

**GLM For Covariance Matrices:
From A Long Series To Many Short Series**

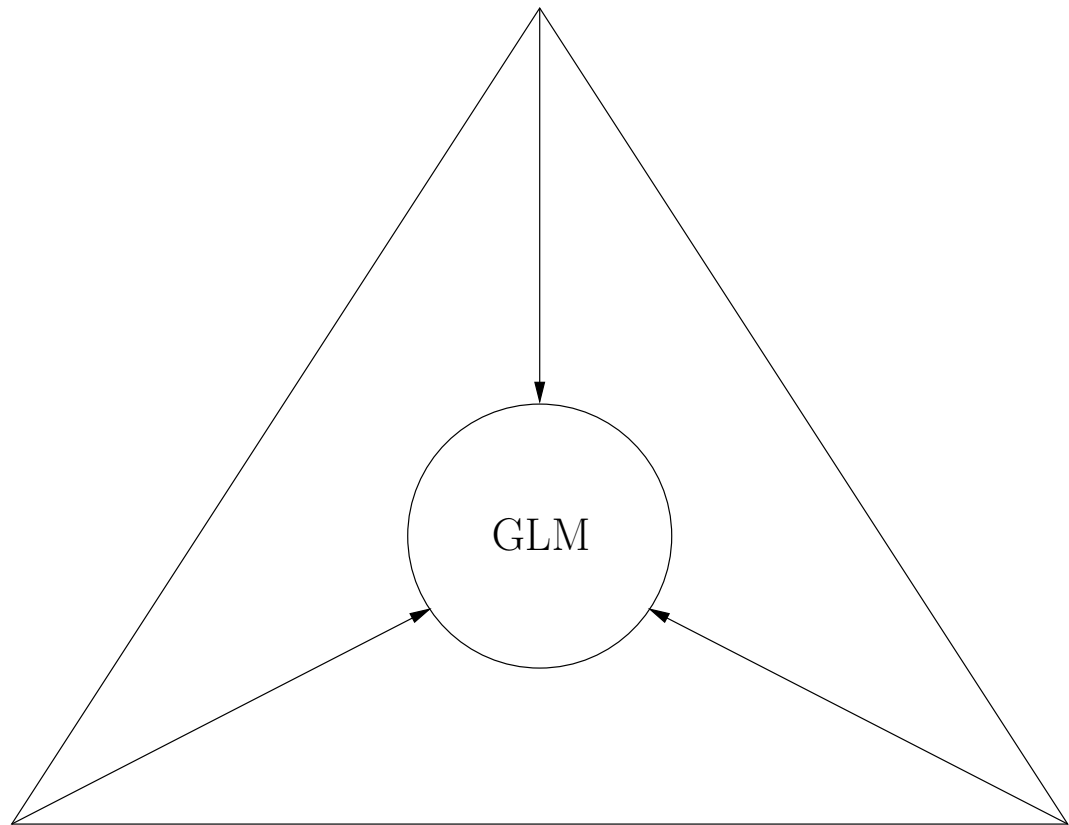
Mohsen Pourahmadi
Division of Statistics
Northern Illinois University

SAMSI
Opening Workshop and Tutorials
Program on High-Dimensional Inference
and Random Matrices

Sept. 17-20, 2006

- Covariance matrices have been studied for over a century in:

Time Series



Multivariate Statistics

Variance Components

- **Parsimonious** covariance models are needed for efficient estimation and inference in regression, for prediction, portfolio selection, assessing risk (ARCH-GARCH), \dots .
- The central issue is how to lift **the curses of dimensionality and positive-definiteness**.

I. One Long Series

- For a **stationary process** $\{X_t\}$, its cov. function (X_{t+k}, X_t) depends only on k :

$$\gamma_k = (X_{t+k}, X_t) = \int_{-\pi}^{\pi} e^{ik\theta} f(\theta) d\theta, \quad k = 0, \pm 1, \dots$$

Its covariance matrix

$$\Gamma = (\gamma_{k-\ell})_{k, \ell \in \mathbb{Z}}$$

is an infinite Toeplitz matrix.

