The Rocky Road to Defining Dependence and Modeling Covariance Matrices

Mohsen Pourahmadi
Department of Statistics
Texas A&M University

Sheldon B. Lubar School of Business
University of Wisconsin-Milwaukee
April, 19, 2013

Covariance and Correlation Matrices

are used in

- Regression and multivariate analysis
- Data mining (clustering, classification, ...).
- Time series, longitudinal and panel data.
- Spatial data (climate, environment, fMRI, ...).
- Business, economics and social sciences.
- Finance (portfolio selection, risk assessment, ...)

Covariance and Correlation Matrices

are used in

- Regression and multivariate analysis
- Data mining (clustering, classification, ...).
- Time series, longitudinal and panel data.
- Spatial data (climate, environment, fMRI, ...).
- Business, economics and social sciences.
- Finance (portfolio selection, risk assessment, ...)

Finding an **unconstrained** and statistically **interpretable** reparam. of the cov. matrix is still an open problem even for three variables.

Outline

1. Measures of Dependence: A Brief Review.

2. Simulating Random Correlation Matrices.

3. Obstacles in Modeling Covariance Matrices.
   - positive-definiteness constraint,
   - high-dimensionality.

4. Modeling a Cov. Matrix Like a Mean Vector (Reg.)

Outline

1. Measures of Dependence: A Brief Review.

2. Simulating Random Correlation Matrices.

3. Obstacles in Modeling Covariance Matrices.
   - positive-definiteness constraint,
   - high-dimensionality.

4. Modeling a Cov. Matrix Like a Mean Vector (Reg.) Linear and Log-Lin. Models.

Outline

1. Measures of Dependence: A Brief Review.

2. Simulating Random Correlation Matrices.

3. Obstacles in Modeling Covariance Matrices.
   - positive-definiteness constraint,
   - high-dimensionality.

4. Modeling a Cov. Matrix Like a Mean Vector (Reg.) Linear and Log-Lin. Models.

5. The **Regression Concept** and GLM.

6. Summary.

1. Galton's Regression

1. How does talent run in families?
   How and why is it that talent or quality once it occurred tended to dissipate rather than grow?
   Hereditary Genius (1869)

1. Galton's Regression

1. How does talent run in families?
   How and why is it that talent or quality once it occurred tended to dissipate rather than grow?
   Hereditary Genius (1869)

2. Realized rather early the difficulty in measuring intellectual quality (intelligence).

1. Galton's Regression

1. How does talent run in families?
   How and why is it that talent or quality once it occurred tended to dissipate rather than grow?
   Hereditary Genius (1869)

2. Realized rather early the difficulty in measuring intellectual quality (intelligence).

3. Measured successive generations of the diameter of sweat peas seeds.
4. Statures of parents and offsprings.

1. Relationship between height of fathers (X) and sons (Y):
2. Two regression lines:

$$E(Y|X = x) = \rho x,$$

$$\text{Var}(Y|X = x) = 1 - \rho^2.$$

when both standardized. The ρ was called the "co-relation index".

1. Relationship between height of fathers (X) and sons (Y):
2. Two regression lines:

$$E(Y|X = x) = \rho x,$$

$$\text{Var}(Y|X = x) = 1 - \rho^2.$$

   when both standardized. The ρ was called the "co-relation index".
3. (X, Y) Bivariate Normal (J.H. Dickson).

1. Relationship between height of fathers (X) and sons (Y):
2. Two regression lines:

$$E(Y|X = x) = \rho x,$$

$$\text{Var}(Y|X = x) = 1 - \rho^2.$$

   when both standardized. The ρ was called the "co-relation index".
3. (X, Y) Bivariate Normal (J.H. Dickson).
4. Invented the Quincunx machine to simulate data following the hereditary process.
   Natural Inheritance (1889)
5. Models linking measures of the current generation (Y_t) to the previous one (Y_{t-1}):

$$Y_t = \phi Y_{t-1} + \varepsilon_t.$$

Pearson's Correlation

1. Estimation of ρ was pursued by Pearson (1896) and Edgeworth (1892), leading to Pearson's product-moment estimator r which is the most popular measure of dependence for normal (Gaussian) data.

2. Stigler (1986, 1999).

Association in 2×2 Tables

1. M.H. Doolittle (1887);

2. Having given the number of instances respectively in which **things are thus and so**, in which **they are thus and not so**, in which **they are so and not thus**, and in which **they are neither thus nor so**, it is required to eliminate the general quantitative relativity inhering in the mere **thingness of the things**, and to determine the special quantitative relativity subsisting between the **thusness and the soness of the things**.

