Nonparametric Covariance Estimation for Longitudinal Data

Tayler A. Blake

The Ohio State University, Department of Statistics

July 30, 2018

Background

Let

$$Y = (y_1, \ldots, y_p)', \quad (t_1, \ldots, t_p)'$$

denote the random vector of observations and their associated measurement times, where

$$Cov\,(Y) = \Sigma = [\sigma_{ij}]$$

- Dimensionality: the number of parameters is quadratic in p.

Background

Let

$$Y = (y_1, \ldots, y_p)', \quad (t_1, \ldots, t_p)'$$

denote the random vector of observations and their associated measurement times, where

$$Cov\,(Y) = \Sigma = [\sigma_{ij}]$$

- Dimensionality: the number of parameters is quadratic in p.
- Covariance estimates should satisfy the positive definite constraint:

$$c'\Sigma c = \sum_{i,j=1}^{p} c_i c_j \sigma_{ij} \geq 0$$

Background

Let

$$Y = (y_1, \ldots, y_p)', \quad (t_1, \ldots, t_p)'$$

denote the random vector of observations and their associated measurement times, where

$$Cov\,(Y) = \Sigma = [\sigma_{ij}]$$

- Dimensionality: the number of parameters is quadratic in p.
- Covariance estimates should satisfy the positive definite constraint:

$$c'\Sigma c = \sum_{i,j=1}^{p} c_i c_j \sigma_{ij} \geq 0$$

- Observation times may be irregular or subject-specific.

The Modified Cholesky decomposition

For any positive definite Σ, there exists a unique lower-triangular matrix $C = [c_{ij}]$, $c_{ii} > 0$:

$$\Sigma = CC',$$

Let $D^{1/2} = diag\,(c_{11}, \ldots, c_{pp})$, $L = D^{-1/2}C$, then

$$\Sigma = LDL'.$$

The **modified Cholesky decomposition** (MCD) of Σ is given by

$$D = T\Sigma T', \tag{1}$$

where $T = L^{-1}$. The lower triangular entries of T are *unconstrained*.

Statistical Interpretation of (T, D)

Let $\hat{y}_t$ be the linear least-squares predictor of y_t based on previous measurements $y_{t-1}, \ldots, y_1$ and $\epsilon_t = y_t - \hat{y}_t$ denote the corresponding mean zero prediction error with variance $Var\left(\epsilon_t\right) = \sigma_t^2$. We can find unique scalars ϕ_{tj}:

$$y_t = \begin{cases} \epsilon_t, & t = 1 \\ \sum_{j=1}^{t-1} \phi_{tj} y_j + \epsilon_t, & t = 2, \ldots, p, \end{cases}$$

where $D = Cov\left(\epsilon\right) = diag\left(\sigma_1^2, \ldots, \sigma_p^2,\right)$. Then

$$\underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{p-1} \\ \epsilon_p \end{bmatrix}}_{\epsilon} = \underbrace{\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \ldots & -\phi_{p,p-1} & 1 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p-1} \\ y_p \end{bmatrix}}_{Y}$$

Taking the covariance on both sides gives the MCD (1).

The coefficients and prediction error variances of successive regressions are unconstrained.

The **generalized autoregressive parameters** ϕ_{tj} and **log innovation variances** $\log \sigma_j^2$ are unconstrained but

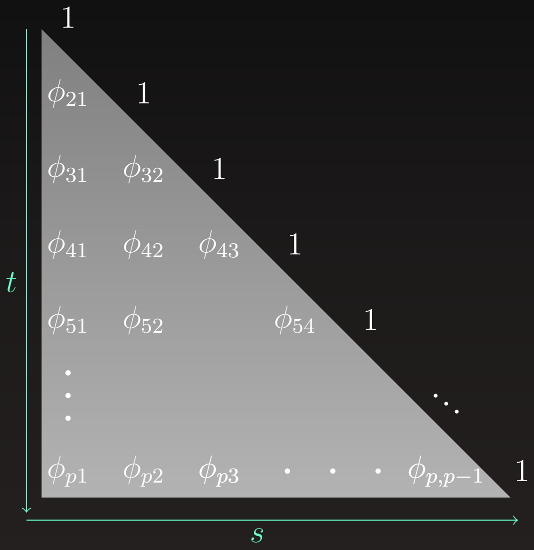
$$\hat{\Sigma}^{-1} = \hat{T}' \hat{D}^{-1} T$$

is guaranteed to be positive definite.