- One usually observes a finite segment X_1, \dots, X_n of $\{X_t\}$ with the covariance matrix

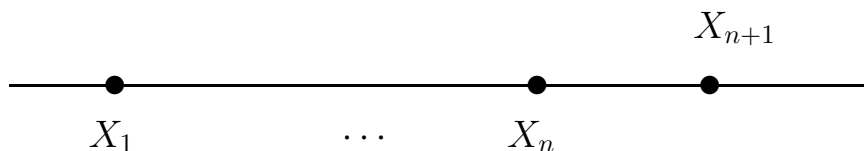
$$\Gamma_n = (\gamma_{k-\ell})_{k, \ell=1:n},$$

which is a finite subsection of Γ . **What happens as $n \rightarrow \infty$?**

Theorem 1 (Kac, Murdock & Szegő, 1953; Widom, 1958). Let $\lambda_{1,n}$ be **the smallest eigenvalue of Γ_n** . Then, under some smoothness conditions on f ,

$$\lambda_{1,n} \longrightarrow \min_{\theta} f(\theta).$$

- Let σ_n^2 be its **one-step ahead prediction error variance**:



then,

$$\lambda_{1,n} \leq \sigma_n^2 \leq \lambda_{n,n}.$$

Theorem 2 (Szegő, 1920). $\sigma_n^2 \rightarrow \exp(\int \log f)$.

- Past & Future

Gelfand & Yaglom (1957), Helson & Szegö (1960), Yaglom (1963), Grenander (1981), Jewell and Bloomfield (1983) studied the sequence of **canonical correlations** $1 \geq \rho_1 \geq \rho_2 \geq \dots \geq 0$ between the past & future:

$$\mathcal{P} = \{\dots, X_{-2}, X_{-1}\} \quad \text{and} \quad \mathcal{F} = \{X_0, X_1, \dots\}$$

of a stationary process with the spectral density

$$f = |\varphi|^2, \quad \varphi \in H^2 \text{ outer.}$$

- Grenander (1981), Jewell and Bloomfield (1983) pointed out the connection between the canonical correlations, the “eigenvalues” and the rank of the Hankel operator H_h on L^2 with the symbol $h = \varphi/\bar{\varphi}$.

Theorem 3. (a) (JB, 1983) Under some conditions, the canon. corr. of $\{X_t\} =$ Spectrum of H_h .

(b) (Kronecker, 1881; JB, 1983): The rank of H_h is finite if and only if f is a rational function or

$$\{X_t\} \text{ is an ARMA process.}$$

- For more details see Sec. 8.6 in

Pourahmadi (2001). Foundations of Time Series Analysis and Prediction Theory. Wiley.

V.V. Peller (2003). Hankel Operators and Their Applications. Springer.

A Century of Research Leading to GLM for Σ

	$\Sigma = (\sigma_{ij})$	$\Sigma^{-1} = (\sigma^{ij})$
Edgeworth (1892)		Parameterized $N(0, \Sigma)$ in terms of entries of the concentration matrix.
Slutsky (1927)	Banded*: Stationary MA(q)	
Yule (1927)		Banded*: Stationary AR(p), $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$.
Gabriel (1962)		Banded*: Nonstationary AR(p), Ante-dependence (AD) structure.
Dempster (1972)		Sparse: Certain $\sigma^{ij} = 0$. Σ^{-1} , the natural param. of MVN. Graphical Models.
Anderson (1973)	Linear	Linear Models
Leonard et al (1992,96)	Log-Linear	

GLM for Σ : Find a link function $g(\cdot)$ so that the entries of $g(\Sigma)$ are unconstrained and interpretable, then write $g(\Sigma) = X\beta$.

* For more recent use of the idea of “banding”, see Bickel and Levina (2004).