Association in 2×2 Tables

1. M.H. Doolittle (1887);

2. Having given the number of instances respectively in which **things are thus and so**, in which **they are thus and not so**, in which **they are so and not thus**, and in which **they are neither thus nor so**, it is required to eliminate the general quantitative relativity inhering in the mere **thingness of the things**, and to determine the special quantitative relativity subsisting between the **thusness and the soness of the things**.

3. J. Fourier:
   Mathematics has no symbols for confused ideas.

Pearson's Tetrachoric Correlation

For 2×2 contingency tables

1. For dichotomous data on two variables Pearson assumed an underlying bivariate normal distribution.

Pearson's Tetrachoric Correlation

For 2×2 contingency tables

1. For dichotomous data on two variables Pearson assumed an underlying bivariate normal distribution.

2. Yule (1871-1951) argued that categorical variables are inherently discrete. Defined the odds ratio (Yule's Q) directly using cell counts.

Pearson's Tetrachoric Correlation

For 2×2 contingency tables

1. For dichotomous data on two variables Pearson assumed an underlying bivariate normal distribution.

2. Yule (1871-1951) argued that categorical variables are inherently discrete. Defined the odds ratio (Yule's Q) directly using cell counts.
3. Time-correlation problem: Introduced Correlogram, AR(2) capturing cycles in economics data.

Dependence Measures: Math Axioms

1. Renyi's (1959) Measures of Dependence, Copula, Distance Correlation, Maximal Information Coefficient,..., are hard to interpret.

Dependence Measures: Math Axioms

1. Renyi's (1959) Measures of Dependence, Copula, Distance Correlation, Maximal Information Coefficient,..., are hard to interpret.

2. Goodman & Kruskal (1979) dependence measures with probabilistic interpretations,

Dependence Measures: Interpretability

1. Reimherr & Nicolae (2013) put more emphasis on **interpretability** of dependence measures mimicking the R^2 in regression:
$$R^2 = \frac{SSR}{SST}.$$
2. Replace SSR by an *information link function*
$$I(X;Y),$$
capturing the information that a variable X contains about another variable Y in a given context.

Dependence Measures: Interpretability

1. Reimherr & Nicolae (2013) put more emphasis on **interpretability** of dependence measures mimicking the R^2 in regression:
$$R^2 = \frac{SSR}{SST}.$$
2. Replace SSR by an *information link function*
$$I(X;Y),$$
capturing the information that a variable X contains about another variable Y in a given context.

3. Examples: Prediction error variance; entropy; Fisher information in the context of missing data.

1. The dependence measure:

$$D(X;Y) = \frac{I(X;Y)}{I(Y;Y)} \in [0,1],$$

has the built-in interpretability as **a reduction or fraction of information**.

1. The dependence measure:

$$D(X;Y) = \frac{I(X;Y)}{I(Y;Y)} \in [0,1],$$

has the built-in interpretability as **a reduction or fraction of information**.

2. *Mutual information index*:

$$\delta^2 = 1 - e^{2M(X,Y)},$$

in Linfoot (1957),..., Ebrahimi, Jalali and Soofi (2013) satisfies many properties of the Reimherr and Nicloae's (2013) proposal.

2. Simulating a Random Correlation Matrix

1. Simulating values of a random variable is a well-understood subject.
2. How does one simulate from a random correlation matrix?
3. Simulating random or typical correlation matrices are important in various areas of statistics (Joe, 2006), operation research and engineering (Holmes, 1991), finance, and numerical analysis.
4. Simulating "typical" nonnormal ARMA models (Lewis and Gaver, 70's).

Simulating a Correlation Matrix

1. Observing the **positive-definiteness** and other constraints is the main obstacle.
2. Is the 2×2 matrix

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

with r randomly selected from $U(-1, 1)$, a *correlation (pd) matrix*?

Simulating a Correlation Matrix

1. Observing the **positive-definiteness** and other constraints is the main obstacle.
2. Is the 2×2 matrix
$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$
with r randomly selected from $U(-1, 1)$, a *correlation (pd) matrix*?
3. YES.

Simulating a Correlation Matrix

1. Is the 3×3 matrix

$$R = (r_{ij}),$$

with r_{ij} simulated independently from $U(-1, 1)$, a *correlation (pd) matrix*?

Simulating a Correlation Matrix

1. Is the 3×3 matrix
$$R = (r_{ij}),$$
with r_{ij} simulated independently from $U(-1, 1)$, a *correlation (pd) matrix*?
2. May be.

Simulating a Correlation Matrix

1. Is the 3×3 matrix
$$R = (r_{ij}),$$
with r_{ij} simulated independently from $U(-1, 1)$, a *correlation (pd) matrix*?
2. May be.
3. The chance of being pd is $\pi^2/16 \approx 61.7\%$.