$$
\begin{array}{cccccc}
y_1 & y_2 & y_3 & \cdots & y_{p-1} & y_p \\
\hline
1 & & & & & \\
\phi_{21} & 1 & & & & \\
\phi_{31} & \phi_{32} & 1 & & & \\
\vdots & \vdots & & \ddots & & \\
\vdots & \vdots & & & \ddots & \\
\phi_{p1} & \phi_{p2} & \cdots & \cdots & \phi_{p,p-1} & 1 \\
\hline
\sigma_1^2 & \sigma_2^2 & \cdots & \cdots & \sigma_{p-1}^2 & \sigma_p^2
\end{array}
$$

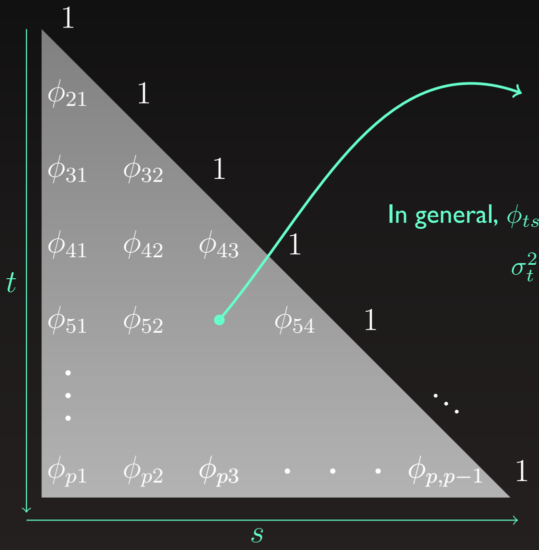
Accommodating Unbalanced Data

Treating ϕ as a continuous bivariate function.



Accommodating Unbalanced Data

Treating ϕ as a continuous bivariate function.



$$\phi_{53} = \tilde{\phi}\left(t_5, t_3\right)$$

In general, $\phi_{ts} = \tilde{\phi}\left(t, s\right), \quad 0 \le s < t \le 1$

$$\sigma_t^2 = \sigma^2\left(t\right), \quad 0 \le t \le 1$$

A functional varying coefficient model for ϕ

Assume measurements $Y_i = (y_{i1}, \ldots, y_{i,p_i})'$ arise from $Y(t)$ observed at

$$t_i = \{t_{i1} < \cdots < t_{i,p_i}\} \subset \mathcal{T} = [0,1]$$

$$\boxed{y(t_{ij}) = \sum_{k<j} \tilde{\phi}(t_{ij}, t_{ik})\, y(t_{ik}) + \epsilon(t_{ij}), \qquad \begin{array}{l} i = 1, \ldots, N \\ j = 2, \ldots, p_i, \end{array}}$$

where $\epsilon(t) \sim N\left(0, \sigma^2(t)\right)$. Transform $l = t - s, \quad m = \frac{t+s}{2}$, let

$$\phi(l, m) = \phi\left(t - s, \frac{1}{2}(s + t)\right) = \tilde{\phi}(t, s),$$

so that

$$-2\ell\left(\phi, \sigma^2 | Y_1, \ldots, Y_N\right) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \left[\log \sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2}\left(y_{ij} - \sum_{k<j} \tilde{\phi}(t_{ij}, t_{ik})\, y_{ik}\right)^2\right]$$

Hilbert spaces for univariate functions

An example: the cubic smoothing spline

The cubic smoothing spline corresponds to model space

$$\mathcal{H} = W_2[0,1] = \{f : [0,1] \to \mathbb{R} \,|\, f,\ f' \text{ abs cnts, } \int (f''(x))^2 dx < \infty\}.$$

and is the solution to the penalized least squares problem with

$$J(f) = \int_0^1 (f''(x))^2 dx.$$

Equip $\mathcal{H}$ with inner product such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $J(f) = 0$ for $f \in \mathcal{H}_0$.

- $\mathcal{H}$ has corresponding reproducing kernel $K = K_0 + K_1$.
- Take functions in null space $\mathcal{H}_0 = \{$linear or constant functions$\}$ to be null models.
- $J(f) = \|P_1 f\|^2$ corresponds to the squared norm of the projection of f to $\mathcal{H}_1$.

