Generalized Linear Models For Covariances : Curses of Dimensionality and PD-ness

Mohsen Pourahmadi
Division of Statistics
Northern Illinois University

MSU

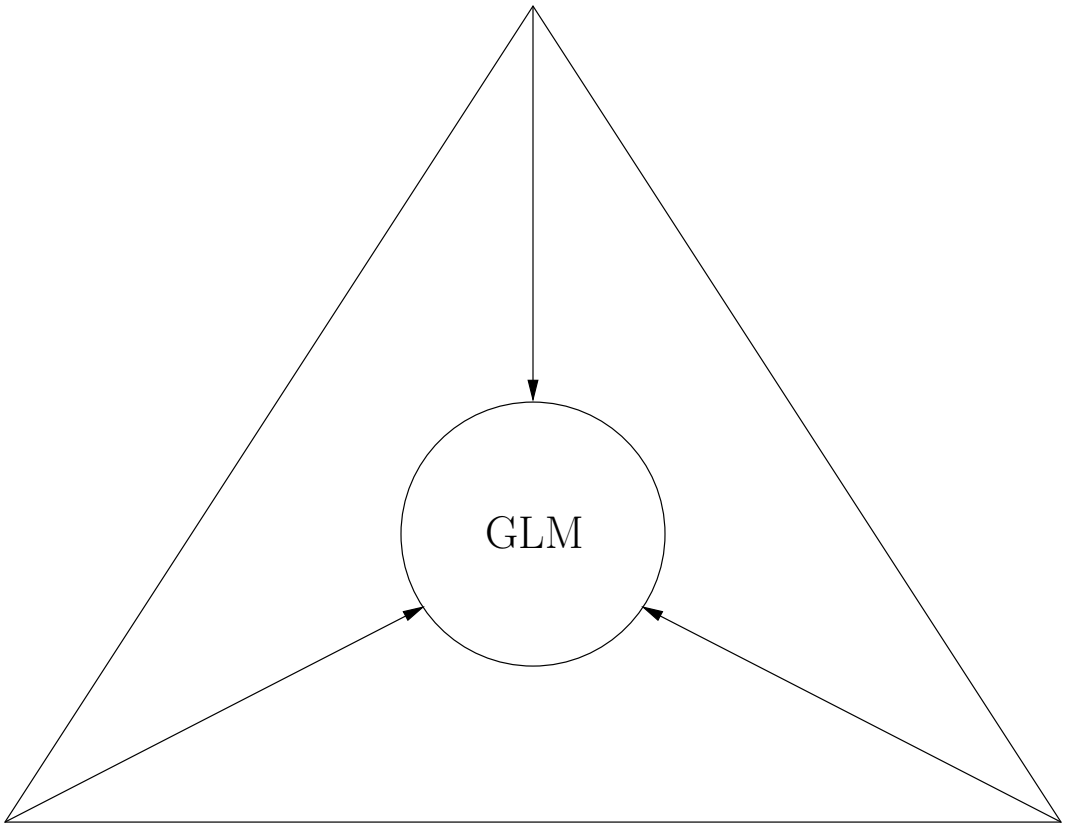
November 1, 2005

Outline

1. Prevalence of Covariance Modeling
2. The Shape of Correlated Data
3. The Cattle & Call Center Data; 11×11 & 102×102
4. Linear and Log-Linear Covariance Models
5. Time Series & Cholesky Decomposition
6. Generalized Linear Models (GLM)
   - Model Formulation (Regressogram)
   - Estimation and Diagnostics
   - Data Analysis
7. Bayesian, Nonparametric and Penalized Likelihood
8. Conclusions

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

Multivariate Statistics



Time Series **Variance Components**

• Parsimonious cov. is needed for efficient est. and inference in regression, for prediction, portfolio selection, assessing risk (ARCH-GARCH), $\cdots$.