Simulating a Correlation Matrix

1. Is the 3×3 matrix
$$R = (r_{ij}),$$
with r_{ij} simulated independently from $U(-1, 1)$, a *correlation (pd) matrix*?
2. May be.
3. The chance of being pd is $\pi^2/16 \approx 61.7\%$.
4. For a 4×4, the chance of being pd reduces to 18.3%.

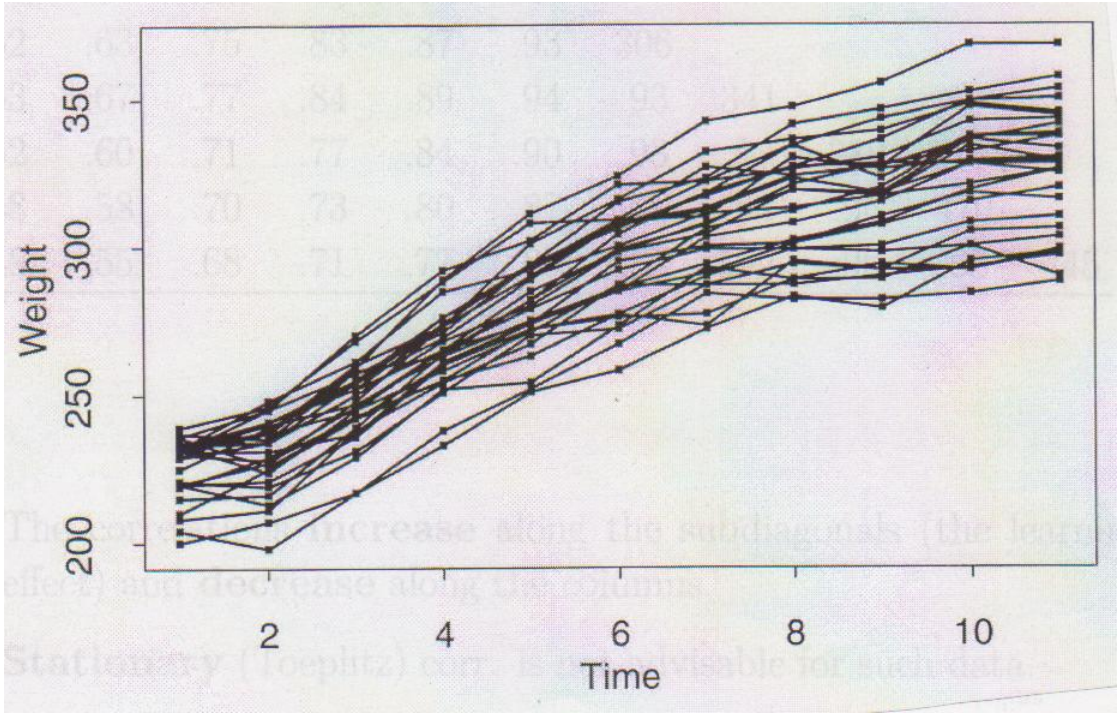
Simulating a Correlation Matrix

1. Is the 3×3 matrix
$$R = (r_{ij}),$$
with r_{ij} simulated independently from $U(-1, 1)$, a *correlation (pd) matrix*?
2. May be.
3. The chance of being pd is $\pi^2/16 \approx 61.7\%$.
4. For a 4×4, the chance of being pd reduces to 18.3%.

5. Ignoring the positive-definiteness can be costly. Reparameterization in terms of the partial correlation is helpful (Joe, 2006; MP and Daniels, 2009).