One-way functional ANOVA models

Given

1. averaging operator $A(f) = \int_0^1 f(x)dx$,
2. basis $k_1(\cdot)$ for linear functions in $\mathcal{H}_0$ such that $A(k_1) = 0$,

$\mathcal{H}$ admits decomposition

$$\begin{aligned}\mathcal{H} &= \{1\} \oplus \{k_1(\cdot)\} \oplus \mathcal{H}_1 \\ &= \text{mean} \oplus \begin{matrix}\text{parametric} \\ \text{main effect}\end{matrix} \oplus \begin{matrix}\text{nonparametric} \\ \text{main effect}\end{matrix}\end{aligned}$$

and representation

$$f(x) = c_0 + c_1 k_1(x) + f_1(x); \quad f \in \mathcal{H}$$

where $c_0, c_1 \in \mathbb{R}$ and $f_1 \in \mathcal{H}_1$.

A Two-way functional ANOVA Model for ϕ

via tensor product

Let
$$\phi \in \mathcal{H} = \mathcal{H}^{[l]} \otimes \mathcal{H}^{[m]}$$

constructed from building blocks

$$\mathcal{H}^{[l]} = \{1\} \oplus \{k_1(l)\} \oplus \mathcal{H}_1^{[l]}$$
$$\mathcal{H}^{[m]} = \{1\} \oplus \mathcal{H}_1^{[m]}.$$

with RK K. The subspaces of $\mathcal{H}$ define unique decomposition

$$\phi(l, m) = \mu + \phi_1(l) + \phi_2(m) + \phi_{12}(l, m).$$

	$\{1\}$	$\{k_1\}$	$\mathcal{H}_1^{[l]}$
$\{1\}$	mean	p-main effect	np-main effect
$\mathcal{H}_1^{[m]}$	np-main effect	$np \times p$-interaction	$np \times np$-interaction

Penalized log likelihood for ϕ

Define transformed pairs of observation times

$$\boldsymbol{v}_{ijk} = \left(t_{ij} - t_{ik}, \frac{1}{2}\left(t_{ij} + t_{ik}\right)\right) = \left(l_{ijk}, m_{ijk}\right),$$
$$V = \bigcup_{i,j,k} \left\{\boldsymbol{v}_{ijk}\right\} = \left\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{|V|}\right\}.$$

Fixing $\sigma_{ij}^2 = \sigma^2\left(t_{ij}\right)$, find ϕ minimizing

$$-2\ell\left(\phi | Y_1, \ldots, Y_N, \sigma^2\right) + \lambda J\left(\phi\right) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \frac{1}{\sigma_{ij}^2} \left(y_{ij} - \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y_{ik}\right)^2 + \lambda J\left(\phi\right) \tag{2}$$

where $J\left(\phi\right) = ||P_1 \phi||^2$ and $\boldsymbol{v}_{ijk} \in \mathcal{V} = [0,1]^2$,

A Representer Theorem

Theorem

Let $\{\nu_1, \ldots, \nu_{\mathcal{N}_0}\}$ span $\mathcal{H}_0$, the null space of $J(\phi) = ||P_1\phi||^2$. Let B denote the $|V| \times \mathcal{N}_0$ matrix having i^{th} column equal to ν_i evaluated at the observed $\boldsymbol{v} \in V$, and assume that B has full column rank. Then the minimizer ϕ_λ of (2) *is given by*

$$\phi_\lambda(\boldsymbol{v}) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(\boldsymbol{v}) + \sum_{j=1}^{|V|} c_j K_1(\boldsymbol{v}_j, \boldsymbol{v}), \tag{3}$$

where $K_1(\boldsymbol{v}_j, \boldsymbol{v})$ denotes the reproducing kernel for $\mathcal{H}_1$ evaluated at $\boldsymbol{v}_j$, the j^{th} element of V.

Obtaining the solution ϕ_λ

By the Representer Theorem, (2) becomes

$$-2\ell\left(c, d|\tilde{Y}, \tilde{B}, \tilde{K}_V\right) + \lambda J(\phi) = \left[\tilde{Y} - \tilde{B}d - \tilde{K}_V c\right]'\left[\tilde{Y} - \tilde{B}d - \tilde{K}_V c\right] + \lambda c' K_V c,$$

where

$$Y = \left(Y_1', Y_2', \ldots, Y_N'\right)' = \left(y_{12}, y_{13}, \ldots, y_{1p_1}, \ldots, y_{N2}, \ldots, y_{Np_N}\right)',$$
$$D = diag\left(\sigma_{12}^2, \sigma_{13}^2, \ldots, \sigma_{1p_1}^2, \ldots, \sigma_{N2}^2, \ldots, \sigma_{Np_N}^2\right)$$
$$X = \begin{bmatrix} X_1' & X_2' & \ldots & X_N' \end{bmatrix}'$$

$X_i = (p_i - 1) \times |V|$ matrix of AR covariates for Subject i,

$K_V = |V| \times |V|$ matrix with (i, j) element $K_1(\boldsymbol{v}_i, \boldsymbol{v}_j)$

$B = |V| \times \mathcal{N}_0$ matrix with (i, j) element $\nu_j(\boldsymbol{v}_i)$

and $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, $\tilde{K}_V = D^{-1/2}XK_V$.

