SUPPLEMENTAL MATERIAL

# 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

## Supplement I: Theoretical Considerations

### S.1 Theoretical Results for Stringing

We provide theoretical justifications for Stringing under certain assumptions. Assume that data are generated by a latent predictor process $\{Z \equiv Z(s,\omega), s \in [0,1], \omega \in \Omega\}$, $\Omega$ a probability space, where $\sup_{s\in[0,1]} |Z(s)| < \infty$ a.s. and $Z$ is square integrable with $EZ(t) = 0$ and $EZ^2(t) < \infty$ for all $t \in [0,1]$. Let $s_{j,p}, j = 1,\ldots,p$, denote the (unknown) position of the $j$th predictor. The observed data of subject $i$ correspond to unordered predictor values $x_{i1},\ldots,x_{ip}$, where we postulate that

$$x_{ij} = Z_i(s_{j,p}), \tag{7}$$

with $Z_i$ a realization of $Z$ for the $i$th subject.

Our basic assumption for the theoretical analysis is that the observed data are a randomly permuted version of the original data. Denote $\{\pi_p(1),\ldots,\pi_p(p)\}$ as the (unknown) permutation of $\{1,\ldots,p\}$ such that

$$0 = s_{\pi_p(1),p} < \ldots < s_{\pi_p(p),p} = 1, \tag{8}$$

where the index $p$ will be allowed to vary. Assume for the moment that the true distances $D_{jk}$ between pairs of predictors $(j,k)$ are known, denote by $\{\hat{s}^*_{1,p},\ldots,\hat{s}^*_{p,p}\}$ the 1-d configuration obtained by unidimensional scaling (UDS) based on these true distances, and let $\{\psi_p(1),\ldots,\psi_p(p)\}$ be the permutation of $\{1,\ldots,p\}$ that orders the $\hat{s}^*_{j,p}$, i.e., for which

$$\hat{s}^*_{\psi_p(1),p} < \ldots < \hat{s}^*_{\psi_p(p),p}. \tag{9}$$

We will prove that $\psi_p(j) = \pi_p(j)$ for all $j$ and $p$, $1 \le j \le p$, implying that the underlying order can be identified.

For these theoretical justifications, we require regularity conditions, some of which may only be approximately satisfied in real data settings: The spacing of the underlying support points $s_{j,p}$ should not be too irregular, and the distance between predictors should be monotone increasing with the distance between the true underlying predictor locations.
(A1) For all $p$, $\max_{1\le j\le p}|s_{j,p}-\frac{j-1}{p-1}|\le\frac{C}{p}$, for a constant $C>0$.
(A2) $|s_{j,p}-s_{k,p}|<|s_{j',p}-s_{k',p}|$ if and only if $D_{jk}<D_{j'k'}$ for all $j,k,j',k'$.
(A3) For the position configuration $\{\hat{s}^*_{1,p},\ldots,\hat{s}^*_{p,p}\}$, resulting from UDS, the following holds: If the predictor distances satisfy $D_{jk}>D_{j'k'}$, then $|\hat{s}^*_{j,p}-\hat{s}^*_{k,p}|>|\hat{s}^*_{j',p}-\hat{s}^*_{k',p}|$ for all $j,k,j',k'$.

**Theorem 1.** *Under (A1) - (A3), the rank orders $\psi_p(j)$ of the elements of the solution $(\hat{s}^*_{1,p},\ldots,\hat{s}^*_{p,p})$ of UDS satisfy $\psi_p(j)=\pi_p(j)$, $1\le j\le p$, for all $p$, i.e., UDS identifies the underlying ordering of the predictor locations.*

The process $Z$ has a Karhunen-Loève expansion (Ash and Gardner 1975)

$$Z(s)=\sum_{j=1}^{\infty}\xi_j\phi_j(s),\quad s\in[0,1],\tag{10}$$

where $E\xi_j=0$, $E\xi_j^2=\lambda_j$, with $\sum_{j=1}^{\infty}\lambda_j<\infty$. The functional principal components $\xi_j$ are uncorrelated r.v.s and the $\phi_j$ are the eigenfunctions of the auto-covariance operator of $Z$, defined as $(A_Gf)(s)=\int f(r)G(r,s)\,dr$, which are orthonormal, i.e., $\int\phi_j(s)\phi_k(s)\,ds=1$ for $j=k$, and $=0$ for $j\neq k$. Further assumptions on $Z$ are
(A4) $Z$ is continuous, $Z\in C([0,1])$ and $Z$ is normalized, with $EZ(s)=0$ and $EZ^2(s)=1$ for all $s\in[0,1]$.
(A5) There exist constants $L_j>0, j=1,2,\ldots$, and a constant $C>0$, such that

$$\sup_{s,t\in[0,1]}|\phi_j(t)-\phi_j(s)|\le L_j|t-s|,\quad\sum_{j=1}^{\infty}\lambda_jL_j\le C<\infty.\tag{11}$$