- Nelder and Wedderburn's (1972) GLM unifies
  - normal linear regressions (Legendre, 1805; Gauss, 1809),
  - logistic (probit, ...) binary regressions, Poisson regressions, log-linear models for contingency tables,
  - variance component estimation using ANOVA sum of squares,
  - joint modelling of mean and dispersion (Nelder & Pregibon, 1987)
  - survival function (McCullagh & Nelder, 1989),
  - spectral density estimation in time series using periodogram ordinates (Cameron & Tanner, 1987),
  - generalized additive models (Hastie & Tibshirani, 1986); nonparametric methods,
  - hierarchical GLMs (Lee & Nelder, 1996),
  - Bayesian GLMs (Dey et al. 2000).

•• The Three Pillars of GLM Are:

I. Work with unconstrained (canonical) parameter,

II. Work with models that are additive in the covariates,

III. Rely on MLE / IRWLS .

History: Linear Covariance Model (LCM)

	$\Sigma = (\sigma_{ij})$	$\Sigma^{-1} = (\sigma^{ij})$
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), $y_t = \phi_{t1} y_{t-1} + \phi_{t2} y_{t-2} + \varepsilon_t$, **Ante-dependence** (AD) structure.
Dempster (1972)		**Sparse:** Certain $\sigma^{ij} = 0$. Σ^{-1}, the **natural param.** of MVN. Graphical Models. Matrix completion problem in LA.
Anderson (66, 69, 73)	**Linear**	**Linear Models**

- **Ideal Shape of Correlated Data:**

		Occasion					
		1	2	$\cdots$	t	$\cdots$	n
Unit	1	y_{11}	y_{12}	$\cdots$	y_{1t}	$\cdots$	y_{1n}
	2	y_{21}	y_{22}	$\cdots$	y_{2t}	$\cdots$	y_{2n}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	$(y_{i1}$	y_{i2}	$\cdots$	y_{it}	$\cdots$	$y_{in}) = Y_i$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	m	y_{m1}	y_{m2}	$\cdots$	y_{mt}	$\cdots$	y_{mn}

Special Cases in Increasing Order of Difficulty:

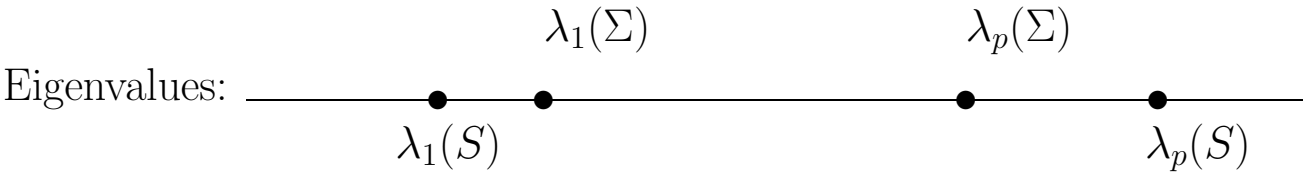
I. **Time Series Data:** $m = 1$, n large.
II. **Multivariate Data:** $m > 1$, n small to moderate; rows are indep.
Longitudinal Data, Panel Data, Cluster Data.
III. **Multiple Time Series:** $m > 1$, n large, rows are dependent.
IV. **Spatial Data:** m & n are hopefully large, rows are dependent.

- Many Short Time Series.
- "Time" - order is required for the Cholesky decomposition.

- **Improving The Sample Covariance Matrix**

Balanced Data: $Y_1, \ldots, Y_m$ are i.i.d. $N(\mu, \Sigma)$.

Sample Cov. Matrix: $S = \frac{1}{m} \sum_{i=1}^{m} (Y_i - \bar{Y})(Y_i - \bar{Y})'$.



Stein's Estimator: Shrinks the eigenvalues of S

Loss Functions: $L(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2$, if Σ invertible,

$L(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma} - \Sigma)^2$, Otherwise.