II. Many Short/Moderate Series

- Anderson's **Linear Covariance Model** (LCM):

$$\Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q,$$

where U_i 's are known symmetric matrices (covariates) and α_i 's are unknown **constrained** parameters so that Σ is positive-definite.

- **Every Σ has a representation as LCM:**

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \sigma_{11} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sigma_{22} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \sigma_{12} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

it is broad enough to include virtually all time series models, mixed models, factor models, multivariate GARCH models, ...

- A major drawback of LCM is the *constraint* on $\alpha = (\alpha_1, \dots, \alpha_q)$, which translates into the **root constraint** in time series, and **nonnegative variance/coefficients** in variance components, factor analysis, etc.
- LCM and many other techniques in the literature, essentially act **componentwise** on Σ , Diggle & Verbyla (1998); Yao, Müller and Wang (2005), ..., and cannot guarantee the positive-definiteness of Σ .

• **Leonard et al's Log-Linear Models (LLM):**

$$\log \Sigma = \alpha_1 U_1 + \cdots + \alpha_q U_q,$$

where U_i 's are as in LCM and α_i 's are unconstrained.

Q. How does one define $\log \Sigma$?

A. $\log \Sigma = A \Leftrightarrow \Sigma = e^A = I + \frac{A}{1!} + \frac{A^2}{2!} + \cdots .$

OR

– If $\Sigma = P' \Lambda P$, then $\log \Sigma = P' \log \Lambda P$.

Lemma: Σ is pd $\Leftrightarrow \log \Sigma$ is real and symmetric.

– A major drawback of LLM is the lack of *statistical interpretability* of the entries of $\log \Sigma$.

- **Time Series & Cholesky Decomposition:**

The AR(2) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t,$$

for $t = 1, 2, \dots, n$ can be written as

$$\begin{aligned} y_1 &= \phi_1 y_0 + \phi_2 y_{-1} + \varepsilon_1, \\ y_2 - \phi_1 y_1 &= \phi_2 y_0 + \varepsilon_2, \\ &\vdots \\ y_n - \phi_1 y_{n-1} - \phi_2 y_{n-2} &= \varepsilon_n. \end{aligned}$$

Setting $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ and $e = (y_{-1}, y_0)$, it becomes the **regression-like model**

$$TY = \varepsilon + Ce,$$

where

$$T = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -\phi_1 & 1 & 0 & \cdots & \cdots & 0 \\ -\phi_2 & -\phi_1 & 1 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\phi_2 & -\phi_1 & 1 \end{bmatrix}, C = \begin{bmatrix} \phi_2 & \phi_1 \\ 0 & \phi_2 \\ \cdots \\ 0 & \cdots & 0 \\ \vdots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} C_1 \\ \cdots \\ 0 \end{bmatrix}.$$

When ε and e are independent (causality assumption), it follows that

$$\begin{aligned} T_{\text{cov}(Y)}T' &= \sigma^2 I_n + \begin{pmatrix} C_1 \text{cov}(e)C_1' & 0 \\ 0 & 0 \end{pmatrix} \\ &= \text{A nearly } \mathbf{\textit{diagonal}} \text{ matrix.} \end{aligned}$$

- **Reg./G.-Schmidt/Chol./Szegö/Bartlett/DL/KF**
Regress y_t on its predecessors:

$$y_t = \phi_{t,t-1}y_{t-1} + \cdots + \phi_{t1}y_1 + \varepsilon_t,$$

y_1	y_2	y_3	\cdots	y_{n-1}	y_n
σ_1^2	σ_2^2	σ_3^2	\cdots	σ_{n-1}^2	σ_n^2
ϕ_{21}	ϕ_{32}	ϕ_{31}	\cdots	$\phi_{n,n-1}$	ϕ_{n1}
\vdots	\vdots	\vdots	\cdots	\vdots	\vdots
ϕ_{n1}	ϕ_{n2}	\cdots	\cdots	$\phi_{n,n-1}$	σ_n^2