Obtaining the solution ϕ_λ

Setting derivatives equal to zero, for fixed λ, c and d satisfy

$$\begin{bmatrix} \tilde{B}'\tilde{Y} \\ \tilde{K}_V'\tilde{Y} \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{K}_V \\ \tilde{K}_V'\tilde{B} & \tilde{K}_V'\tilde{K}_V + \lambda K_V \end{bmatrix}}_{C'C} \begin{bmatrix} d \\ c \end{bmatrix}$$

$$\Longrightarrow \begin{bmatrix} \hat{d} \\ \hat{c} \end{bmatrix} = C^{-1}(C')^{-1} \begin{bmatrix} \tilde{B}' \\ \tilde{K}_V' \end{bmatrix} \tilde{Y}.$$

The fitted values are given by $\widehat{\tilde{Y}} = \tilde{A}_{\lambda,\boldsymbol{\theta}}\tilde{Y}$, where the smoothing matrix is

$$\tilde{A}_{\lambda,\boldsymbol{\theta}} = \begin{bmatrix} \tilde{B} & \tilde{K}_V \end{bmatrix} C^{-1}(C')^{-1} \begin{bmatrix} \tilde{B}' \\ \tilde{K}_V' \end{bmatrix}$$

and can be used to compute model selection criteria.

A RKHS Framework for $\log \sigma^2$

For fixed $\phi(\cdot)$, let $Z_i = (z_{i1}, \ldots, z_{ip_i})'$, $z_{ij} = \epsilon_{ij}^2$, where

$$\epsilon_{ij} = y_{ij} - \sum_{k<j} \phi(\boldsymbol{v}_{ijk})\, y_{ik}.$$

The log likelihood of the squared working innovations $Z_1, \ldots, Z_N$ coincides with a Gamma distribution with scale parameter $\alpha = 2$:

$$-2\ell\left(\sigma^2 | Z_1, \ldots, Z_N\right) = \sum_{i=1}^{N}\sum_{j=1}^{p_i} \eta_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{p_i} z_{ij} e^{-\eta_{ij}},$$

where $\eta_{ij} = \eta(t_{ij}) = \log \sigma^2(t_{ij})$.

A RKHS Framework for $\log \sigma^2$

Take the estimator of $\eta(t) = \log \sigma^2(t) \in \mathcal{H}$ to minimize

$$-2\ell(\eta|Z_1, \dots, Z_N) + \lambda J(\eta) = \sum_{i=1}^{N} \sum_{j=1}^{p_i} \eta(t_{ij}) + \sum_{i=1}^{N} \sum_{j=1}^{p_i} z_{ij} e^{-\eta(t_{ij})} + \lambda J(\eta).$$

Let $\mathcal{T} = \bigcup_{i,j} \{t_{ij}\}$. By Theorem 1, the minimizer has the form

$$\eta_\lambda(t) = \sum_{i=1}^{\mathcal{N}_0} d_i \nu_i(t) + \sum_{j=1}^{|\mathcal{T}|} c_j K_1(t_j, t),$$

where $\{\nu_i\}$ span $\mathcal{H}_0$, $K_1(t_j, t)$ is the RK for $\mathcal{H}_1$ evalutated at t_j, the j^{th} element of $\mathcal{T}$.

Simulation Studies

1. Simulations with complete data varying
   - generating model Σ
   - number of subjects $N = 50, 100$
   - dimension $p = 10, 20, 30$