(Ledoit et al., 2000+): $\hat{\Sigma} = \alpha S + (1 - \alpha) I$, $0 \leq \alpha \leq 1$.

In some applications such as in finance and microarray data,

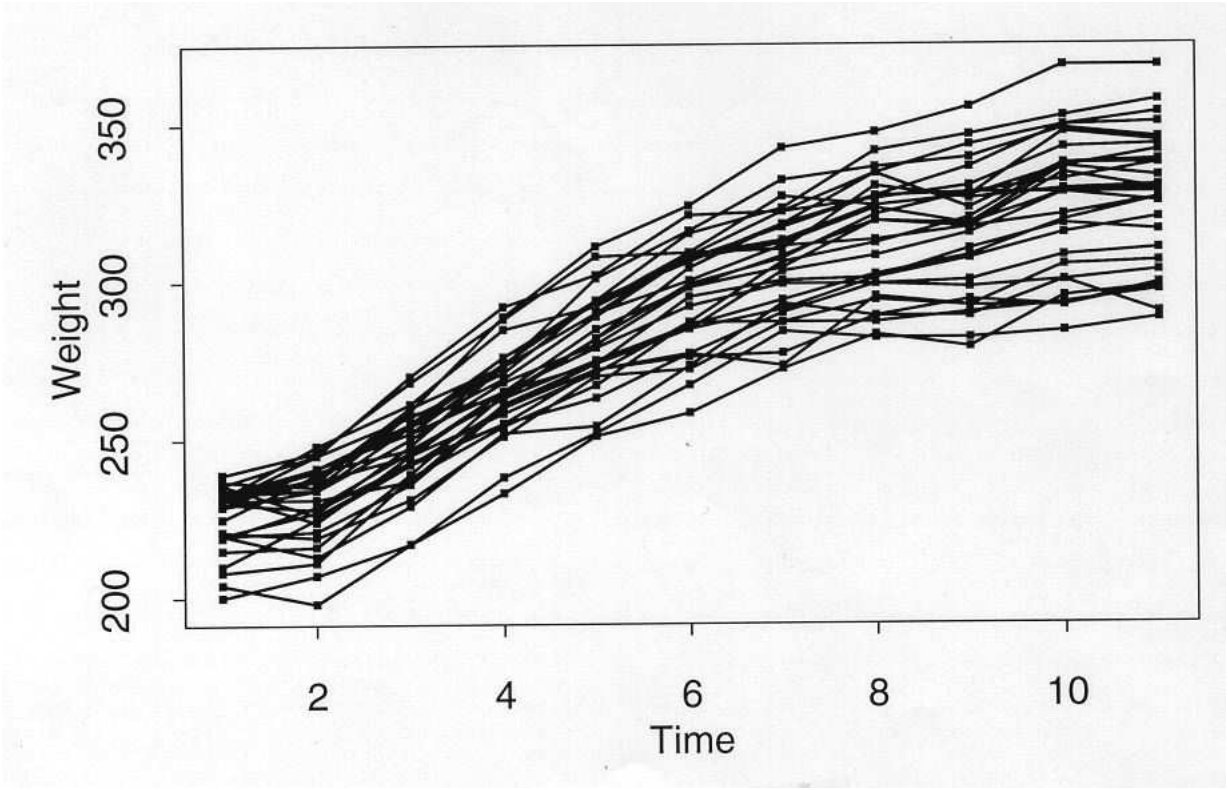
$$n >> m,$$

and S is singular.

Ledoit & Wolf (2004). **Honey, I shrunk the sample covariance matrix.**
J. Portfolio Manag., 4, 110-119.

- Kenward's (1987) Cattle Data:

An experiment to study effect of treatments on intestinal parasites. $m = 30$ animals received treatment A, they were weighed $n = 11$ times, the first 10 measurements were made at two-week intervals and the final measurement was made after a one week interval. The times are rescaled to $t_j = 1, 2, \cdots, 10, 10.5$.