in matrix form

$$\begin{bmatrix} 1 & & & & & \\ -\phi_{21} & 1 & & & & \\ -\phi_{31} & -\phi_{32} & 1 & & & \\ \vdots & & & \ddots & & \\ -\phi_{n1} & -\phi_{n2} & \cdots & -\phi_{n,n-1} & 1 & \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- ϕ_{tj} and $\log \sigma_t^2$ are unconstrained. Call them the **generalized autoregressive parameters** (GARP) and **innovation variances** (IV) of Y or Σ .
- This idea reduces the unintuitive task of covariance modeling to that of a sequence of regressions (with varying-order and varying-coefficients). (Pourahmadi, 1999, 2000).

- **Generalized Linear Models (GLM):**

For Σ pd, there are unique T and D with positive diagonal entries such that

$$T\Sigma T' = D.$$

Note. $\Sigma \longleftrightarrow (T, D)$.

Link functions: $g(\Sigma) = 2I - T - T' + \log D$,

a symmetric matrix with unconstrained and statistically meaningful entries.

Strategy: Model T “linearly” as in Anderson (1973)

$$\log D \quad \text{”} \quad \text{”} \quad \text{”} \quad \text{Leonard et al. (92,96).}$$

OR replace “linearly” by parametrically/nonparam./Bayesian
...

Bonus: The estimate $\hat{\Sigma} = \hat{T}^{-1} \hat{D} \hat{T}'^{-1}$ is always pd, where \hat{T} and \hat{D} are estimates of **parsimoniously** modeled T and D .

Q. How to identify models for (T, D) ?

A. Use covariates OR shrink to zero the smaller entries of T using penalized likelihood, priors, etc.. (Smith and Kohn, 2002; Pourahmadi and Daniels, 2002; Huang et al. 2006; Bickel and Levina, 2004, 2006).

Pearson



$$P\Sigma P' = \Lambda$$

Edgeworth

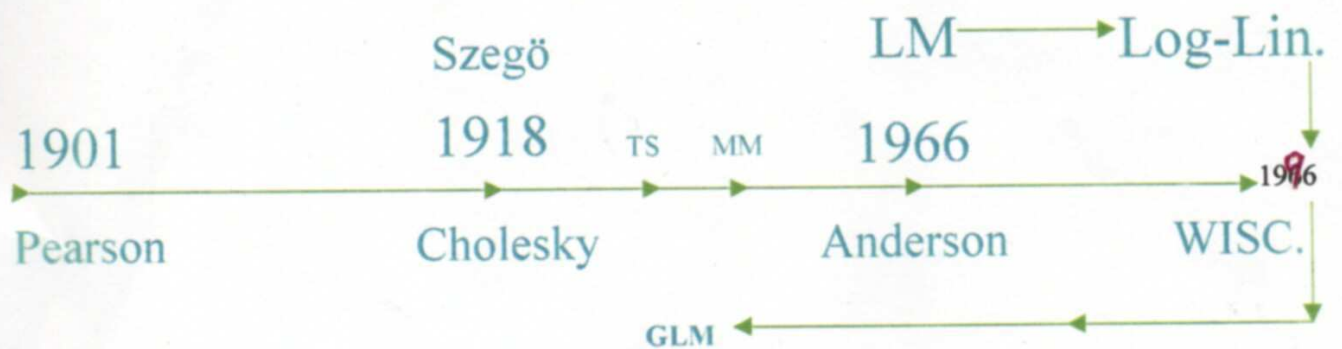


$$\Sigma^{-1}$$

Yule



AR(2), Correlogram,
Odds ratio



$$\text{GEE: } D' \Sigma^{-1} (Y - \mu) = 0$$



LS



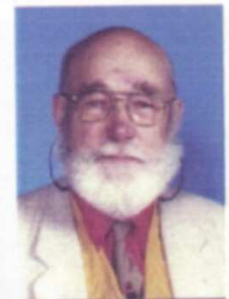
Legendre



Gauss



Galton



Nelde

Wedderburn