3. Modeling Covariance Matrices: An Example

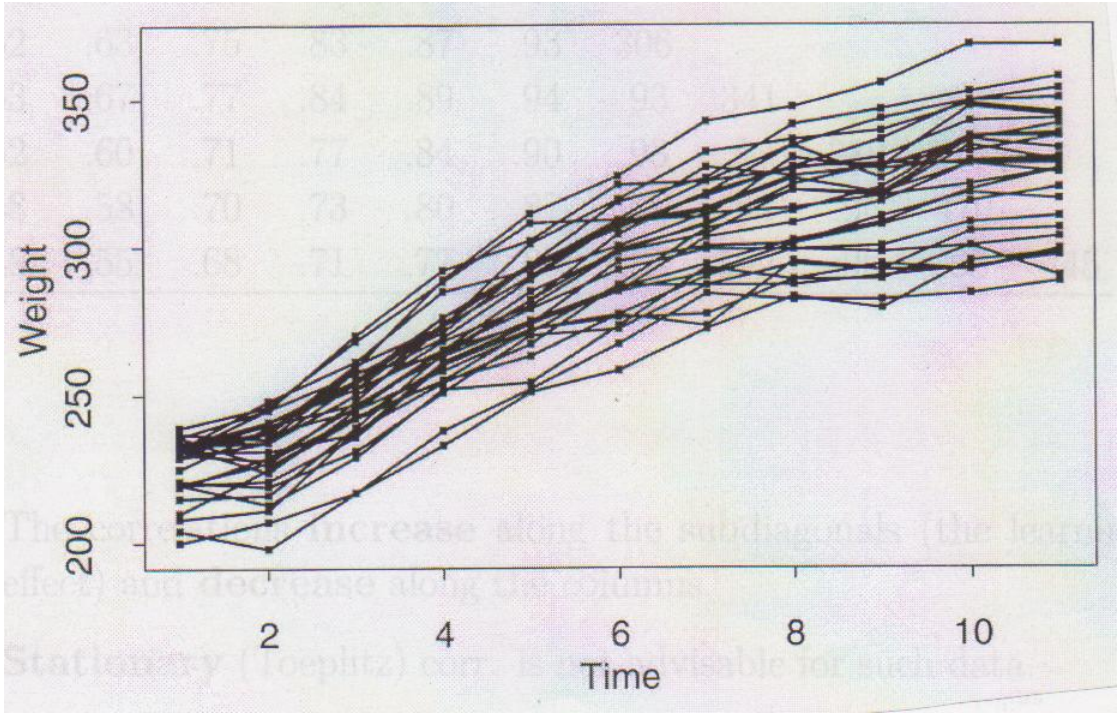
The cattle data (Kenward, 1987) was collected to study the effect of a treatment on intestinal parasites; $n = 30$ animals received a treat., they were weighed for $p = 11$ times.



- Clearly, means & variances **increase** over time,

3. Modeling Covariance Matrices: An Example

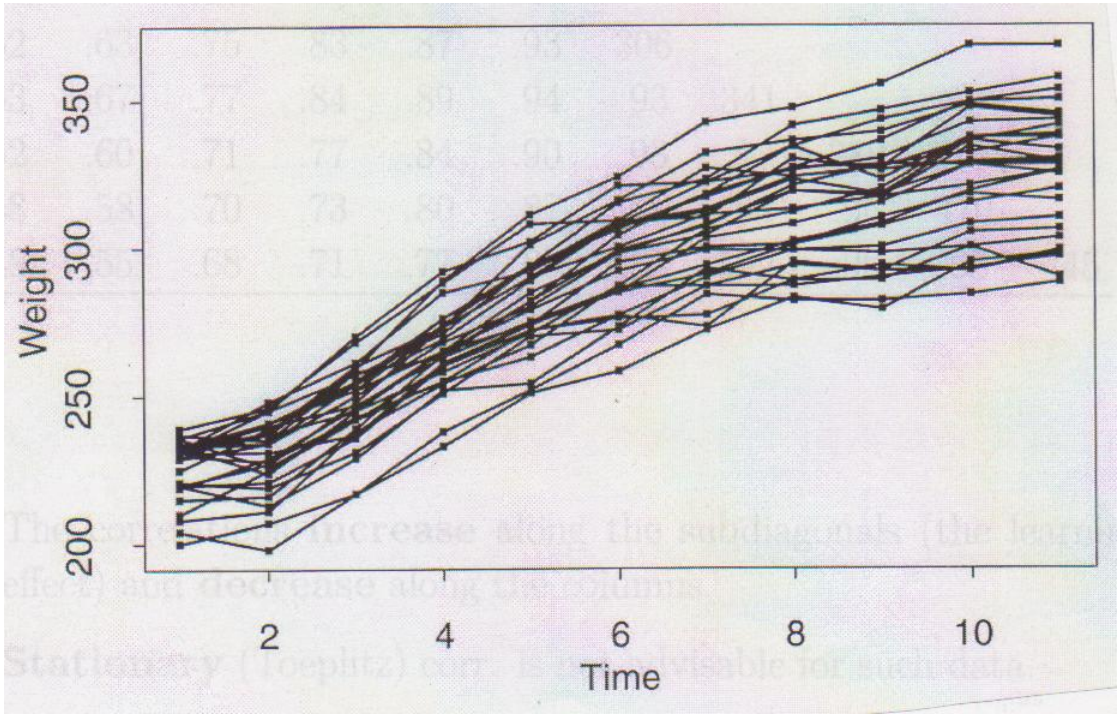
The cattle data (Kenward, 1987) was collected to study the effect of a treatment on intestinal parasites; $n = 30$ animals received a treat., they were weighed for $p = 11$ times.