Using the UDS-derived series of permutations $\psi_p$ (see (14)), and noting $U_{\psi_p(j),p}=Z(s_{\psi_p(j),p})$, we define a stochastic process $Z_p$ for each $p$ by

$$Z_p(s)=\sum_{j=1}^{p-1}\left[((p-1)s-j+1)\,U_{\psi_p(j+1),p}+(j-(p-1)s)U_{\psi_p(j),p}\right]I_{[\frac{j-1}{p-1},\frac{j}{p-1}]}(s),\quad s\in[0,1],\tag{12}$$

linearly interpolating the recorded data between any two grid points on $[0,1]$. Denoting weak convergence on the space of continuous random functions $C([0,1])$ by $\Rightarrow$, we obtain the following convergence result for increasing dimensions $p\to\infty$.

**Theorem 2.** *Under assumptions (A1)-(A5), processes $Z_p$ (12) converge weakly to $Z$ (10),*

$$Z_p(s) \Rightarrow Z(s) \text{ on } C([0,1]), \text{ as } p \to \infty. \tag{13}$$

We note that assumption (A5) is always satisfied if process $Z$ has only finitely many components, as then the summation in (11) only extends over finitely many summands. Theorem 2 implies in particular that the distributional properties of the process $Z$ can be inferred from the observed data, if $p$ is sufficiently large and UDS is effective. Since $\mathrm{Cov}(Z(s), Z(t)) = f(s,t)$ is a smooth function in $s$, $t$, we also conclude that $\mathrm{Cov}(Z_p(s'), Z_p(t')) \to \mathrm{Cov}(Z(s), Z(t))$ as $s' \to s$, $t' \to t$ for large $p$. This implies an additional error for the covariance estimator of order $O(\frac{1}{p})$, uniformly in $s$ and $t$, so that known results on covariance function estimation and their relation to eigenvector and eigenvalue estimation are applicable (Yao et al. 2005; Hall and Hosseini-Nasab 2006).

## S.2 Proof of Theorem 1

We first observe that UDS will place the two predictors with maximum distance at the two ends of the interval by (A3), namely,

$$(\psi_p(1), \psi_p(p)) = \underset{1 \le j,k \le p}{\operatorname{argmax}} D_{jk}.$$

Then, by (A2) and (A3), $(Z_{\psi_p(1)p}, Z_{\psi_p(p)p}) = (Z(s_{\pi_p(1),p}), Z(s_{\pi_p(p),p}))$. There are two possible configurations in (6); Configuration A: $\psi_p(1) = \pi_p(1), \psi_p(p) = \pi_p(p)$; and Configuration B: $\psi_p(1) = \pi_p(p), \psi_p(p) = \pi_p(1)$. These two configurations are just mirror images of each other. We restrict our discussion to Configuration A, for which $\psi_p(1) = \pi_p(1), \psi_p(p) = \pi_p(p)$ holds; the argument for Configuration B is analogous. The remaining predictors are then arranged according to their distances with respect to one of the two predictors indexed by $\psi_p(1)$ and $\psi_p(p)$. By (A3), UDS positions the predictor with smallest distance to predictor $\psi_p(1)$ adjacent to predictor $\psi_p(1)$. Therefore,

$$\psi_p(2) = \underset{1 \le k \le p, k \ne \psi_p(1), \psi_p(p)}{\operatorname{argmin}} D_{\psi_p(1)k},$$

and there exists a unique minimum, thus $\psi_p(2) = \pi_p(2)$.

At the $l$th iteration, the locations of the $l-1$ predictors adjacent to the first predictor $\psi_p(1)$ have been identified by UDS. Then the predictor with the smallest distance to the predictor $\psi_p(l-1)$ among the remaining predictors is placed as its nearest neighbor, i.e.,

$$\psi_p(l) = \underset{1 \le k \le p, k \notin \{\psi_p(1), \ldots, \psi_p(l-1), \psi_p(p)\}}{\operatorname{argmin}} D_{\psi_p(l-1)k}, \tag{14}$$

so that $\psi_p(l) = \pi_p(l)$. Continuing this iteration until $p-1$, the correct order of predictors is successively identified. □

