Nonparametric Covariance Estimation with Shrinkage Toward Stationary Models

Yoonkyung Lee*¹
Department of Statistics
The Ohio State University
*joint work with Tayler Blake

August 2-7, 2014
Joint Statistical Meetings, Boston, MA

¹Research supported in part by NSF
Covariance Estimation for Functional Data

- Data:
  “time-ordered” measurements from individual $i$
  $y_i = (y(t_{i1}), y(t_{i2}), \ldots, y(t_{i,n_i}))$ with $t_{i1} < t_{i2} < \cdots < t_{i,n_i}$
  for $i = 1, 2, \ldots, N$.

- Goal:
  covariance matrix of the random vector $y_i$ taken as
  realizations of $Y(t)$ at $t_i = (t_{i1}, \ldots, t_{i,n_i})$ with smooth covariance function, $\gamma(s, t) = \text{Cov}(Y(s), Y(t))$

- Inference, dimension reduction or feature extraction for
  effective representation of functional data, classification or clustering
Challenges

- High dimensionality for densely observed data ($n_i \gg N$) ⇒ regularization

- Sparse or irregular “time points” ($t_i \cap t_j \approx \emptyset$) ⇒ functional approach

- Positive definiteness of covariance ⇒ alternative unconstrained parameterization

- Simplicity in dependence structure (stationarity, short-term dependence, etc.) ⇒ banding, tapering, penalization
Unconstrained Parameterization

- Cholesky decomposition of the covariance matrix $\Sigma$ of a random vector $\mathbf{y} = (y_1, \ldots, y_n)$ (Pourahmadi 1999):
  \[
  T \Sigma T^\top = D,
  \]
  where $T$ is a unit lower triangular matrix and $D = \text{diag}(\sigma_j^2)$.

- Autoregressive model behind the decomposition:
  \[
  y_j = \sum_{k=1}^{j-1} \phi_{jk} y_k + \epsilon_j, \text{ for } j = 2, \ldots, n,
  \]
  where $T_{jk} = -\phi_{jk}$ for $j > k$, 1 for $j = k$, and 0 for $j < k$ and \(\text{var}(\epsilon_j) = \sigma_j^2\).

- Estimate the generalized autoregressive (GAR) parameters $\{\phi_{jk}\}$ and the innovation variances (IV) $\{\sigma_j^2\}$. 
Estimation of \( \{\phi_{jk}\} \) and \( \{\sigma_j^2\} \)

\[
y_j = \sum_{k=1}^{j-1} \phi_{jk} y_k + \epsilon_j, \text{ for } j = 2, \ldots, n
\]

Under the assumption that \( \epsilon_j \sim N(0, \sigma_j^2) \) independently, the negative log-likelihood is given by

\[
-2L(\{\phi_{jk}\}, \{\sigma_j^2\} | y) = \sum_{j=1}^{n} \log \sigma_j^2 + \sum_{j=2}^{n} \frac{(y_j - \sum_{k=1}^{j-1} \phi_{jk} y_k)^2}{\sigma_j^2}
\]

A two-stage estimation procedure can be developed by alternating estimation of \( \{\phi_{jk}\} \) and \( \{\log(\sigma_j^2)\} \).
GAR Parameters and IV as a Function

- View $\{\phi_{jk}\}$ and $\{\sigma^2_j\}$ as values of continuous GAR coefficient function $\phi(s, t)$ and IV function $\sigma^2(t)$ at observed time points:

  $$\phi_{jk} = \phi(t_j, t_k) \quad \text{for} \quad t_j > t_k$$
  $$\sigma^2_j = \sigma^2(t_j)$$

- Take covariance estimation as smoothing:

  $$\gamma(s, t) \Rightarrow \phi(s, t) \text{ and } \sigma^2(t)$$

- $T$ or $\phi(s, t)$ characterizes the dependence structure. e.g. $T$: a Toeplitz matrix $\Rightarrow$ stationary
Parsimonious Models for GAR Coefficients

- Truncate \( \{ \phi_{jk} \} \) at certain time lag:
  e.g. Levina et al. (2008)

- Model \( \{ \phi_{jk} \} \) as low-order polynomials of time differences \( (t_j - t_k) \):
  e.g. Pourahmadi (1999), Pan and Mackenzie (2003)

- Smooth down the sub-diagonals of \( T \):
  e.g. Wu and Pourahmadi (2003), Huang et al. (2007)
Reparameterization of $\phi$

- To shrink $\phi(s, t)$ toward stationary models that depend on time lag only, consider transformation of the pair of time points $(s, t)$ for $s > t$:

$$\ell = s - t \quad \text{and} \quad m = \frac{s + t}{2}$$

(lag) (additive direction)

- Re-express $\phi$ in terms of the new arguments $\ell$ and $m$:

$$\phi^*(\ell, m) = \phi^*(s - t, \frac{s + t}{2}) = \phi(s, t)$$
Functional ANOVA Model for $\phi^*$

- Model $\phi^*$ in a structured function space:

$$\phi^*(\ell, m) = \mu + \phi^*_1(\ell) + \phi^*_2(m) + \phi^*_{12}(\ell, m)$$

mean + main effects + interaction

- Smoothing spline ANOVA models (Gu, 2002):

$$\phi^* \in \mathcal{H}_\ell \otimes \mathcal{H}_m = \{1\} \oplus \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_{12}$$

e.g. Take the 2nd-order Sobolev space for $\mathcal{H}_\ell$ and $\mathcal{H}_m$ and decompose $W_2(0, 1) = \{1\} \oplus W^*_2$ via the averaging operator $A(f) = \int f(x)dx$. 
Penalization Approach

Find a function by minimizing

\[ \text{Lack-of-fit criterion for } f + \lambda \cdot J(f) \]

The penalty \( J(f) \) pulls the solution toward the null models with \( J(f) = 0 \).

\[ J(f) = \int (f'')^2 \, dx = \| P_1(f) \|^2 \]
yields a linear function as the null model in smoothing splines (Wahba, 1990).