- Clearly, means & variances **increase** over time,
- Any dependence (correlation) over time?

3. Modeling Covariance Matrices: An Example

The cattle data (Kenward, 1987) was collected to study the effect of a treatment on intestinal parasites; $n = 30$ animals received a treat., they were weighed for $p = 11$ times.



- Clearly, means & variances **increase** over time,
- Any dependence (correlation) over time?
- If so, is it stationary?

Sample variances on the diagonal, **correlations** off-diagonal:

106									
.82	155								
.76	.91	165							
.66	.84	.93	185						
.64	.80	.88	.94	243					
.59	.74	.85	.91	.94	284				
.52	.63	.75	.83	.87	.93	306			
.53	.67	.77	.84	.89	.94	.93	341		
.52	.60	.71	.77	.84	.90	.93	.97	389	
.48	.58	.70	.73	.80	.87	.88	.94	.96	470
.48	.55	.68	.71	.77	.83	.86	.92	.96	.98	445

- The correlations **increase** along the subdiagonals, **stationary** cov. matrix is not advisable.

Sample variances on the diagonal, **correlations** off-diagonal:

106										
.82	155									
.76	.91	165								
.66	.84	.93	185							
.64	.80	.88	.94	243						
.59	.74	.85	.91	.94	284					
.52	.63	.75	.83	.87	.93	306				
.53	.67	.77	.84	.89	.94	.93	341			
.52	.60	.71	.77	.84	.90	.93	.97	389		
.48	.58	.70	.73	.80	.87	.88	.94	.96	470	
.48	.55	.68	.71	.77	.83	.86	.92	.96	.98	445

- The correlations **increase** along the subdiagonals, **stationary** cov. matrix is not advisable.
- SAS PROC Mixed and other software packages provide a long menu of corr. structures. Choosing the suitable one is difficult even for the experts!

LCM: Linear Covariance Models (1892-1992)

- Set $\Sigma = \beta_1 U_1 + \cdots + \beta_q U_q$,
  where U_i's are known symmetric matrices (covariates) and β_i's are unknown scalar parameters so that Σ is positive-definite.
- The parameters are easy to **interpret**, but are **constrained**.

LCM: Linear Covariance Models (1892-1992)

- Set $\Sigma = \beta_1 U_1 + \cdots + \beta_q U_q$,
  where U_i's are known symmetric matrices (covariates) and β_i's are unknown scalar parameters so that Σ is positive-definite.
- The parameters are easy to **interpret**, but are **constrained**.

- LCM is broad enough to include virtually all time series models, mixed models, factor models,GARCH models, ...

LLM: Log-Linear Models (1992+)

Set

$$\log \Sigma = \beta_1 U_1 + \cdots + \beta_q U_q,$$

where U_i's are as in LCM and β_i's are unconstrained (Leonard & Hsu, 1992; Chiu, Leonard & Tsui, 1996).

LLM: Log-Linear Models (1992+)

Set

$$\log \Sigma = \beta_1 U_1 + \cdots + \beta_q U_q,$$

where U_i's are as in LCM and β_i's are unconstrained (Leonard & Hsu, 1992; Chiu, Leonard & Tsui, 1996).

A major drawback of LLM is the lack of **statistical interpretability** of entries of $\log \Sigma$ (Brown, Le & Zidek, 1994).

4. GLM: Cholesky Decompositions

- In most textbooks and software packages, the **Cholesky decomposition** of a pd matrix is introduced by

$$\Sigma = CC',$$

where C is a unique **lower triangular matrix** with positive diagonal entries (Cholesky, 1918; Bartlett, 1933).

- **Interpretation** of the entries of $C = (c_{ij})$ is difficult (Bates and Pinheiro, 1996).
- However, reducing C to unit lower triangular matrices makes the task of interpretation much easier.

A Time Series Motivation

$AR(2)$: $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t,\ 1 \leq t \leq p,$
can be written as

$$TY = \varepsilon + Ke,$$

A Time Series Motivation

$AR(2)\ :\ y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t,\ 1 \le t \le p,$
can be written as

$$TY = \varepsilon + Ke,$$

$$T = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -\phi_1 & 1 & 0 & \cdots & \cdots & 0 \\ -\phi_2 & -\phi_1 & 1 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\phi_2 & -\phi_1 & 1 \end{bmatrix}, \ T\mathrm{Cov}(Y)T' \approx D.$$

and $Y = (y_1, \ldots, y_p)', \varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)', e = (y_{-1}, y_0)'$ and K is a small matrix.