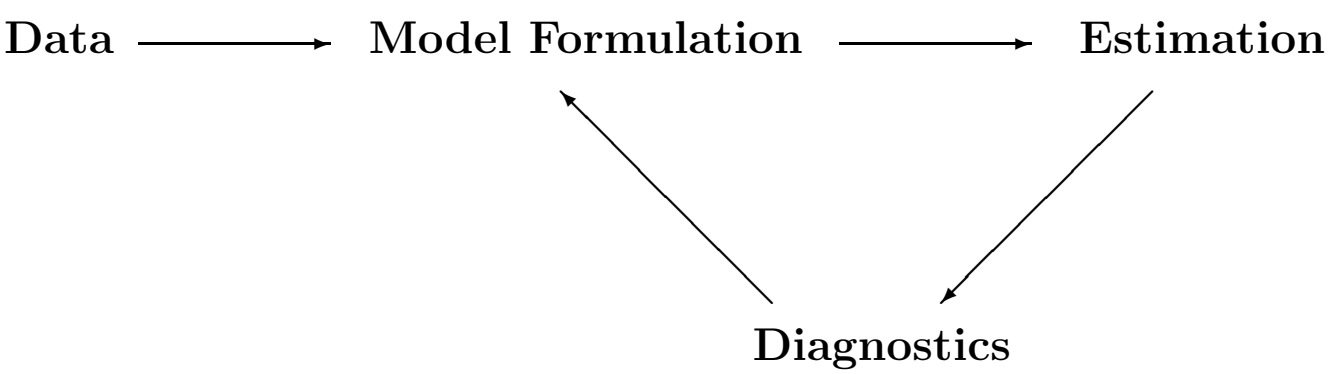
- Clearly, variances **increase** over time,
- Are equidistant measurements equicorrelated?

TABLE 1. Sample variances are along the main diagonal and correlations are off the main 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 (the learning effect) and **decrease** along the columns.
- **Stationary** (Toeplitz) covariance is not advisable for such data.
- SAS PROC MIXED and other software packages provides a long menu of covariance structures to choose from.
- How to view a 102×102 cov. matrix?

Goal: Model a covariance matrix using covariates similar to modeling the mean vector in regression analysis.



- Models:

– **Generalized Linear Models** (GLM) for the mean vector $\mu = E(Y)$ amounts to setting

$$g(\mu) = X\beta,$$

where g acts *componentwise* on the vector μ.

– GLM for the covariance matrix

$$\Sigma = E(Y - \mu)(Y - \mu)',$$

requires finding $g(\cdot)$ so that entries of $g(\Sigma)$ are **unconstrained**, then one may set

$$g(\Sigma) = Z\alpha.$$

- $g(\cdot)$ acting *componentwise* cannot remove the positive-definiteness constraint.

- 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 $\alpha = (\alpha_1, \ldots, \alpha_q)$, which gets translated 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 work **componentwise**, Diggle & Verbyla (1998); Yao, Müller and Wang (2005), ...

- **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 $\log \sum$?

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.

– When Σ is **diagonal**, then the LLM reduces to modeling **variance heterogeneity** (Cook & Weisberg, 1983).

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

Ex. If $\log \Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$, then

$$\sigma_{11} = \frac{1}{2\sqrt{\Delta}} \exp\left(\frac{\alpha+\gamma}{2}\right) \left\{\sqrt{\Delta}\, u^{+} - (\alpha-\gamma)u^{-}\right\},$$

where $\Delta = (\alpha-\gamma)^2 + 4\beta^2$,

$$u^{\pm} = \exp\left(\frac{\sqrt{\Delta}}{2}\right) \pm \exp\left(-\frac{\sqrt{\Delta}}{2}\right).$$

Ref.