## S.3 Proof of Theorem 2

(i) Tightness of $Z_p$: Observe that for any given $u, v \in [0,1]$ there exist indices $1 \le j_1, j_2 \le p$ and $0 \le \alpha_p(u), \alpha_p(v) \le 1$, s.t. $u = \alpha_p(u)s_{j_1,p} + (1-\alpha_p(u))s_{j_1+1p}$ and $v = \alpha_p(v)s_{j_2,p} + (1-\alpha_p(v))s_{j_2+1p}$. Then one can find nonnegative constants $0 < \beta_1, \beta_2, \beta_3 < C_0$ for some constant $C_0 > 0$, such that

$$
\begin{aligned}
& E(Z_p(u) - Z_p(v))^2 \\
\le\ & E\big[\beta_1(Z_p(s_{j_1p}) - Z_p(s_{j_2}p)) + \beta_2(Z_p(s_{j_1+1,p}) - Z_p(s_{j_2+1,p})) + \beta_3(Z_p(s_{j_1+1,p}) - Z_p(s_{j_2,p}))\big]^2 \\
\le\ & \Big\{\beta_1^2 E\big[Z(s_{j_1p}) - Z(s_{j_2p})\big]^2 + \beta_2^2 E\big[Z(s_{j_1+1p}) - Z(s_{j_2+1p})\big]^2 + \beta_3^2 E\big[Z(s_{j_1+1}\,p) - Z(s_{j_2p})\big]^2 \\
& +2\beta_1\beta_2\big\{E[Z(s_{j_1p}) - Z(s_{j_2p})]^2 E[Z(s_{j_1+1,p}) - Z(s_{j_2+1,p})]^2\big\}^{1/2} \\
& +2\beta_1\beta_3\big\{E[Z(s_{j_1p}) - Z(s_{j_2p})]^2 E[Z(s_{j_1+1,p}) - Z(s_{j_2,p})]^2\big\}^{1/2} \\
& +2\beta_2\beta_3\big\{E[Z(s_{j_1+1,p}) - Z(s_{j_2+1,p})]^2 E[Z(s_{j_1+1,p}) - Z(s_{j_2,p})]^2\big\}^{1/2}\Big\}. \qquad (15)
\end{aligned}
$$

For any $u, v \in [0,1]$, according to (11),

$$
\begin{aligned}
E\big[Z(u) - Z(v)\big]^2 &\le E\big\{\sum \xi_j(\phi_j(u) - \phi_j(v))\big\}^2 \\
&\le \sum \lambda_j L_j |u-v|^2 \le C|u-v|^2. \qquad (16)
\end{aligned}
$$

Plugging this into (15), we find

$$
\begin{aligned}
E(Z_p(u) - Z_p(v))^2 &\le C\big(\beta_1(s_{j_1p} - s_{j_2p}) + \beta_2(s_{j_1+1,p} - s_{j_2+1,p}) + \beta_3(s_{j_1+1,p} - s_{j_2,p})\big)^2 \\
&\le \tilde{C}|u-v|^2, \qquad (17)
\end{aligned}
$$

for a constant $\tilde{C}$. According to criterion (12.51) of Billingsley (1968), this implies tightness of processes $Z_p$.

(ii) Convergence of finite-dimensional distributions: Let $u_1, \ldots, u_k \in [0,1]$, for any arbitrary but fixed $k \ge 1$. Let $s_{j_1(l)p}, s_{j_2(l)p}$ be such that $s_{j_1(l)p} = \max\{s_{jp} : s_{jp} < u_l\}$ and $s_{j_2(l)p} = \min\{s_{jp} : s_{jp} \ge u_l\}$. Then $j_2(l) = j_1(l) + 1$ and $s_{j_2(l)p} - s_{j_1(l)p} = \frac{1}{p-1}$ and $s_{j_1(l)p} \le u_l \le s_{j_2(l)p}$, for $1 \le l \le k$. Due to the a.s. continuity of the trajectories of $Z$, there exists a $v_{jp}$ such that $s_{j_1(l)p} \le v_{jp} \le s_{j_2(l)p}$ and $Z(v_{jp}) = Z_p(u_j)$. Since $|u_j - v_{jp}| \le \frac{1}{p-1}$, property (16) implies that $Z(v_{jp}) - Z(u_j) = O_p(1/p)$, $1 \le j \le k$, and therefore for all $a_1, \ldots, a_k$, $P(Z_p(u_1) \le a_1, \ldots, Z_p(u_k) \le a_k) = P(Z(v_{1p}) \le a_1, \ldots, Z(v_{kp}) \le a_k) \to P(Z(u_1) \le a_1, \ldots, Z(u_p) \le a_k)$, as $p \to \infty$. Tightness and convergence of finite-dimensional distributions then imply the weak convergence stated in Theorem 2.

## REFERENCES


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

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

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.