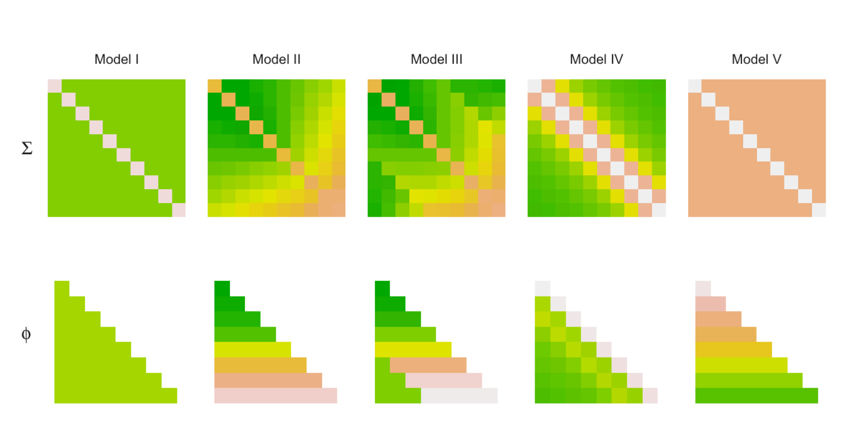
   to compare performance to the oracle estimator, Polynomial MCD GLM $\hat{\Sigma}_{poly}$, the sample covariance matrix $S = [s_{ij}]$, Shrinkage estimators S^{ω}, S^{λ}
2. Simulations with unbalanced data to study robustness to missing or irregular data
   - generating model Σ
   - dimension $p = 10, 20$
   - proportion of missing data 10%, 20%, 30%

Use Monte Carlo simulation to estimate risk $R\left(\Sigma, \hat{\Sigma}\right) = E_{\Sigma}\left[\Delta\left(\Sigma, \hat{\Sigma}\right)\right]$ under loss functions

$$\Delta_1\left(\Sigma, \hat{\Sigma}\right) = tr\left(\left(\Sigma^{-1}\hat{\Sigma} - \mathrm{I}\right)^2\right), \quad \Delta_2\left(\Sigma, \hat{\Sigma}\right) = tr\left(\Sigma^{-1}\hat{\Sigma}\right) - log|\Sigma^{-1}\hat{\Sigma}| - p$$

