Polynomial Optimization in Multivariate Statistics

Maximum Likelihood Estimation in Seemingly Unrelated Regressions

Mathias Drton

Department of Statistics
University of Washington
drton@stat.washington.edu

Outline 1

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

Linear regression

- Random variables (observed for several years $t = 1, \ldots, n$):

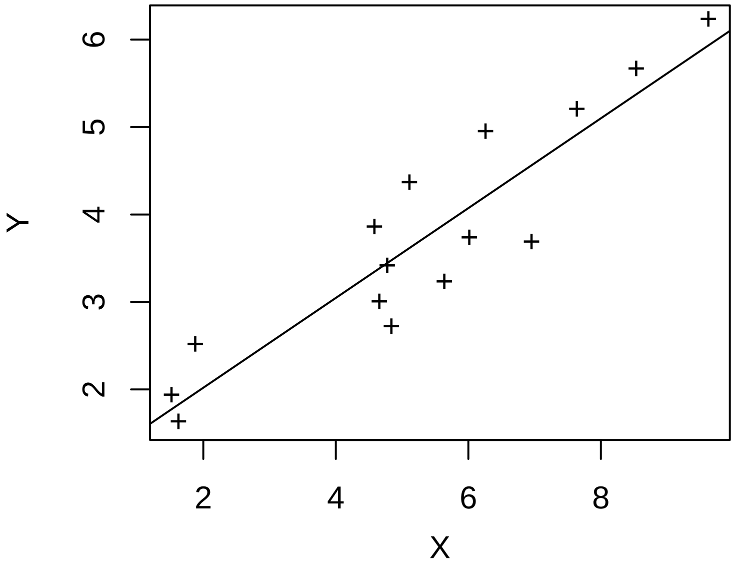
  Y_t : company's investment in year t

  X_t : company's market value at end of year $t-1$

- How does Y_t depend on X_t?

$$Y_t = \alpha + \beta X_t + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

- Least squares regression line:



Y_t : response variable

X_t : covariate

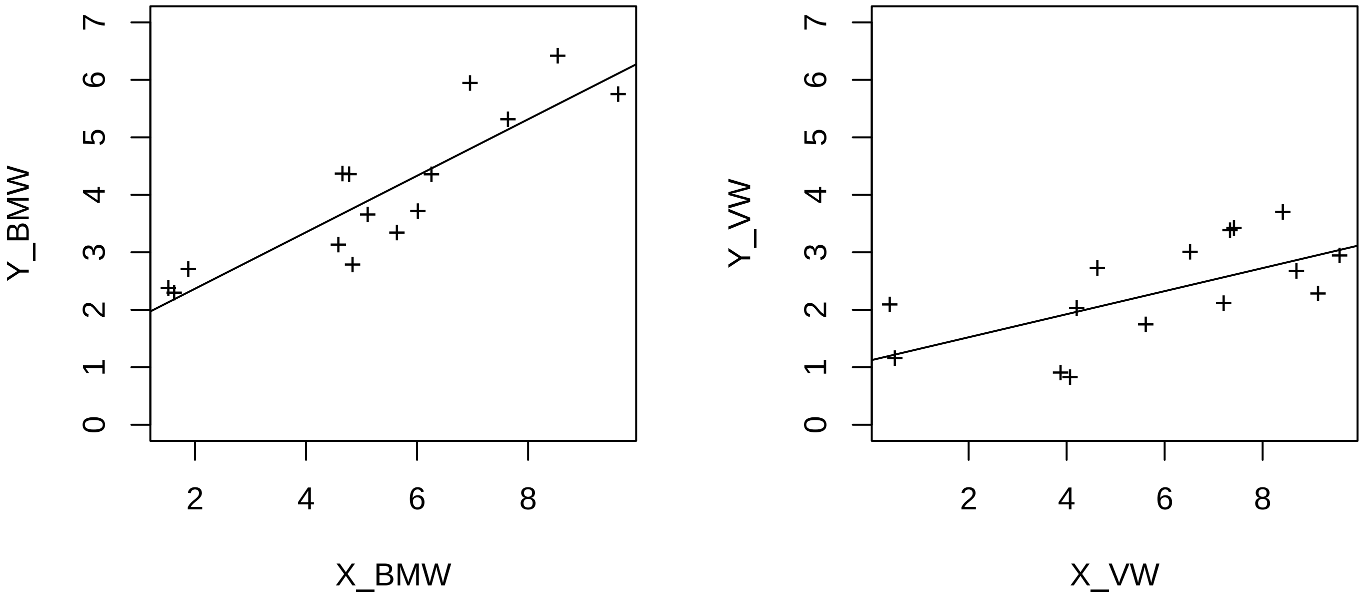
ε_t : residual

Seemingly Unrelated Regressions

- Two companies:

$$\begin{aligned} Y_t^{\mathrm{BMW}} &= \alpha^{\mathrm{BMW}} + \beta^{\mathrm{BMW}} X_t^{\mathrm{BMW}} + \varepsilon_t^{\mathrm{BMW}} \\ Y_t^{\mathrm{VW}} &= \alpha^{\mathrm{VW}} + \beta^{\mathrm{VW}} X_t^{\mathrm{VW}} + \varepsilon_t^{\mathrm{VW}} \end{aligned}$$