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.

## Supplement II: Additional Simulation Results

We provide here additional boxplots for the results obtained in Simulation 3, which is described in Section 3. This complements the results reported in Figure 3 and in Table 3.

The following boxplots demonstrate that for less sparse scenarios, Stringing not only has smaller median MSE but also smaller interquartile range and fewer outliers, respectively smaller maxima of MSE. Thus Stringing appears to provide more reliable and better results for these cases, especially when the ratio of $p/n$ is large, as in $p = 100$ and $n = 30$. Overall, the advantages of Stringing are more pronounced for the binary response case. The analogue to Figure 3, which is for continuous responses and uniformly generated covariances, for the situation of binary responses is a case in point. The corresponding boxplots are shown in Figure 9. Note that the third quartiles of the Misclassification Rate observed for Stringing are always below the medians of those for Lasso, except for one panel in the lower right corner, where $\beta$ is sparse, with nonzero percentage of 10% for $\beta$, and the sample size is $n = 60$ and predictor dimension is $p = 50$. Even for this case, the maximal misclassification rate observed for Lasso is larger than the maximum observed for Stringing.

The case where predictor covariance structure is $0.5^{\sqrt{|i-j|}}$, corresponding to relatively low predictor correlations, is most favorable for Lasso, relative to Stringing, when compared with the other covariance structures, for which the advantages of Stringing are more pronounced. But even in this least favorable case, Stringing still performs reasonably well. This is demonstrated by the boxplots in Figure 10 for the continuous response case and in Figure 11 for the binary response case. What is particularly remarkable, is that for small $n$, large $p$, such as $p = 100$, $n = 30$, Stringing still performs much better than Lasso even in such very sparse cases with low predictor correlation. As expected, the relative performance of Stringing as compared to Lasso deteriorates as the regression parameters $\beta$ become sparser, and this feature is observed across all simulation scenarios.

For predictors with covariance structures $0.9^{\sqrt{|i-j|}}$ and $\min(i, j)$, the boxplots in Figures 12 - 15 reflect similar results as those obtained for the random covariance structure $\mathcal{U}(0, 1)$ (Figure 3 in Section 3 and Figure 9). The few cases where Lasso performs better, are characterized by very sparse predictors (10%) or relatively low $p/n$ ratio.

Overall, these additional results confirm the observations described in Section 3. In particular, one finds that Stringing works better than Lasso for situations where either one of the following holds: (a) regression parameters are less sparse; (b) predictor correlations are higher; (c) predictor dimension $p$ is large.