Regression/GS/ 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_{p-1}	y_p
σ_1^2					
ϕ_{21}	σ_2^2				
ϕ_{31}	ϕ_{32}	σ_3^2			
$\vdots$	$\vdots$		$\ddots$		
ϕ_{p1}	ϕ_{p2}	$\cdots$	$\cdots$	$\phi_{p,p-1}$	σ_p^2

In matrix form

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \cdots & -\phi_{p,p-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix},$$

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

- ϕ_{tj} and $log\sigma_t^2$ are **unconstrained** and called the **gen. autoregressive parameters** (GARP) and **innovation variances** (IV) of Y.
- Swap the constrained, high-dimensional parameter Σ with the unconstrained pair (T, D).

Modeling Strategy

- Write parametric (regression) models for T and $logD$.
- **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.
- This process reduces the unintuitive task of modeling a covariance matrix to that of a sequence of regressions. Pourahmadi (1999, 2000, 2007).

The Statistical Model Fitting Process

Is iterative (Box and Jenkins, 1970) and cycles through model

- **formulation**,
- **estimation**,
- **diagnostics**.

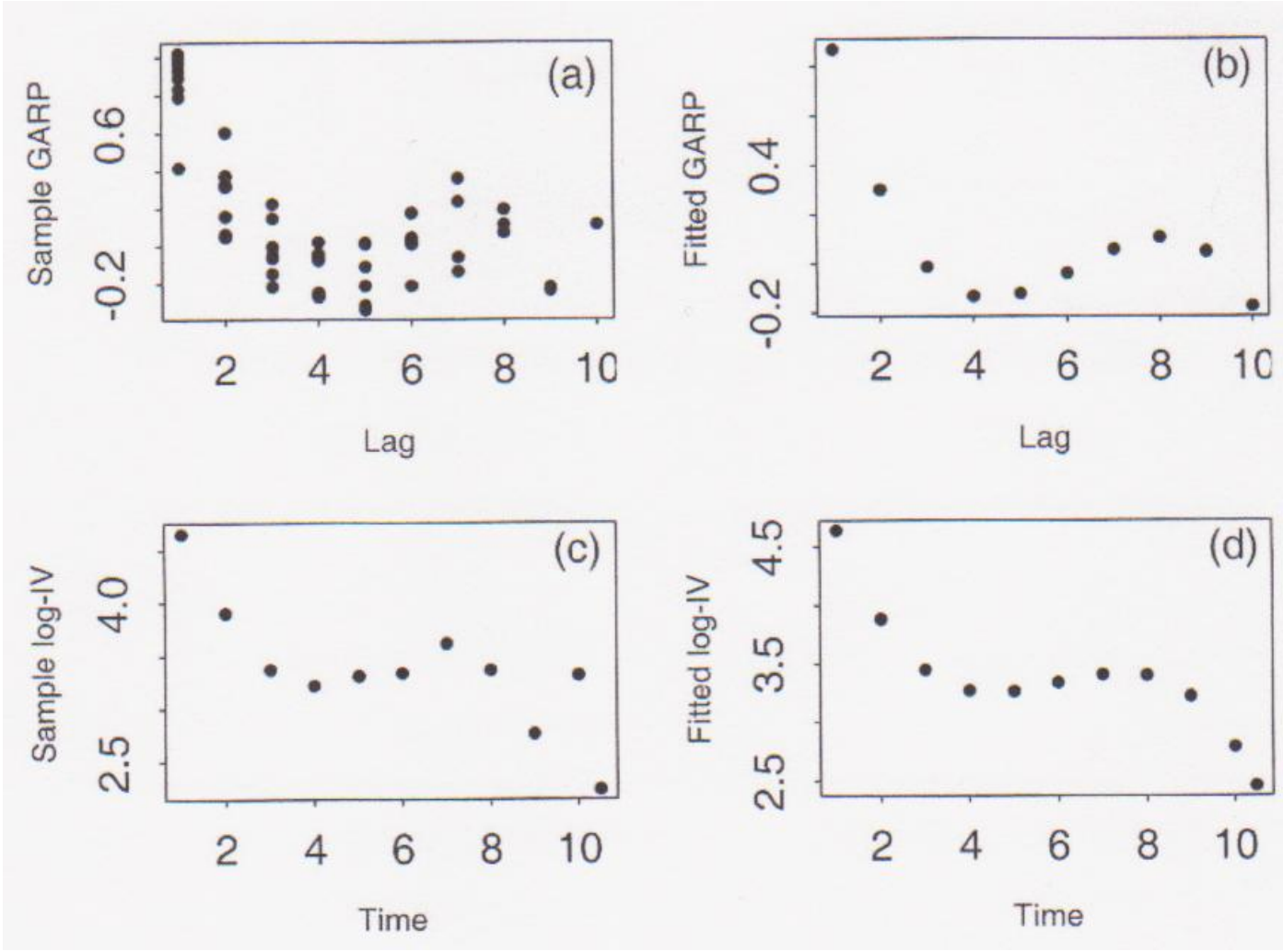
Cov. Model Formulation: Regressogram

The Regressogram* plays roles similar to the scatterplot in regression, and correlogram in time series.