- Two separate regression lines:

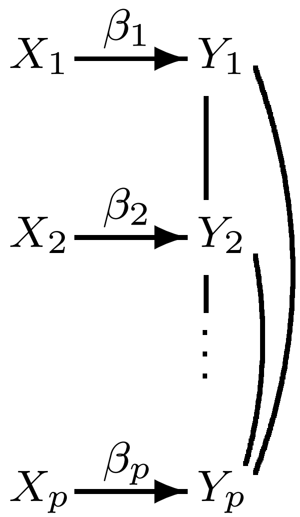


- **BUT:**

  If $\mathrm{corr}[\varepsilon_t^{\mathrm{BMW}}, \varepsilon_t^{\mathrm{VW}}] \neq 0$, then better to estimate two lines simultan.

Seemingly Unrelated Regressions

- More general SUR (Zellner, 1962 & 1963):



$$\begin{aligned} Y_{1t} &= \beta_1 X_{1t} + \varepsilon_{1t} \\ \vdots \quad & \quad \vdots \\ Y_{pt} &= \beta_p X_{pt} + \varepsilon_{pt} \end{aligned}$$

where $t = 1, \ldots, n$ and

$$\varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{pt} \end{pmatrix} \overset{i.i.d.}{\sim} \mathcal{N}_p(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{1p} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathbb{R}^{p \times p}$$

- "Central role in contemporary econometrics." (Goldberger, 1991)

Gaussian Model

- Observations (Data)

$$Y = \begin{pmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \vdots & \vdots \\ Y_{p1} & \cdots & Y_{pn} \end{pmatrix}, \qquad X = \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \vdots & \vdots \\ X_{p1} & \cdots & X_{pn} \end{pmatrix}$$

- X fixed, Y random as

$$Y \sim \mathcal{N}\left(BX, \Sigma \otimes I_n\right),$$

where

$$B = \begin{pmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & 0 & \cdots \\ & & \ddots & \\ 0 & \cdots & 0 & \beta_p \end{pmatrix}$$

General SUR

- SUR model = family of distributions of $Y \in \mathbb{R}^{p \times n}$ for fixed $X \in \mathbb{R}^{q \times n}$:

$$\text{Model} \;=\; \big(\mathcal{N}(BX, \Sigma \otimes I_n) \mid (B, \Sigma) \in \mathbb{B} \times \mathbb{P}\big),$$

  where
  - $\mathbb{P}$ = cone of positive definite (real) $p \times p$ matrices,
  - $\mathbb{B}$ = set of real $p \times q$ matrices with an *a priori* zero pattern,
    i.e. $\beta_{ij} = 0$ for some set of (i, j).
- Terminology: (B, Σ) parameters, $\mathbb{B} \times \mathbb{P}$ parameter space

- Graphical representation:
  - Undirected edges between all Y_i, $i = 1, \ldots, p$,
  - Directed edge $X_j \to Y_i$ if $\exists B \in \mathbb{B}$ s.t. $\beta_{ij} \neq 0$

Outline 2

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

Likelihood Function

- Data on Y and X (two matrices in $\mathbb{R}^{p\times n}$ and $\mathbb{R}^{q\times n}$, resp.)
- "Maximum likelihood" principle to estimate parameters (B,Σ):

  maximize the likelihood function of the model!
- Prob. density function $f_{(B,\Sigma)}:\mathbb{R}^{p\times n}\to(0,\infty)$ of $\mathcal{N}(BX,\Sigma\otimes I_n)$:

$$f_{(B,\Sigma)}(Y)=\frac{1}{\sqrt{(2\pi)^{pn}|\Sigma|^n}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(Y-BX)(Y-BX)'\right]\right\}$$

- Likelihood function of the model for fixed data Y:

$$\begin{aligned} L:\mathbb{B}\times\mathbb{P} &\to (0,\infty) \\ (B,\Sigma) &\mapsto f_{(B,\Sigma)}(Y) \end{aligned}$$

Likelihood Equations

- More convenient to maximize: log-likelihood function

$$\ell(B,\Sigma) = \log L(B,\Sigma) \propto -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(Y-BX)(Y-BX)'\right]$$

- Partial derivatives:

$$\frac{\partial \ell(B,\Sigma)}{\partial \Sigma} = 0 \iff \Sigma = \frac{1}{n}(Y-BX)(Y-BX)'$$

$$\frac{\partial \ell(B,\Sigma)}{\partial \beta} = 0 \iff \beta = \left[A'(XX'\otimes\Sigma^{-1})A\right]^{-1} A'\mathrm{vec}(\Sigma^{-1}YX')$$

Here, $\beta \in \mathbb{R}^a$ is the vector of unrestricted elements in B, and $A = \dfrac{\partial B}{\partial \beta}$ is a matrix of ones and zeros that satisfies $\mathrm{vec}(B) = A\beta$.