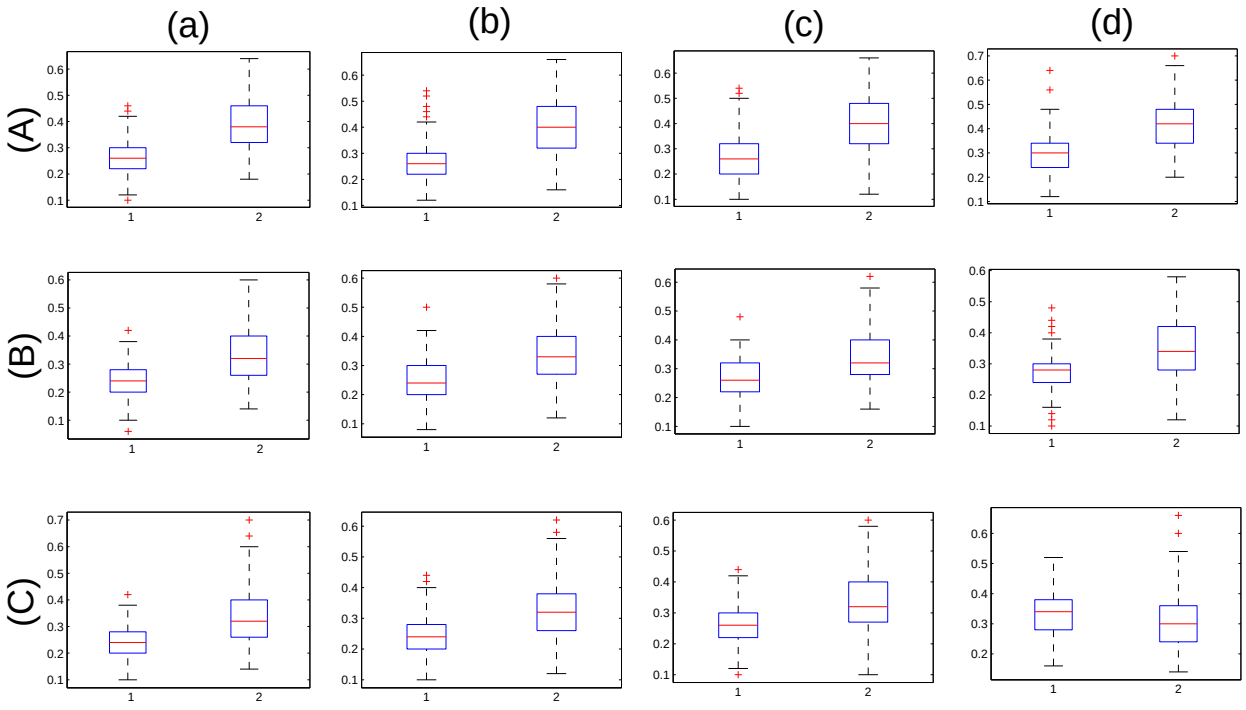


Figure 9: Boxplots of Misclassification Rates obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with binary responses and predictor covariance structure $\mathcal{U}(0,1)$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

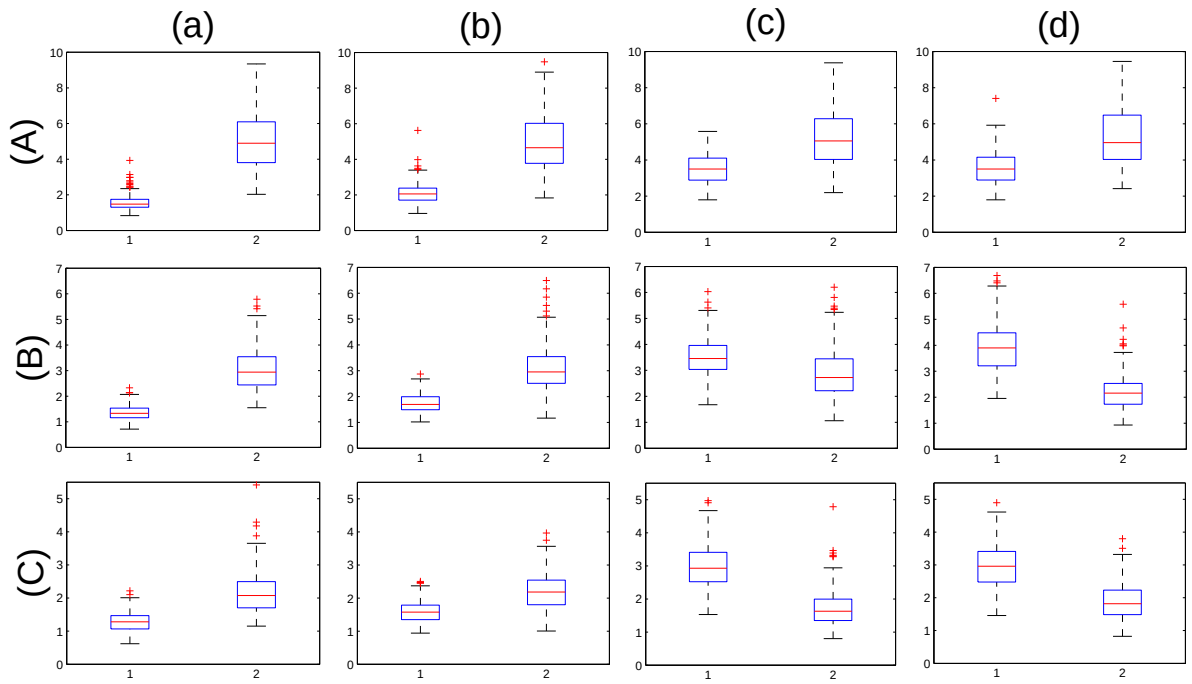


Figure 10: Boxplots of MSEs obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with continuous responses and predictor covariance structure $0.5^{\sqrt{|i-j|}}$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