Simulation Studies

Data Generation Models

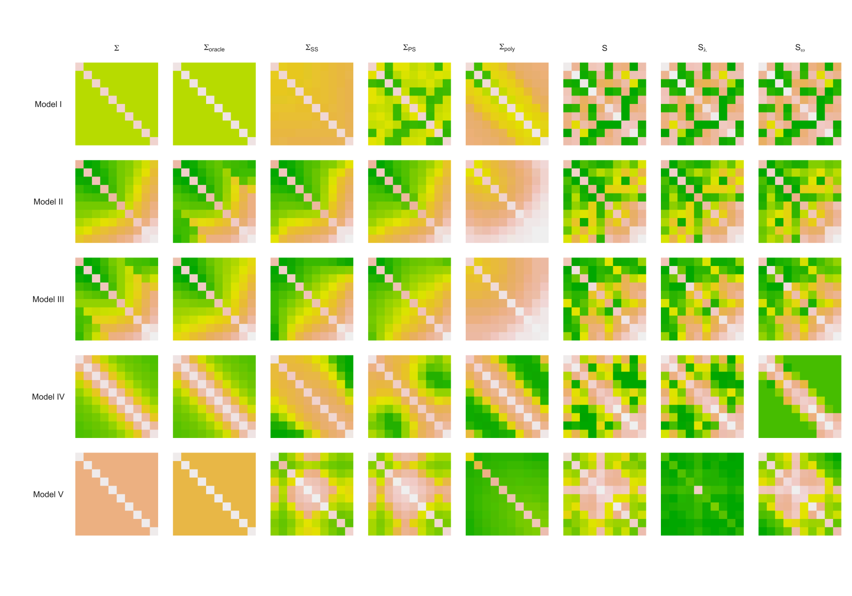


Simulation Studies

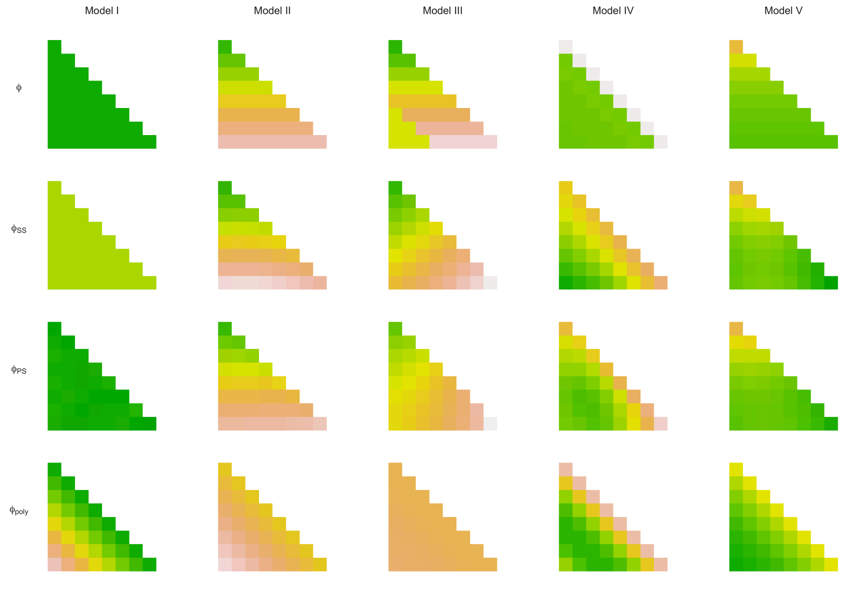
Data Generation Settings

I. $\Sigma = \mathrm{I}$	$\phi(t,s) = 0, 0 \leq s < t \leq 1$	$\sigma^2(t) = 1, 0 \leq t \leq 1$
II. $\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = t - \frac{1}{2}, \quad 0 \leq t \leq 1$	$\sigma^2(t) = 0.1^2, 0 \leq t \leq 1$
III. $\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = \begin{cases} t - \frac{1}{2}, & t - s \leq 0.5 \\ 0, & t - s > 0.5 \end{cases}$	$\sigma^2(t) = 0.1^2, 0 \leq t \leq 1$
IV. $\Sigma = [\sigma_{ij}]$	$\sigma_{ij} = \left(1 + \frac{(t_i - t_j)^2}{2k^2}\right)^{-1}$	$k = 0.6, 0 < t_i, t_j < 1$
V. $\Sigma = \rho \mathrm{J} + (1-\rho)\mathrm{I}, \ \rho = 0.7$	$\phi(t,s) = \frac{\rho}{1+(t-2)\rho}, t = 2, \ldots, p$	$\sigma^2(t) = 1 - \frac{(\max(t,2)-2)\rho^2}{1+(\max(t,2)-2)\rho}$