Standard Algorithm

- Algorithm for ML estimation:
  - Estimate B for known Σ (generalized least squares)
  - Estimate Σ for known B (covariance matrix of residuals)
  - Iterate until "convergence"

- Literature:
  - Magnus (1978): "this [consistent] root [of the likelihood equations] is the unique ML estimator."
  - Greene (1997, textbook!): "the log-likelihood [of a SUR model] is globally concave."
  - Srivastava/Giles (1987): Limit of the algorithm is unique if the initial estimate is consistent.

Outline 3

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

Profile Likelihood

- Profile log-likelihood

$$\begin{aligned}\ell(\beta) &= \max\{\ell\big(B(\beta), \Sigma\big) : \Sigma \text{ pos.def.}\} \\ &\propto -\frac{n}{2}\log|\hat{\Sigma}(\beta)| - \frac{np}{2},\end{aligned}$$

  where

$$\hat{\Sigma}(\beta) = \frac{1}{n}\big(Y - B(\beta)X\big)\big(Y - B(\beta)X\big)'.$$

- Recall:

$$\ell(B, \Sigma) \propto -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\big[\Sigma^{-1}(Y - BX)(Y - BX)'\big]$$

- Stationary points:

$$\begin{aligned}&(\beta, \Sigma) \text{ is stationary point of } \ell\big(B(\beta), \Sigma\big) \\ \iff \quad &\beta \text{ is stationary point of } \ell(\beta) \text{ and } \Sigma = \hat{\Sigma}(\beta).\end{aligned}$$

Polynomial Optimization

- Profile log-likelihood

$$\ell(\beta) = -\frac{n}{2}\log|\hat{\Sigma}(\beta)| - \frac{np}{2}$$

  is a monotone function of $|\hat{\Sigma}(\beta)|$

- For maximum likelihood estimation we can solve the unconstrained optimization problem

$$\min_{\beta\in\mathbb{R}^a} \left|\big(Y - B(\beta)X\big)\big(Y - B(\beta)X\big)'\right|$$

- Note: $G(\beta) = \left|\big(Y - B(\beta)X\big)\big(Y - B(\beta)X\big)'\right|$ is a polynomial in β!!

Outline 4

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

Bivariate SUR Model

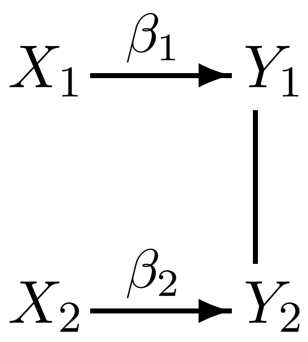
- Bivariate model $Y = BX + \varepsilon$ with

$$Y = \begin{pmatrix} Y_{11} & \dots & Y_{1n} \\ Y_{21} & \dots & Y_{2n} \end{pmatrix}, \quad X = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ X_{21} & \dots & X_{2n} \end{pmatrix}$$

  and

$$B = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix}$$

- Graph



Polynomial Optimization Problem

- Minimize

$$
\begin{aligned}
G(\beta) &= \left| \left[Y - \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix} X \right] \left[Y - \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix} X \right]' \right| \\
&= \sum_{i=0}^{2} \sum_{j=0}^{2} a_{ij} \beta_1^i \beta_2^j
\end{aligned}
$$

- Constraints on coefficients a_{ij}, for example:

$$
\begin{array}{llll}
a_{22} > 0, & a_{20} > 0, & a_{02} > 0, & a_{00} > 0, \\
a_{21}^2 < 4a_{22}a_{20}, & a_{12}^2 < 4a_{22}a_{02}, & a_{10}^2 < 4a_{20}a_{00}, & a_{01}^2 < 4a_{02}a_{00}, \\
|a_{11}| < 2\sqrt{(a_{22}a_{00})} + 2\sqrt{(a_{20}a_{02})} & & &
\end{array}
$$

Zeros of the Derivative

- Partial Derivatives

$$g_1 = \frac{\partial G(\beta)}{\partial \beta_1}, \qquad g_2 = \frac{\partial G(\beta)}{\partial \beta_2}$$

- Maximum Likelihood Ideal $I_G = < g_1, g_2 >$: Dim = 0, Degree = 5
- Lex-order Groebner basis ($\beta_1 > \beta_2$):
  - quintic in β_2
  - linear in β_1 + quartic in β_2
- It is possible that $V_{\mathbb{R}}(I_G) = V_{\mathbb{C}}(I_G) = 5$!