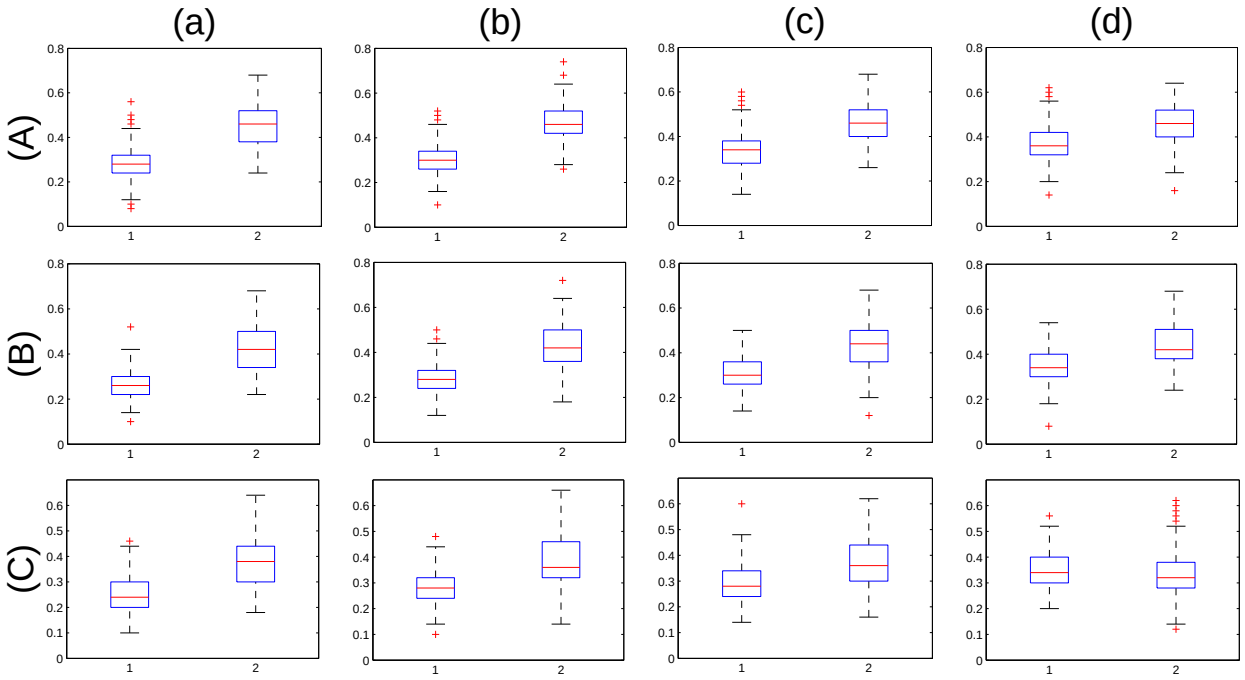


Figure 11: Boxplots of Misclassification Rates obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with binary responses and predictor covariance structure $0.5^{\sqrt{|i-j|}}$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

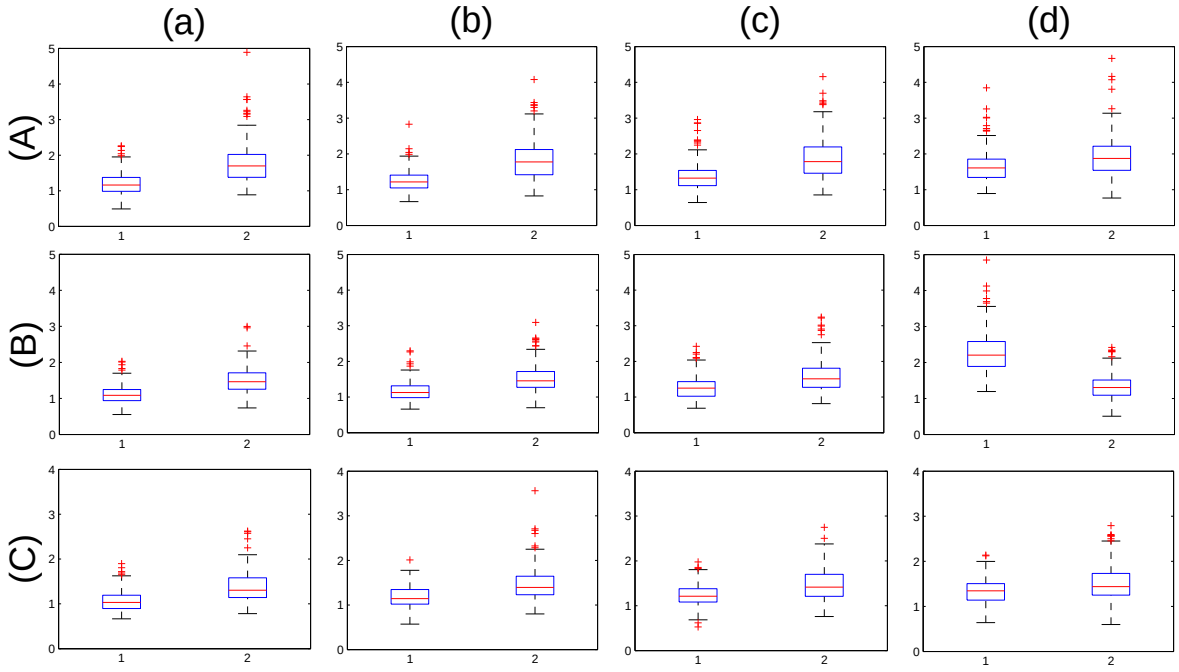


Figure 12: Boxplots of MSEs obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with continuous responses and predictor covariance structure $0.9^{\sqrt{|i-j|}}$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