Simulation Studies

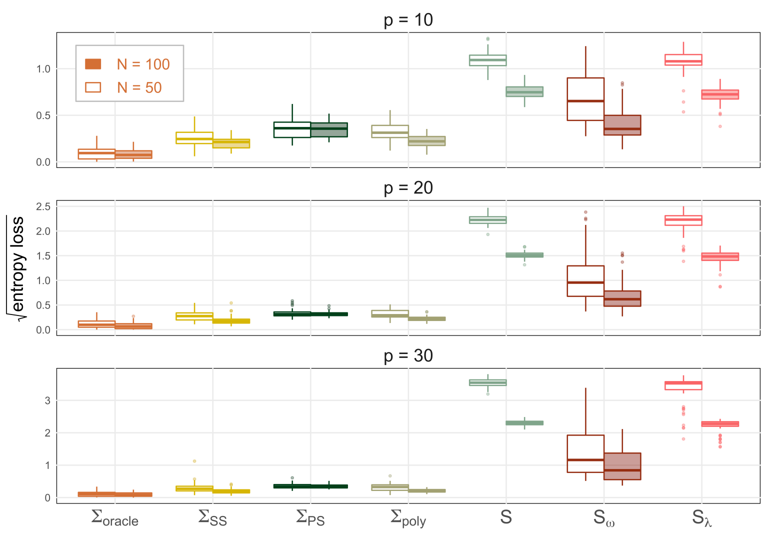


Simulation Studies



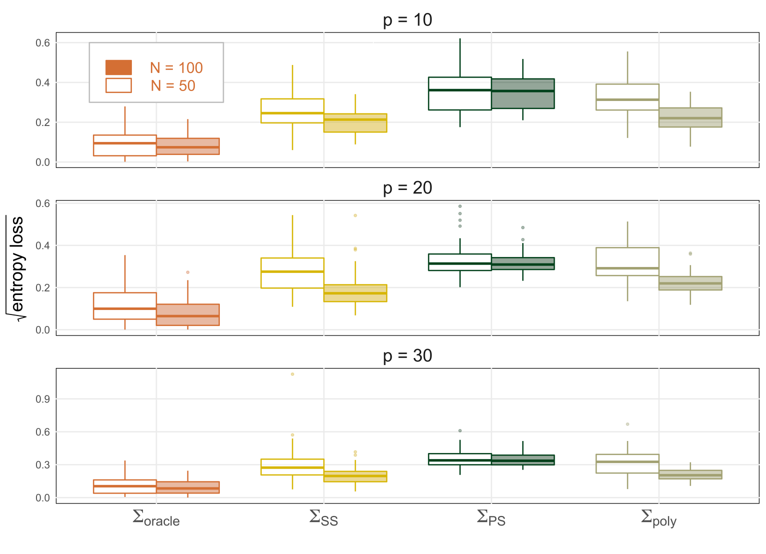
Simulation Studies

Results with complete data: Model I



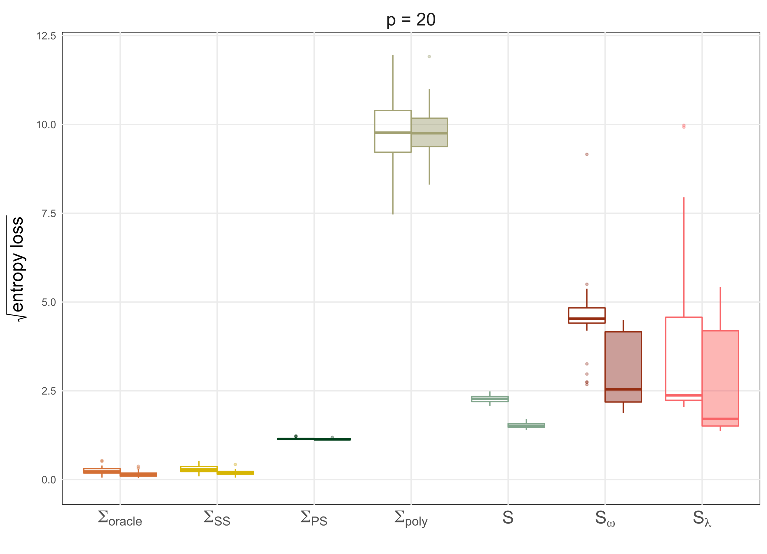
Simulation Studies

Results with complete data: Model I



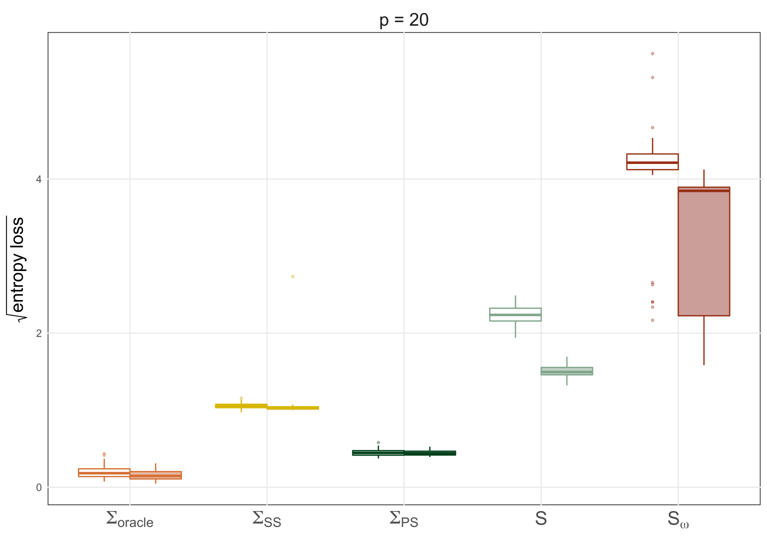
Simulation Studies

Results with complete data: Model II



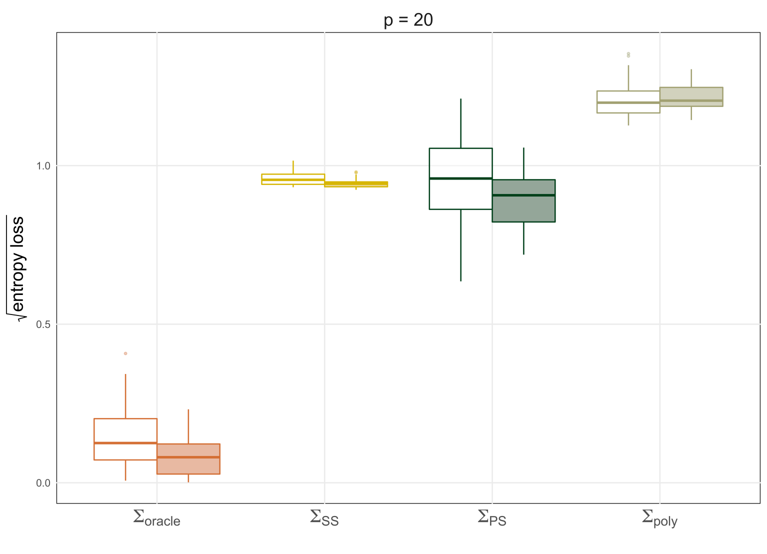
Simulation Studies

Results with complete data: Model III



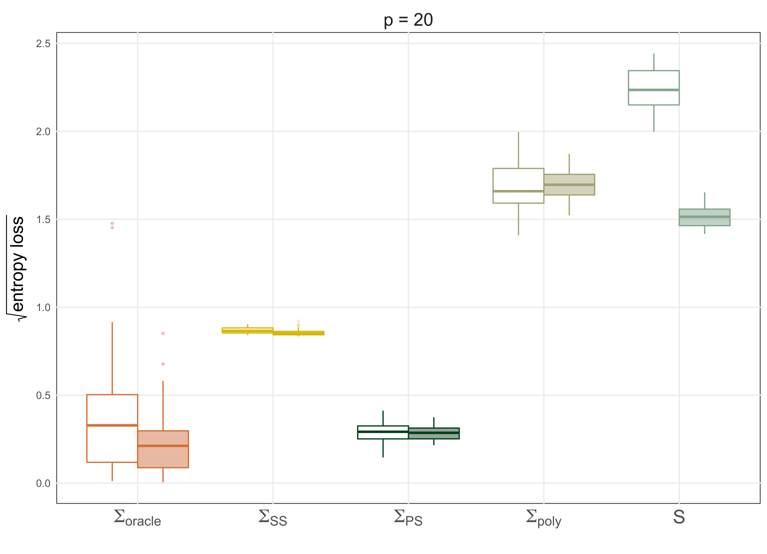
Simulation Studies

Results with complete data: Model IV



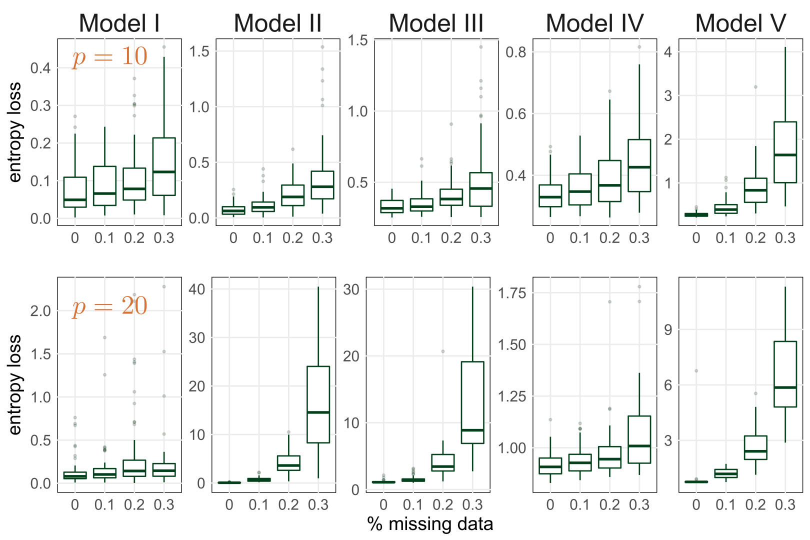
Simulation Studies

Results with complete data: Model V



Simulation Studies

Results with incomplete data, $N = 50$



Thank you!

For additional detail, writing, code, or questions, you can find me here:

tayler.a.blake@gmail.com
https://github.com/taylerablake/
taylerablake.github.io

Parametric Models for the Cholesky Decomposition

Pourahmadi (2000), Pan andMackenzie (2003) suggest modeling ϕ_{tj}, σ_t^2 with covariates, letting

$$\phi_{jk} = x_{jk}'\gamma$$
$$\log \sigma_j^2 = z_j'\lambda.$$

Common choices for the covariates x_{jk} and z_j are

$$x_{jk}' = \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{d-1}\right)',$$
$$z_j' = \left(1, t_j, \ldots, t_j^{q-1}\right)'.$$

Polynomial orders d and q are tuning parameters chosen by a model selection criterion (BIC, AIC).

Simulations

Applying Elementwise Shrinkage to S

Tapering Estimators

- **The Banded Sample Covariance Matrix**

$$B_k(S) = \left[s_{ij} 1\left(|i-j| \leq k\right)\right] = R_B * S, \quad 0 < k < p.$$

- **The Tapered Sample Covariance Matrix**

$$S^{\omega} = \left[\omega_{ij}^{k} s_{ij}\right],$$

where $0 < k < p$, and if $k_h = k/2$

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

- **Soft Thresholding Estimator:**

$$S^{\lambda} = \left[\mathrm{sign}\left(s_{ij}\right)\left(s_{ij} - \lambda\right)_+\right]$$

▸▸ Simulations