1. Leonard & Hsu (1992). Bayesian inference for a covariance matrix. *Ann. of Stat.*, 20, 1669-1696.

2. Chiu, Leonard & Tsui (1996). The matrix-logarithm covariance model. *JASA*, 91, 198-210.

3. Pinheiro & Bates (1996). Unconstrained parameterizations for variance-covariance matrices.
   Stat. Comp., 289-296.

- **Time Series & Cholesky Decomposition:**
  The AR(2) model

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

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

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

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

$$TY = \varepsilon + Ce,$$

where

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

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

$$\begin{array}{rcl} T\mathrm{cov}(Y)T' & = & \sigma^2 I_n + \begin{pmatrix} C_1 \mathrm{cov}(e) C_1' & 0 \\ 0 & 0 \end{pmatrix} \\ & = & \text{A nearly } \textbf{diagonal} \text{ matrix.} \end{array}$$

- **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					
ϕ_{21}	σ_2^2				
ϕ_{31}	ϕ_{32}	σ_3^2			
$\vdots$	$\vdots$		$\ddots$		
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).

Special Cases:

- When $y_1, \cdots, y_n$ are indep., then $\phi_{t,j} \equiv 0, \sigma_t^2 = var(y_t) = \sigma_{t,t}^2$.

  Variance heterogeneity, Cook and Weisberg (1983).

- **Antedependence of order** p, $AD(p)$:

  $y_1, \cdots, y_n$ is $AD(p)$ if the conditional distribution of y_t given $y_{t-1}, \cdots, y_1$ depends on $y_{t-1}, \cdots y_{t-p}$ for all $t \geq p$ (Gabriel, 1962). This is Markovian dependence of order p.

  $$AD(p) \Leftrightarrow \text{ last } n-p-1 \text{ subdiagonals of } T \text{ are zero}$$

  $$\Leftrightarrow \text{ last } n-p-1 \text{ subdiagonals of } \mathbf{\Sigma}^{-1} \text{ are zero.}$$

- **Covariance selection**, Dempster (1972):

- **Gaussian Graphical Models**, Roverato (2000):

  Certain elements of $\mathbf{\Sigma}^{-1}$ are set to zero

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

  For $\sum$ pd, there are unique T and D with positive diagonal entries such that

  $$T \sum T' = D.$$

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

  Link functions: $g(\sum) = 2I - T - T' + logD$,

  a symmetric matrix with unconstrained and statistically meaningful entries.

  Strategy: Model T "linearly" as in Anderson (1966)

  $\log D$ " " " Leonard et al. (92,96).

  OR

  replace "linearly" by parametrically/nonparam. / Bayesian $\cdots$.

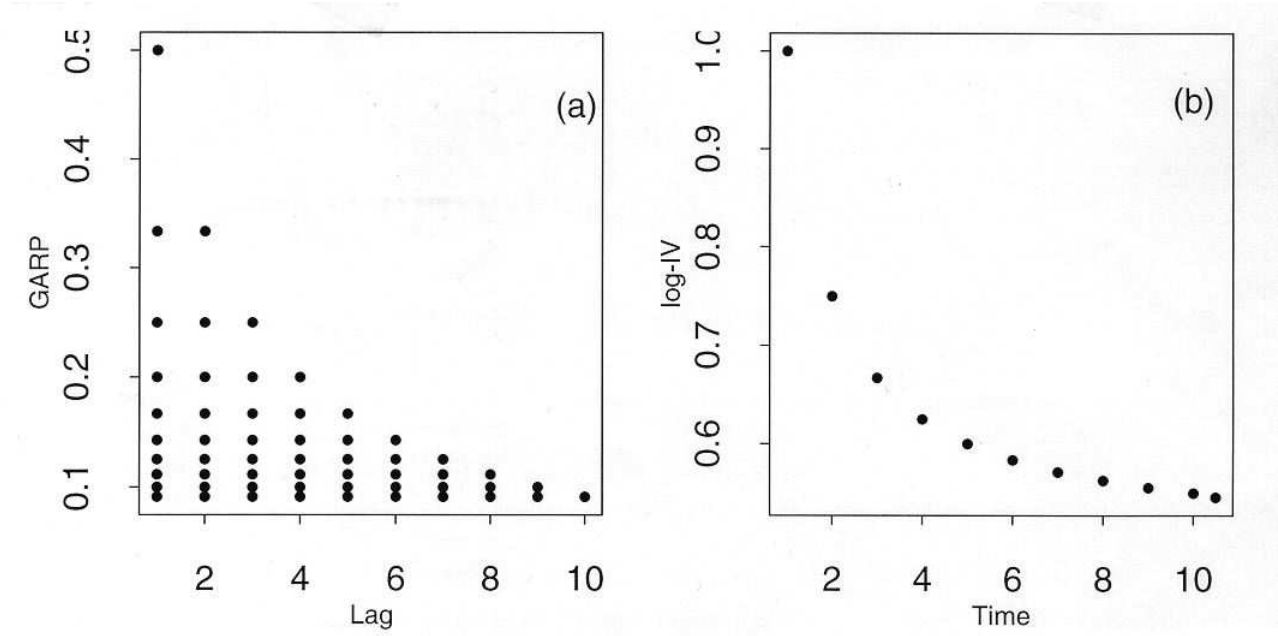
  Bonus: The estimate $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}'^{-1}$ is always pd, where $\hat{T}$ and $\hat{D}$ are estimates of **parsimoniously** modeled T and D.

