Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

February 9, 2018

1 Simulation Studies

In this section, we evaluate the performance of the spline estimator under different simulation settings when the tuning parameters are chosen by the unbiased risk estimate and leave-one-subject-out cross validation. We compare the performance of the maximum penalised likelihood estimator using the classical smoothness penalty under the smoothing spline representation to the performance of the tensor product P-spline estimator for varying orders of the penalty. We also compare performance under complete data to the performance under irregularly sampled data:

- All subjects share a common set of observation times $t_1, \ldots, t_M$.
- Observation times vary across subjects, with subject-specific deviation defined as follows:

1.1 Alternative estimators for benchmarking

For the case of common observation times across all subjects, we also consider three other methods of estimating a covariance matrix for comparison: the sample covariance matrix Σ^*, the soft thresholding estimator of Rothman et al. [2009], and the tapering estimator of Cai et al. [2010]. The soft-thresholding estimator proposed in Rothman et al. [2009] is given by

$$\hat{\Sigma}^{sthresh}\left(\lambda\right) = \left[\mathrm{sign}\left(\sigma^*_{ij}\right)\left(\sigma^*_{ij} - \lambda\right)_+\right],$$

where σ^*_{ij} denotes the i-j^{th} entry of the sample covariance matrix, and λ is a penalty parameter controlling the amount of shrinkage applied to the empirical estimator. The tapering estimator presented in Cai et al. [2010] is defined

$$\hat{\Sigma}^{taper}\left(\lambda\right) = \left[\omega^k_{ij} s_{ij}\right].$$

*The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201
†The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201

The weights ω_{ij}^{λ} are given by

$$\omega_{ij}^{k} = k_h^{-1}\left[(k - |i-j|)_+ - (k_h - |i-j|)_+\right],$$

The superscript on the weights ω_{ij}^{k} serves to indicate that these are controlled by a tuning parameter, k, which controls the amount of shrinkage applied to the elements of the sample covariance matrix. Without loss of generality, we assume that $k_h = k/2$ is even. The weights may be rewritten as

$$\omega_{ij} = \begin{cases} 1, & ||i-j|| \leq k_h \\ 2 - \frac{i-j}{k_h} & k_h < ||i-j|| \leq k, \\ 0 & \text{otherwise} \end{cases}$$

[discuss the MCRE and CVTuningCov package]

[discuss the implementation and R package here]

1.2 Loss functions for evaluating estimators

To assess performance of estimator $\hat{\Sigma}$, we consider two commonly used loss functions:

$$\Delta_1\left(\Sigma, \hat{\Sigma}\right) = tr\left(\Sigma^{-1}\hat{\Sigma}\right) - log|\Sigma^{-1}\hat{\Sigma}| - M, \tag{1}$$

$$\Delta_2\left(\Sigma, \hat{\Sigma}\right) = tr\left(\left(\Sigma^{-1}\hat{\Sigma} - \mathrm{I}\right)^2\right) \tag{2}$$

where Σ is the true covariance matrix and $\hat{\Sigma}$ is an $M \times M$ positive definite matrix. Each of these loss functions is 0 when $\hat{\Sigma} = \Sigma$ and is positive when $\hat{\Sigma} \neq \Sigma$. Both measures of loss are invariant with respect to linear transformations of the data, Y,

$$CY,$$

which corresponds to transformations

$$\hat{\Sigma}^* = C\hat{\Sigma}C', \quad \Sigma^* = C\Sigma C',$$

for a nonsingular matrix C. The first loss Δ_1 is commonly referred to as the entropy loss; it gives the Kullback-Leibler divergence of two multivariate Normal densities with the same mean corresponding to the two covariance matrices. The second loss Δ_2, or the quadratic loss, measures the difference between $\Sigma^{-1}\hat{\Sigma}$ and the identity matrix with the squared Frobenius norm which is given by

$$||A||^2 = \mathrm{tr}\left(AA'\right)||.$$

The quadratic loss consequently penalizes overestimates more than underestimates, so "smaller" estimates are favored more under Δ_2 than Δ_1.

TO DO: define more clearly the notion of an overestimate versus an underestimate - i.e. for matrices A, B, we define $A >> B$ if $A - B$ is non-negative definite.

We obtain the corresponding risk functions by taking expectations,

$$R_i\left(\Sigma, \hat{\Sigma}\right) = E_\Sigma\left[\Delta_i\left(\Sigma, \hat{\Sigma}\right)\right], \quad i = 1, 2.$$

We prefer estimator $\hat{\Sigma}_1$ over another estimator $\hat{\Sigma}_2$ if it has smaller risk. Given Σ, we estimate the risk of an estimator via Monte Carlo approximation.