Null models for \( \phi^* \)?
Parsimonious Models for $\phi^*$

$$\phi^*(\ell, m) = \mu + \phi_1^*(\ell) + \phi_2^*(m) + \phi_{12}^*(\ell, m)$$

- Independence:
  $$\phi^*(\ell, m) = 0$$

- Stationarity:
  $$\phi_2^*(m) = 0 \text{ and } \phi_{12}^*(\ell, m) = 0$$

- Short-term dependence:
  $$\mu + \phi_1^*(\ell) = 0 \text{ for } \ell \geq \ell_0$$

- Diminishing dependence:
  $$\phi_1^* \searrow \text{ in } \ell$$
Penalty

- Independence: $\phi^*(\ell, m) = 0$
  \[ J_I(\phi^*) = \|\phi^*(\ell, m)\|^2 \]

- Stationarity: $\phi^*_2(m) = 0$ and $\phi^*_{12}(\ell, m) = 0$
  \[ J_S(\phi^*) = \|\phi^*_2(m)\|^2 + \|\phi^*_{12}(\ell, m)\|^2 \]

- Short-term dependence: $\mu + \phi^*_1(\ell) = 0$ for $\ell \geq \ell_0$
  \[ J_{\ell_0}(\phi^*) = \int_{\ell \geq \ell_0} [\mu + \phi^*_1(\ell)]^2 d\ell \]

- Diminishing dependence: $\phi^*_1 \downarrow$ in $\ell$
  \[ J_d(\phi^*) = \int [\phi^*_1'(\ell)]_+ d\ell \]
Focusing on estimation of $\phi^*$ given $\sigma^2$,

$$
\min_{\phi^* \in \mathcal{H}} \sum_{i=1}^{N} -2L(\phi^*, \sigma^2 | y_i) + \lambda \cdot J(\phi^*)
$$

For classical $J(\phi^*)$ based on the norms of functional components (e.g. $J_I$ and $J_S$), $\hat{\phi}^*$ admits a finite dimensional representation by the representer theorem:

$$
\hat{\phi}^*(\ell, m) = \sum_k d_k b_k(\ell, m) + \sum_{i,j} c_{ij} B_{ij}(\ell, m),
$$

where $\{b_k\}$ and $\{B_{ij}\}$ are the basis functions for $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$.

Optimization over $\{d_k\}$ and $\{c_{ij}\}$
Additional Non-Norm Penalty

Given the representation of $\hat{\phi}^*(\ell, m)$, consider a sample version of penalty functional:

- In place of $J_{l_0}(\phi^*) = \int_{\ell \geq \ell_0} [\mu + \phi_1^*(\ell)]^2 d\ell$
  for short-term dependence,

  $J^n_{l_0}(\hat{\phi}^*) = \sum_{\ell_i \geq \ell_0} [\mu + \hat{\phi}_1^*(\ell_i)]^2$

- In place of $J_d(\phi^*) = \int [\phi_1^*(\ell')]_+ d\ell$
  for diminishing dependence,

  $J^n_d(\hat{\phi}^*) = \sum_i \left[ \frac{\hat{\phi}_1^*(\ell_{i+1}) - \hat{\phi}_1^*(\ell_i)}{\ell_{i+1} - \ell_i} \right]_+$

  as in nearly-isotonic regression (Tibshirani et al. 2011)
Alternative Basis Functions?

- **Right truncated power basis** for desirable monotonicity and support:

\[
(x - x_0)_-^k = \begin{cases} 
(x_0 - x)^k & \text{for } x \leq x_0 \\
0 & \text{for } x > x_0 
\end{cases}
\]

in contrast with \((x - x_0)_+^k\)

- For instance, \(\mathcal{F}_\ell = \text{span}(\{\ell^k\}_{k=0}^3, \{(\ell - \ell_i)_-^3\}_{i=1}^n)\) and take \(\phi^* \in \mathcal{F}_\ell \otimes \mathcal{F}_m\).

- Various penalties in linear regression (e.g. lasso, grouped lasso) can be combined with this alternative basis expansion.
Comments and Conclusions

- Propose regularization framework with novel covariance penalties for shrinkage toward stationary or short-term dependence models.

- Coupling of the form of penalty with dependence structure would be the key to successful applications.

- Implementational details need to be worked out. (e.g. choice of penalty parameters)
References

Chong Gu.
*Smoothing Spline ANOVA Models.*

Jianhua Z Huang, Linxu Liu, and Naiping Liu.
Estimation of large covariance matrices of longitudinal data with basis function approximations.

Elizaveta Levina, Adam Rothman, and Ji Zhu.
Sparse estimation of large covariance matrices via a nested lasso penalty.

Mohsen Pourahmadi.
Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation.

Ryan J. Tibshirani, Holger Hoeffling, and Robert Tibshirani.
Nearly-isotonic regression.

Grace Wahba.
*Spline Models for Observational Data.*

Wei Biao Wu and Mohsen Pourahmadi.
Nonparametric estimation of large covariance matrices of longitudinal data.