It simply plots the GARP $\{\phi_{t,t+k}\}$ vs the lags $k = 1, 2, \cdots, t-1$, and $log\sigma_t^2$ vs the times $t = 1, 2, \cdots, p$.

*Tukey (1961).
Curves as parameters, and touch estimation. 4th Berkeley Symp., 681-694.

Regressogram of the Cattle Data



For the cattle data the regressogram suggests the following cubic models for the IVs and GARPs:

$$\begin{cases} \log \hat{\sigma}_t^2 = \lambda_1 + \lambda_2 t + \lambda_3 t^2 + \lambda_4 t^3 + \epsilon_{t,v}, \\ \\ \phi_{t,j} = \gamma_1 + \gamma_2 (t-j) + \gamma_3 (t-j)^2 + \gamma_4 (t-j)^3 + \epsilon_{t,d}. \end{cases}$$

In general, however, these and μ_t can be modeled **linearly** as

$$\mu_t = x_t' \beta, \log \sigma_t^2 = z_t' \lambda, \phi_{t,j} = z_{t,j}' \gamma,$$

where $x_t, z_t, z_{t,j}$ are $p \times 1, q \times 1$ and $d \times 1$ vectors of covariates, $\beta = (\beta_1, \cdots, \beta_p)', \lambda = (\lambda_1, \cdots, \lambda_q)'$ and $\gamma = (\gamma_1, \cdots, \gamma_d)'$ are parameters corresponding to the **means**, **innovation variances** and **correlations**.

Model Estimation: $\theta = (\beta', \lambda', \gamma')'$

The normal likelihood function has **three representations** corresponding to the three components of θ:

$$
\begin{aligned}
-2L(\theta) &= n \log |\Sigma| + \sum_{i=1}^{n} (Y_i - X_i\beta)' \Sigma^{-1} (Y_i - X_i\beta) \\
&= n \sum_{t=1}^{p} \log \sigma_t^2 + \sum_{t=1}^{p} \frac{RSS_t}{\sigma_t^2} \\
&= n \sum_{t=1}^{p} \log \sigma_t^2 + \sum_{i=1}^{n} \{r_i - Z(i)\gamma\}' D^{-1} \{r_i - Z(i)\gamma\},
\end{aligned}
$$

where $r_i = Y_i - X_i\ \beta, RSS_t$ and $Z(i)$ depend on r_i and other covariates and parameter values.

- Note that the log-likelihood function is **quadratic** in β and γ, but not in λ. The Newton-Raphson algorithm can be used to compute the MLE $\hat{\theta}$.

Asymptotic Distribution of the MLE

Theorem. (MP, 2000, 2007) For normal data and the usual regularity conditions on the design matrices of the three submodels,

(a) The MLE $\hat{\theta}$ is strongly consistent as $n \to \infty$,

(b) $\sqrt{n}(\hat{\theta} - \theta) \to N(0, I_\theta^{-1})$,

$$I_\theta = \text{ block diagonal } (I_\beta, I_\lambda, I_\gamma).$$

- Note the **parameter orthogonality** (Cox and Reid, 1987).

- For nonnormal data, see
  Ye and Pan (2006). Modelling covariance structures in **generalized estimating equations** for longitudinal data.

Cattle Data:
Poly(v,d)=Poly. of deg. v for $logD$ and d for T.

Model	L_{max}	No. of Parameters	BIC
Unstructured Σ	-1019.69	66	75.35
Poly (3,3)	-1049.01=L_1	8	70.84
Poly (3,2)	-1080.08=L_0	7	72.80
Poly (3,1)	-1131.61	6	76.09
Poly (3,0)	-121235	5	81.59
Poly (3)= Diag. Σ	-1377.43	4	92.28
Unstructured AD(2)	-1035.98	30	72.47
Structured AD(2)	-1054.13	8	71.18
Stationary AR(2)	-1062.89	3	71.20
Structured AD(2)	-1054.20	6	70.96
with $\lambda_1 = \lambda_2 = 1$	Zimmerman		

Model Checking

Likelihood Ratio Test of Poly (3,3) vs Poly (3,2):

$$2(L_1 - L_0) \sim \chi^2_1.$$

Since $2(L_1 - L_0) = 62.14$, the simpler Poly(3,2) model is rejected so $(t - j)^3$ is kept in the model.