  Q. How to identify models for (T, D) ?
  A. Use covariates OR shrink to zero the smaller entries of T using penalized likelihood, etc..

- **Regressogram*** :

  Plays roles similar to the correlogram. For a $t \geq 2$, simply plot the GARP $\phi_{t,j}$ vs the lags $j = 1, 2, \cdots, t-1$, and plot log σ_t^2 vs $t = 1, 2, \cdots, n$.

  Ex. Compound Symmetry Covariance ($\rho = .5, \sigma^2 = 1$):



Ex. $AR(p)$, $AD(p)$.

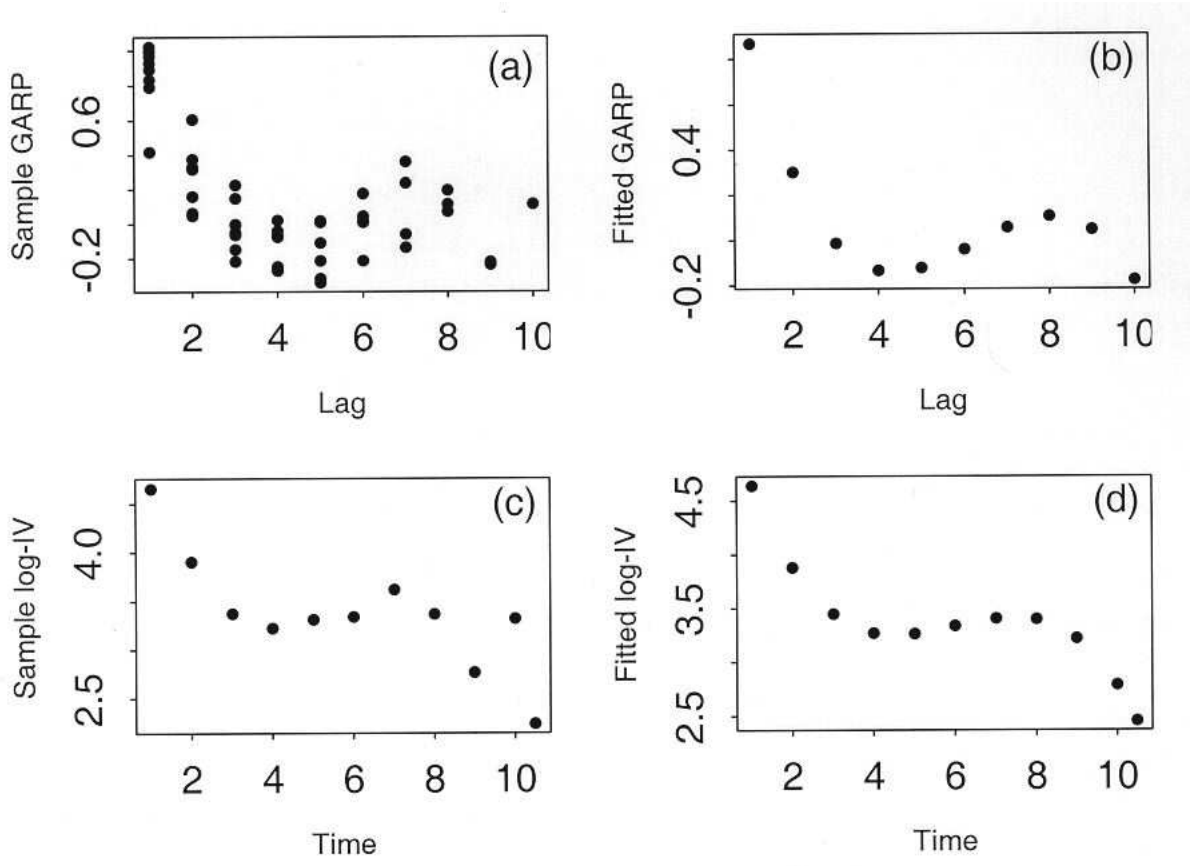
Other Graphical Tools: Scatterplot Matrices;

