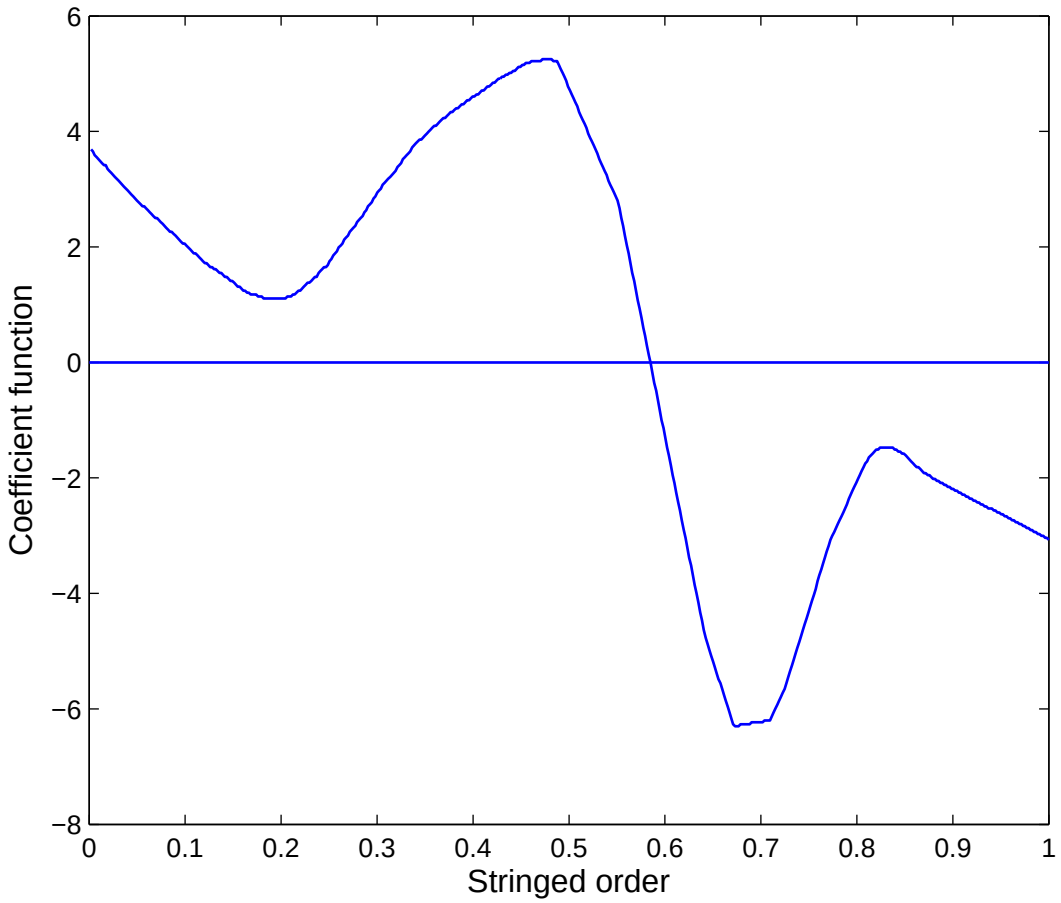

Figure 8: Coefficient function for the stringed functional Cox regression model, as obtained for one random split of the DLBCL gene expression data.