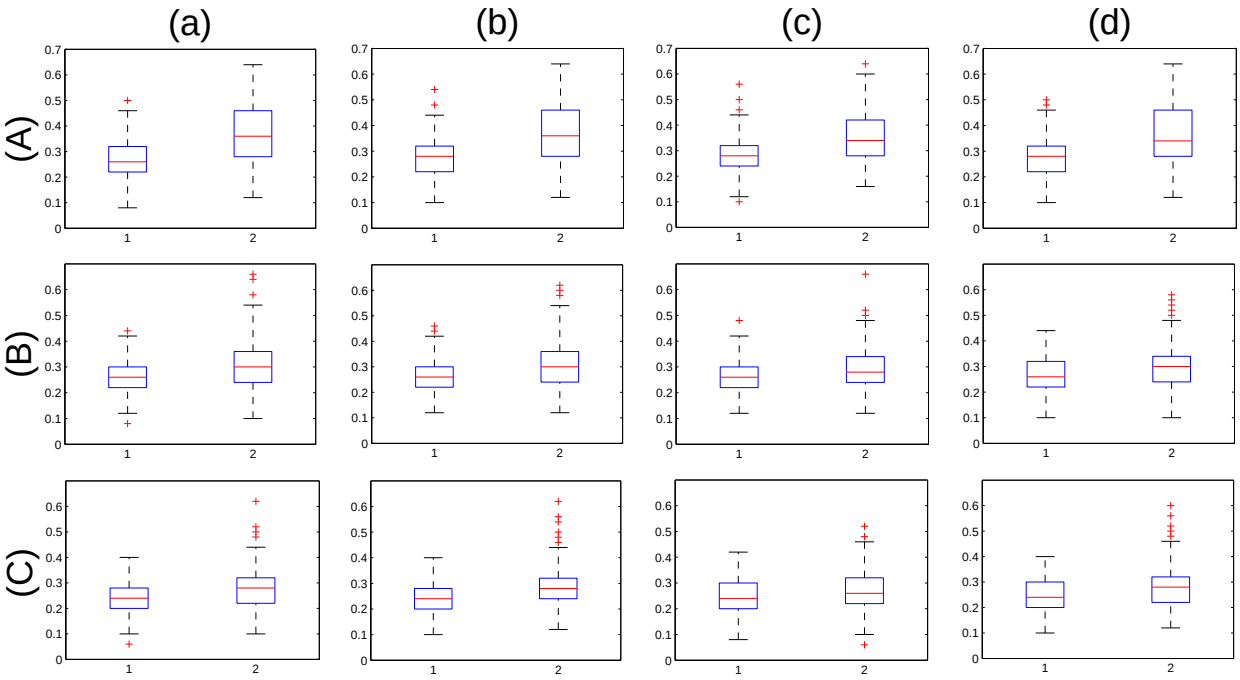


Figure 13: Boxplots of Misclassification Rates obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with binary responses and predictor covariance structure $0.9^{\sqrt{|i-j|}}$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