Model Checking

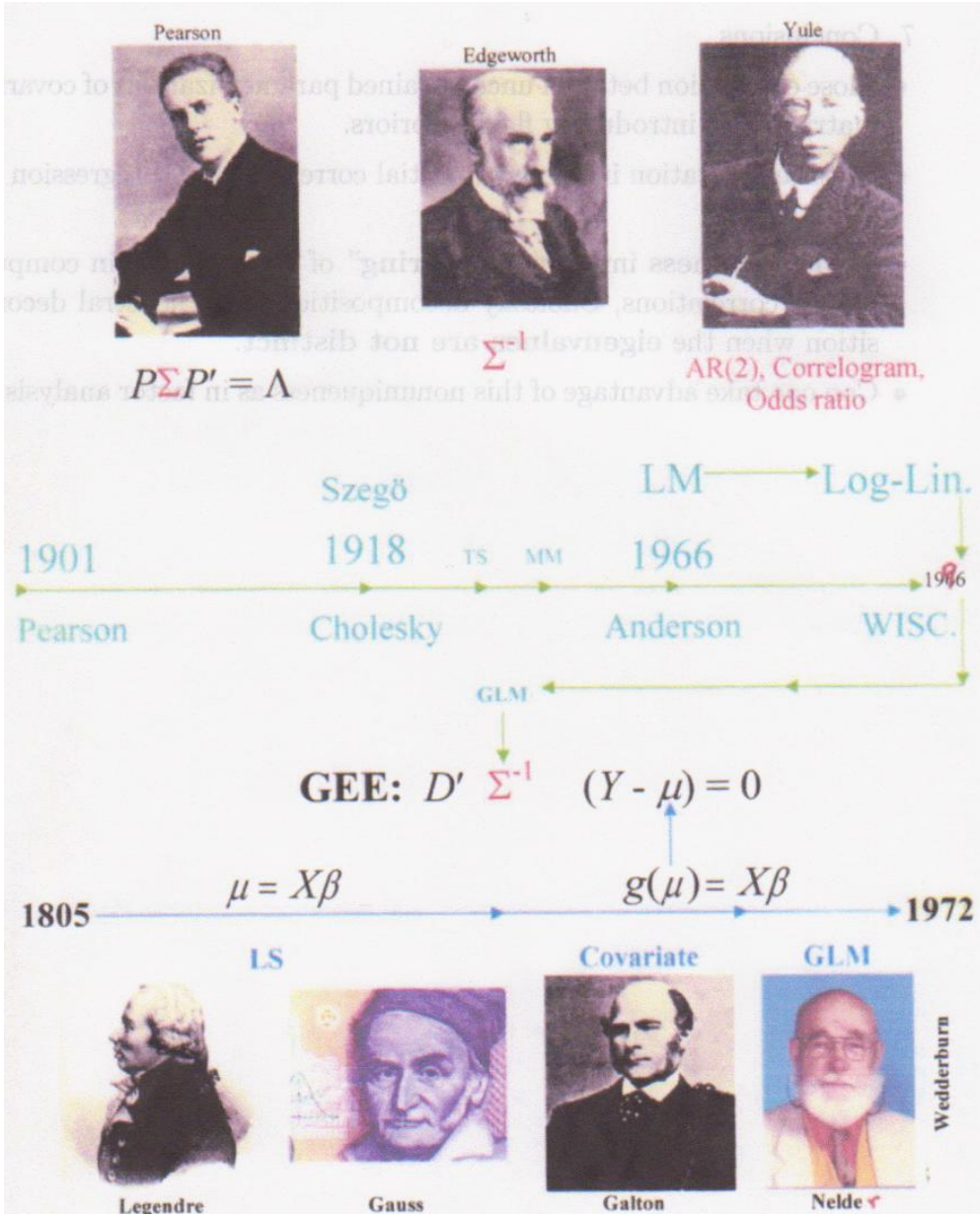
Likelihood Ratio Test of Poly (3,3) vs Poly (3,2):

$$2(L_1 - L_0) \sim \chi^2_1.$$

Since $2(L_1 - L_0) = 62.14$, the simpler Poly(3,2) model is rejected so $(t - j)^3$ is kept in the model.

T. Garcia, P. Kohli and MP (2012). Regressograms and mean-covariance models for incomplete longitudinal data. *The American Statistician*, 66, 85-91.

A Pictorial Summary



Coming Soon to a Bookstore Near You