1.3 Simulation study design

TO DO: new leading sentence to introduce the 5 simulation settings. we consider four Cholesky covariance structures for the underlying generating distribution:

I. Mutual independence: $\Sigma = T^{-T}D^2T^{-1} = \mathrm{I}$ where

$$\begin{aligned} \phi(t,s) &= 0, \quad 1 \le t < s \le M; \\ \sigma^2(t) &= 1, \quad t = 1, \ldots, M. \end{aligned}$$

II. Linear varying coefficient model with constant innovation variance: $\Sigma_2 = T^{-T}D^2T^{-1}$ where

$$\begin{aligned} \phi(t,s) &= t - \frac{1}{2M}, \quad 1 \le s < t \le M \\ \sigma^2(t) &= 0.1, \quad t = 1, \ldots, M. \end{aligned}$$

III. AR(k) model with linear varying coefficient: $\Sigma_3 = T^{-T}D^2T^{-1}$ where $k = \lfloor M/2 \rfloor + 1$ and

$$\begin{aligned} \phi(t,s) &= \begin{cases} t - \frac{1}{2M}, & t - s \le \lfloor M/2 \rfloor + 1 \\ 0, & t - s > 1 \end{cases}, \\ \sigma^2(t) &= 0.1, \quad t = 1, \ldots, M. \end{aligned}$$

IV. AR(1) model with linear varying coefficient: $\Sigma_3 = T^{-T}D^2T^{-1}$ where

$$\begin{aligned} \phi(t,s) &= \begin{cases} t - \frac{1}{2M}, & t - s = 1 \\ 0, & t - s > 1 \end{cases}, \\ \sigma^2(t) &= 0.1, \quad t = 1, \ldots, M. \end{aligned}$$

V. The compound symmetry model: $\Sigma_4 = \sigma^2\left(\rho \mathrm{J} + (1-\rho)\,\mathrm{I}\right),\ \rho = 0.7,\ \sigma^2 = 1.$

$$\begin{aligned} \phi_{ts} &= -\frac{\rho}{1 + (t-1)\rho}, \quad t = 2, \ldots, M, \ \ s = 1, \ldots, t-1 \\ \sigma_t^2 &= 1 - \frac{(t-1)\rho^2}{1 + (t-1)\rho}, \ \ t = 2, \ldots, M. \end{aligned}$$

TO DO - this was moved from the introductory sentence for this subsection: using $N_{sim} = 100$ simulation runs for each scenario outlined above. Estimation is performed on data generated according to an M-dimensional multivariate Normal distribution with mean zero.

1.4 Results

Table 1: Simulation results for $\Sigma_1 = \mathrm{I}$ under quadratic loss, Δ_1. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0010	0.0013	0.4702	0.3926	0.3871
	20	0.0007	0.0006	0.8495	0.8301	0.8287
	30	0.0003	0.0004	1.1449	1.1926	1.1924
$N = 100$	10	0.0004	0.0004	0.2072	0.1802	0.1777
	20	0.0002	0.0002	0.3920	0.3858	0.3817
	30	0.0001	0.0001	0.5712	0.6191	0.6109

Table 2: Simulation results for Σ_2, the linear varying coefficient AR model, under quadratic loss, Δ_1. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0314	0.0411	0.5726	0.5810	0.7758
	20	0.3266	0.7265	2.3130	5.5964	2.7545
	30	5.0696	4.9073	15.1096	765.7206	28.6820
$N = 100$	10	0.0156	0.0147	0.2479	0.2501	0.3544
	20	0.1894	0.2017	1.3177	5.1945	4.7634
	30	2.3876	1.6465	9.8175	488.6801	85.9508

Table 3: Simulation results for Σ_3, the k-banded linear varying coefficient AR model with $k = \lfloor M/2 \rfloor + 1$, under quadratic loss, Δ_1. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0562	0.0547	0.5237	0.5810	0.5313
	20	0.7832	0.8934	2.1419	9.5721	9.1421
	30	8.2650	10.6855	15.2842	407.3659	129.7459
$N = 100$	10	0.0376	0.0449	0.2546	0.2556	0.2661
	20	0.6260	0.5967	1.3751	3.3281	1.2759
	30	5.7635	6.2824	7.4750	203.6710	10.0634

Table 4: Simulation results for Σ_4, the 2-banded linear varying coefficient AR model, under quadratic loss, Δ_1. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0134	0.0145	0.4169	0.3987	0.3985
	20	0.0590	0.0574	0.8810	0.9078	0.9073
	30	0.1351	0.1362	1.2571	1.2570	1.2575
$N = 100$	10	0.0077	0.0078	0.2263	0.2111	0.2104
	20	0.0549	0.0534	0.4309	0.4127	0.4120
	30	0.1331	0.1320	0.6819	0.6579	0.6565