  (Up to five stationary points of likelihood of this bivariate SUR model)

Simulations from True Model

	$\rho = -0.92$		$\rho = 0$		$\rho = 0.92$	
n	N_b	N_i	N_b	N_i	N_b	N_i
5	261	80	328	166	262+1	83
6	79	21	120	53	96	25
7	20	7	44	19	29	13
8	16	5	23	15	7	2
9	3	2	11	3	2	0
10	1	0	2	1	0	0
11	1	1	1	1	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0

Parameters: $\Sigma^X = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$, $\beta = (1, 2)$, $\Sigma = \begin{pmatrix} 1 & \rho\sqrt{2} \\ \rho\sqrt{2} & 2 \end{pmatrix}$

Simulations from Wrong Model

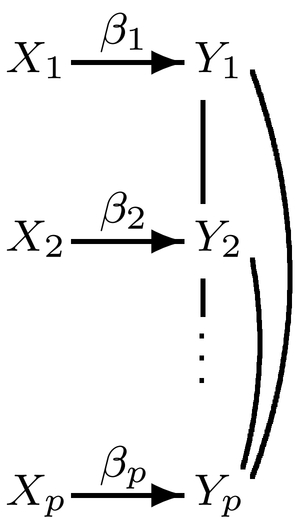
	$\rho = -0.92$		$\rho = 0$		$\rho = 0.92$	
n	N_b	N_i	N_b	N_i	N_b	N_i
5	2803	1363	3362	1233	2698	1335
6	2095+2	1032	2991	1016	2172	1056
7	1721	830	2771	852	1729	867
8	1448	717	2634	731	1459	679
9	1253	563	2617	722	1174	592
10	994	446	2467	683	985	423
11	831	382	2390	621	874	389
12	691	309	2338	590	734	348
13	599	255	2208	558	611	265

Parameters: $\Sigma^X = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 5 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 & 0.4 \\ 1.8 & 2 \end{smallmatrix}\right)$, $\Sigma = \begin{pmatrix} 1 & \rho\sqrt{2} \\ \rho\sqrt{2} & 2 \end{pmatrix}$

Outline 5

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

SUR with diagonal B



$$B = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_p \end{pmatrix}$$

$$G(\beta) = \sum_{i_1=0}^{2} \cdots \sum_{i_p=0}^{2} a_{i_1 \cdots i_p} \beta_1^{i_1} \times \cdots \times \beta_p^{i_p}$$

- ML-Ideal I_G:

$p = 3$, Dim = 0, Degree = 29
$p = 4$, Dim = 1, Degree = 32
$p = 5$, Dim = 2, Degree = 80

How many real points?

Other Bivariate SUR Models

	Dim	Degree
$B = \begin{pmatrix} \beta_{11} & \beta_{12} & 0 \\ 0 & 0 & \beta_{23} \end{pmatrix}$	Dim = 0	Degree = 9
$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & 0 \\ 0 & 0 & 0 & \beta_{24} \end{pmatrix}$	1	4
$B = \begin{pmatrix} \beta_{11} & \beta_{12} & 0 & 0 \\ 0 & 0 & \beta_{23} & \beta_{24} \end{pmatrix}$	1	8
$B = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & 0 \end{pmatrix}$	0	5
$B = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}$	1	4

Submodels

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & 0 \\ 0 & 0 & \beta_{12} \end{pmatrix}$$

Dim = 0 Degree = 7

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & 0 \\ 0 & 0 & 0 & \beta_{13} \end{pmatrix}$$

0 11

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & 0 & 0 \\ 0 & 0 & \beta_{12} & \beta_{24} \end{pmatrix}$$

0 23

$$B = \begin{pmatrix} \beta_{1} & 0 & 0 & 0 \\ 0 & \beta_{1} & 0 & 0 \\ 0 & 0 & \beta_{3} & 0 \\ 0 & 0 & 0 & \beta_{4} \end{pmatrix}$$

0 63

Outline 6

- Seemingly unrelated regressions
- Maximum likelihood estimation
- Transformation into polynomial optimization problem
- Bivariate Example
- Other Examples
- Conclusion

Conclusion

- Can we find data such that $V_{\mathbb{R}}(I_G) = V_{\mathbb{C}}(I_G)$?
- Can we have $\#V_{\mathbb{R}}(I_G) = \infty$?
- Try out global optimization methods...