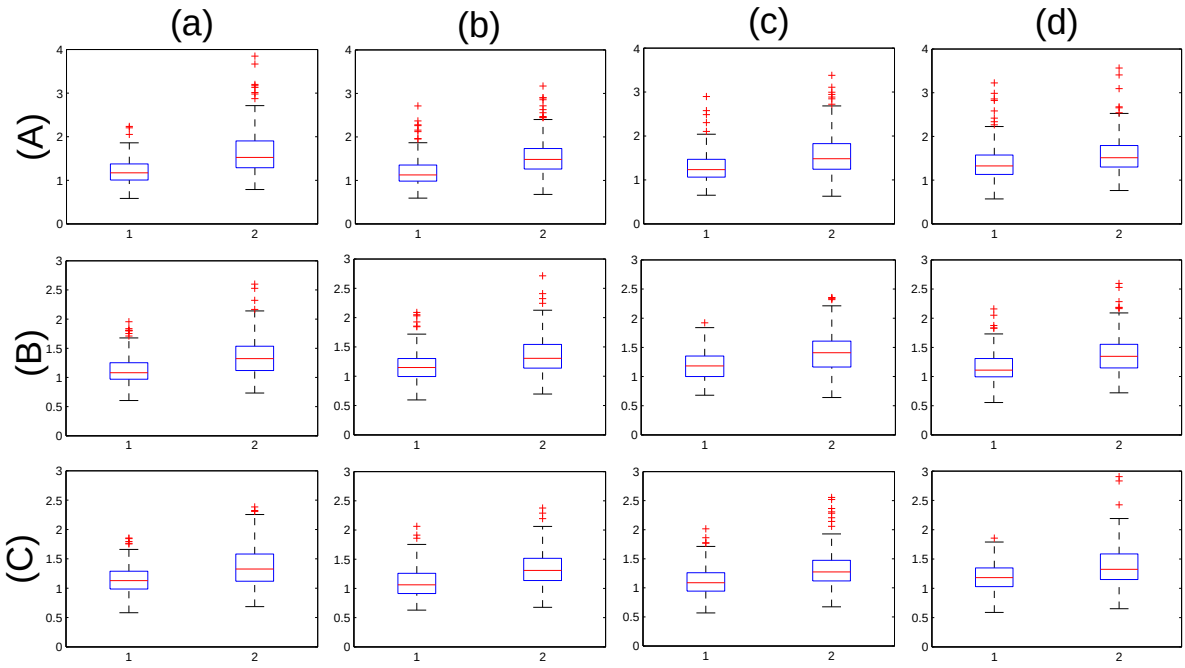


Figure 14: Boxplots of MSEs obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with continuous responses and predictor covariance structure $\min(i, j)$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).

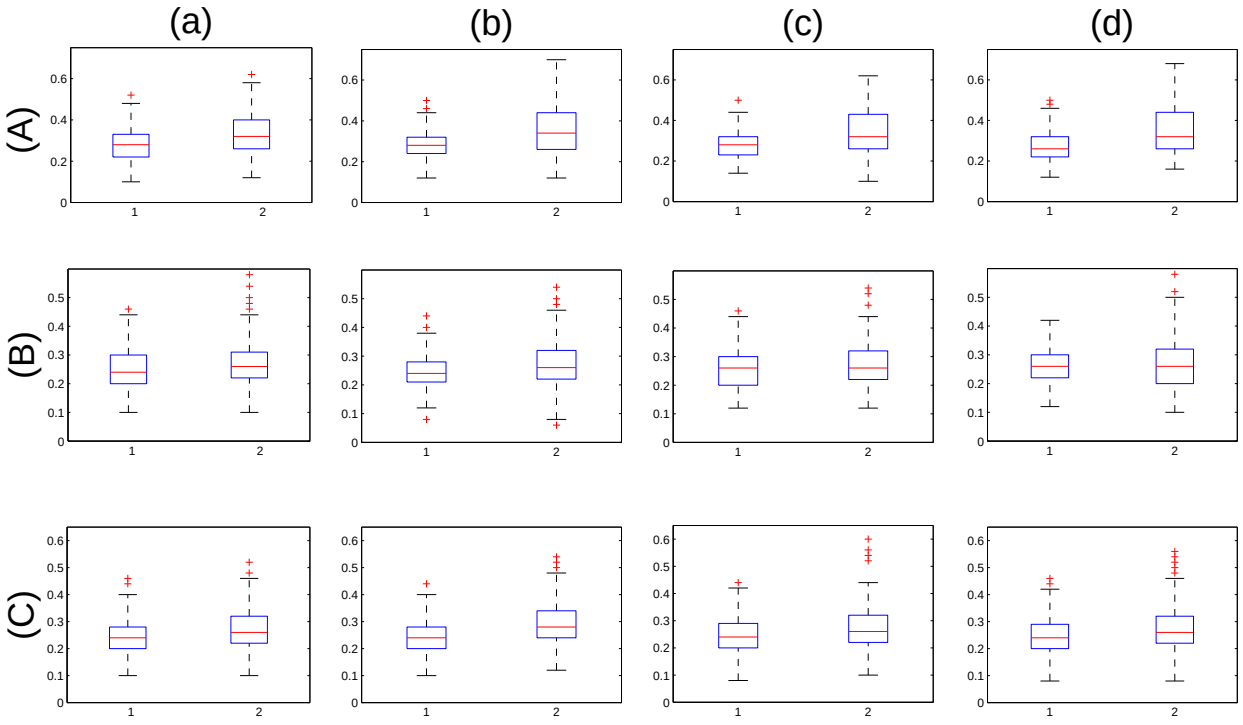


Figure 15: Boxplots of Misclassification Rates obtained from 200 simulation runs (*Simulation 3*), comparing Stringing and Lasso, for samples of $n$ high-dimensional predictor $p$-vectors with binary responses and predictor covariance structure $\min(i, j)$. Columns indicate level of sparsity (percentage of non-zero $\beta$): (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).