Table 5: Simulation results for Σ_5, the compound symmetry model, under quadratic loss, Δ_1. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.3688	0.3599	0.7872	0.8058	1.4774
	20	0.9770	0.9954	1.6167	1.7840	3.4516
	30	1.6067	1.6151	2.5548	2.4837	4.9027
$N = 100$	10	0.3210	0.3168	0.3913	0.3819	0.8958
	20	0.9793	0.9774	0.8714	0.8479	2.2110
	30	1.6177	1.6032	1.2967	1.2293	3.4968

Table 6: Simulation results for $\Sigma_1 = \mathrm{I}$ under entropy loss, Δ_2. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0684	0.0678	1.2339	0.4451	1.1760
	20	0.0799	0.0720	5.0827	1.6504	4.7847
	30	0.0668	0.0740	12.5162	1.9975	11.0434
$N = 100$	10	0.0405	0.0379	0.5854	0.1783	0.5201
	20	0.0356	0.0378	2.3038	0.4394	1.9637
	30	0.0396	0.0322	5.2641	0.6717	4.5410

Table 7: Simulation results for Σ_2, the linear varying coefficient AR model, under entropy loss, Δ_2. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.0647	0.0696	1.2431	1.4242	1.1195
	20	0.0884	0.0969	5.0437	17.0220	13.5290
	30	0.0702	0.0894	12.4559	39.9769	159.0521
$N = 100$	10	0.0307	0.0302	0.5403	0.7659	0.5609
	20	0.0357	0.0350	2.3195	10.0140	12.1431
	30	0.0372	0.0334	5.2817	35.0353	108.1015

Table 8: Simulation results for Σ_3, the k-banded linear varying coefficient AR model with $k = \lfloor M/2 \rfloor + 1$, under entropy loss, Δ_2. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.3354	0.3174	1.1947	1.1073	1.1649
	20	1.1144	1.1143	5.0966	17.0220	12.6171
	30	2.3247	2.3168	12.4905	50.3684	101.8245
$N = 100$	10	0.2826	0.2955	0.5446	0.5410	0.5531
	20	1.0690	1.0627	2.3514	12.8490	11.4934
	30	2.2737	2.2767	5.4204	27.2736	30.5818

Table 9: Simulation results for Σ_4, the 2-banded linear varying coefficient AR model, under entropy loss, Δ_2. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.2605	.2743	1.1692	0.5899	1.1126
	20	0.8836	.8764	5.0899	1.8834	4.6363
	30	1.6087	1.6195	12.5844	3.1902	11.4818
$N = 100$	10	0.2193	0.2183	0.5642	0.2902	0.5456
	20	0.8468	0.8491	2.2607	0.7869	2.2028
	30	1.5743	1.5802	5.2437	1.1974	4.8555

Table 10: Simulation results for Σ_5, the compound symmetry model, under entropy loss, Δ_2. The risk functions for the sample covariance matrix, the tapered estimator, the soft thresholding estimator, the SSANOVA Cholesky estimator, and the tensor product P-spline Cholesky estimator. The tuning parameters for the tapering estimator and the soft thresholding estimator were chosen using $K = 10$-fold cross validation. The performance of the spline estimators is evaluated when both the unbiased risk estimate and leave-one-subject-out cross validation are used to select the smoothing parameters.

		$\hat{\Sigma}_{ssanova}$		S	S_{st}	S_{taper}
	M	LosoCV	URE			
$N = 50$	10	0.2837	0.2766	1.1943	17.3871	1.2122
	20	0.7551	0.7657	5.0283	35.4067	5.1671
	30	1.1936	1.1927	12.5871	46.5337	12.4110
$N = 100$	10	0.2449	0.2390	0.5734	16.2705	0.5796
	20	0.7231	0.7299	2.2678	31.3226	2.2988
	30	1.1780	1.1813	5.2562	39.2108	5.2592

1.5 Discussion

Like other estimators based on shrinking the sample covariances matrix, the soft thresholding estimator is not guaranteed to be positive definite, though Rothman et al. [2009] established that in the limit, soft thresholding produces a positive definite estimator with probability tending to 1. We observed simulations runs which yielded a soft thresholding estimator that was indeed not positive definite. Evaluation of the entropy loss 2 is undefined at an estimator having at least one eigenvalue that is not greater than zero. To enable the evaluation of the entropy loss, we coerced these estimates to the "nearest" positive definite estimate via application of the technique presented in Cheng and Higham [1998]. For a symmetric matrix A, which is not positive definite, a modified Cholesky algorithm produces a symmetric perturbation matrix E such that $A + E$ is positive definite.

See Pourahmadi [2011] section 3.1 for further discussion of loss functions

References

T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Sheung Hun Cheng and Nicholas J Higham. A modified cholesky algorithm based on a symmetric indefinite factorization. *SIAM Journal on Matrix Analysis and Applications*, 19(4):1097–1110, 1998.

Mohsen Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.