Partial Scatterplot Matrices (Zimmerman, 2000)

Variogram (Diggle, 1988); Lorelogram (Heagerty & Zeger, 1998).

⋮

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



Sample and Fitted Regressograms for the Cattle Data. (a) Sample GARP, (b) Fitted GARP, (c) Sample log-IV and (d) Fitted log-IV.

- **Model Formulation:**

Regressogram suggests cubic models for the GARP and log IV for the cattle data with 8 param. For $t = 1, 2, \cdots, 11$, and $j = 1, 2, \cdots, t-1$.

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

In general, these and μ_t can be modeled 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.

Pourahmadi (1999). Joint mean-covariance models with
applications to longitudinal data;
Unconstrained parameterization.
Biometrika, 86, 677-690.

• **Estimation of** $\theta = (\beta', \lambda', \gamma')$:

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

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

where $r_i = Y_i - X_i\ \beta = (r_{it})_{t=1}^{n}$, RSS_t and $Z(i)$ depend on r_i and other covariates and parameter values.

For the estimation algorithm and asymptotic distribution of the MLE of θ, see Theorem 1 in

Pourahmadi (2000). MLE of GLMs for MVN covariance matrix.
Biometrika, 87, 425-435.

Example. Cattle Data

Table 2: Values of L_{max}, NO. of parameters and BIC for several models. The last four rows are from Zimmerman & Núñez-Antón (97).

Model	L_{max}	NO. of Parameters	BIC
Unstructured $\sum$	-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)	-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) with $\lambda_1 = \lambda_2 = 1$	-1054.20	6	70.96

Likelihood Ratio Test:

$$2(L_1 - L_0) = 62.14 \sim \chi^2_1,$$

so $(t-j)^3$ is kept in the model.

Other developments:

– Covariate-selection (Pan & MacKenzie, 2003). Relies on AIC & BIC, not the regressogram.

– Bayesian (Daniels & Pourahmadi, 02, 03; Kohn and Smith 02).

$$g(\Sigma) \sim N(\quad , \quad).$$

– Nonparametric (Wu & Pourahmadi, 2003). Smooth (T, D) using

$$\log \sigma_t^2 = \sigma^2(t/n),$$

$$\phi_{t,t-j} = f_j(t/n),$$

where $\sigma^2(\cdot)$ and $f_j(\cdot)$ are smooth functions on $[0, 1]$.

- This formulation is fairly standard in the nonparametric regression literature where one pretends to observe $\sigma^2(\cdot)$ and $f_j(\cdot)$ on finer grids as n gets large.

- One may restrict attention to the first p subdiagonals of T for a small p.

- Amounts to approximating T by the varying-coefficients AR:

$$y_t = \sum_{j=1}^{p} f_j(t/n) y_{t-j} + \sigma(t/n)\varepsilon_t.$$

– Nonparametric/Penalized likelihood (MP, Huang, Liu & Liu, 05).

$$Y_1, \cdots, Y_m \text{ a sample from } N(0, \Sigma)$$

- Log-likelihood function

$$\begin{aligned} -2L(\Sigma) &= m \log |\Sigma| + \sum_{i=1}^{m} Y_i' \Sigma^{-1} Y_i \\ &= \sum_{t=1}^{n} \left(m \log \sigma_t^2 + \sum_{i=1}^{m} \frac{\varepsilon_{it}^2}{\sigma_t^2} \right), \end{aligned}$$

  where

$$\varepsilon_{it} = y_{it} - \sum_{j=1}^{t-1} \phi_{tj} \; y_{ij} \; , \; t = 1, \cdots, n.$$

- Penalized likelihood with L_p penalty,

$$-2L(\Sigma) + \lambda \sum_{t=2}^{n} \sum_{j=1}^{t-1} |\phi_{tj}|^p,$$

  where $\lambda > 0$ is a tuning parameter.

- $p = 2$, corresponds to **Ridge Regression**,
- $p = 1$, **LASSO** (Tibshirani, 1996): **L**east **a**bsolute **s**hrinkage and **s**election **o**perator.

– Use of L_1 norm, allows LASSO to do **variable selection**–it can produce coefficients that are **exactly** zero.

– LASSO is most effective when there are a small to moderate number of moderate-sized coefficients.

- **Bridge Regression** ($p > 0$), Frank & Friedman (1993), Fu (1998); Fan & Li (2001).

- For the Call Center Data with $n = 102$ and 5151 parameters in T, about 4144 are essentially zero.

  L. Brown (2005). Statistical Analysis of a Telephone Call Center:

  A Queueing Science Perspective. JASA, 36-50.

•• Simultaneous Modeling of Several Covariance Matrices
(Pourahmadi, Daniels, Park, 2005).
Applications to Cluster Analysis, Classification, $\cdots$.

REFERENCES

Anderson, T.W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1**, 135-141.

Dempster, A.M. (1972). Covariance selection, *Biometrics*, **28**, 157-175.

Diggle, P.J., Verbyla, A.P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, **54**, 401-415.

Gabriel, K.R. (1962). Ante-dependence analysis of an ordered set of variables. *Ann. Math. Statist.*, **33**, 201-212.

Kenward, M.G. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296-308.

Pinheiro, J.D. and Bates, D.M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statist. Comp.* **6**, 289-296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterization. *Biometrika*, **86**, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425-435.

Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*, John Wiley, New York.

Pourahmadi, M. and Daniels, M. (2002). Dyanamic conditionally linear mixed models for longitudinal data. *Biometrics*, **58**, 225-231.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, **87**, 99-112.

Zimmerman, D.L. and V. Núñez-Antón (1997). Structured antedependence models for longitudinal data. In *Modelling Longitudinal and Spatially Correlated Data. Methods, Applications, and Future Directions*, 63-76 (T.G. Gregoine, et al., eds.) Springer-Verlag, New York.

Pearson



$P\Sigma P' = \Lambda$

Edgeworth



Σ^{-1}

Yule



AR(2), Correlogram,
Odds ratio

Szegö

LM → Log-Lin.

1901 — 1918 — TS — MM — 1966 — 1996

Pearson — Cholesky — Anderson — WISC.

GLM

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

1805 — $\mu = X\beta$ — $g(\mu) = X\beta$ — **1972**

LS



Legendre



Gauss

Covariate



Galton

GLM



Nelde **Wedderburn**