Statistical Analysis with Missing Data

Module 7

Maximum Likelihood/Bayes

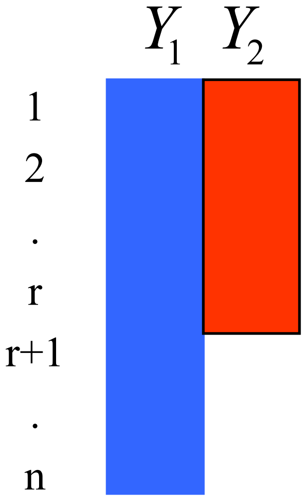
Factored likelihood methods

UNIVERSITY OF MICHIGAN

Computational tools

- Tools for monotone patterns
  - Maximum likelihood based on factored likelihood
  - Draws from Bayesian posterior distribution based on factored posterior distributions
  - Multiple imputation for monotone patterns
  - Creating propensity weights for monotone patterns

Bivariate Monotone Data



Under MAR, loglikelihood for monotone bivariate data is

$$\ell(\theta \mid y_{(0)}) = \sum_{i=1}^{r} \log f(y_{i1}, y_{i2} \mid \theta) + \sum_{i=r+1}^{n} \log f(y_{i1} \mid \theta)$$

E.g. for bivariate normal data:

$$\ell(\mu, \Sigma \mid y_{(0)}) =$$

$$-0.5r \log\left|\Sigma\right| - (1/2)\sum_{i=1}^{r}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)$$

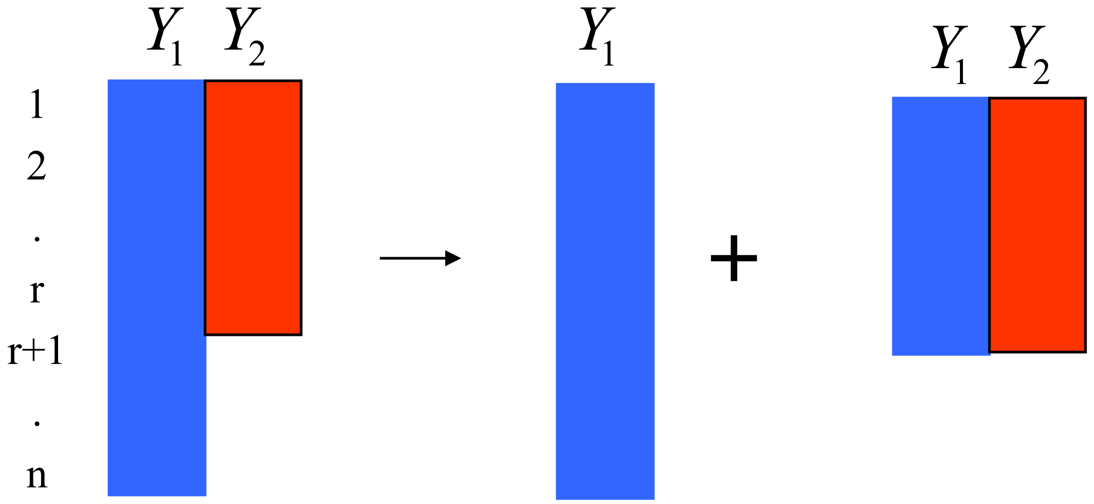
$$-0.5(n-r)\log\sigma_{11} - (1/2)\sum_{i=r+1}^{n}(y_{i1} - \mu_1)^2/\sigma_{11}$$

ML seems to require iterative methods...

Anderson showed that explicit estimates are possible

Bivariate Monotone Data

- Maximum Likelihood by <u>factoring the likelihood</u> (Anderson 1957: Little and Rubin 2019, chapter 7):



Joint of Y_1, Y_2 Marginal of Y_1 x Conditional of Y_2 on Y_1

$$f(y_1,y_2|\theta) = f(y_1|\phi_1(\theta)) \times f((y_2|y_1,\phi_2(\theta))$$

$$\Rightarrow L(\theta|Y_{obs}) = L_1\left(\phi_1 \mid \{y_{i1} : i=1,\ldots,n\}\right) \times L_2\left(\phi_2 \mid \{(y_{i1},y_{i2}), i=1,\ldots r\}\right)$$

- Method only works when parameters ϕ_1 and ϕ_2 are distinct

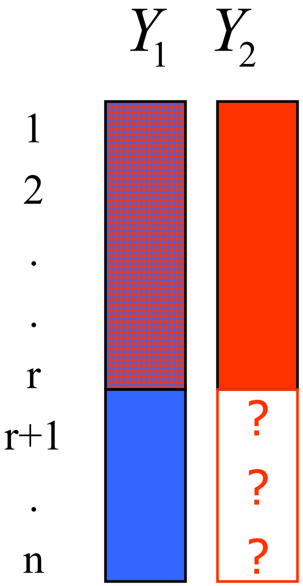
Bivariate Normal Monotone Data

- For Bivariate Normal data:

$$\phi = (\phi_1, \phi_2) \quad \phi_1 = (\mu_1, \sigma_{11}) \; \phi_2 = (\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$$

$$f(y_1, y_2 | \theta) = f(y_1 | \mu_1, \sigma_{11}) \times f((y_2 | y_1, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}))$$

$$\Rightarrow L(\theta | Y_{obs}) = L_1\left(\mu_1, \sigma_{11} \,|\, \{y_{i1} : i = 1, \ldots, n\}\right) \times$$
$$L_2\left(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1} \,|\, \{(y_{i1}, y_{i2}), i = 1, \ldots r\}\right)$$



Yields two standard complete-data problems:

$L_1 \rightarrow$ univariate normal sample of size *n* on Y_1

$L_2 \rightarrow$ regression of Y_2 on Y_1 with sample size *r*

Bivariate Normal Data ctd

- Hence ML estimates are:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}, \hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - \hat{\mu}_1)^2$$

$$\left.\begin{aligned} \hat{\beta}_{21\cdot 1} &= s_{12} / s_{11} \\ \hat{\beta}_{20\cdot 1} &= \bar{y}_2 - \hat{\beta}_{21\cdot 1}\bar{y}_1 \\ \hat{\sigma}_{22\cdot 1} &= s_{22\cdot 1} = s_{22} - s_{12}^2 / s_{11} \end{aligned}\right\}$$

Least squares on *r* complete cases

Estimates of other parameters

- ML for other parameters by expressing them as functions of ϕ and applying transformation property of ML:

$$\mu_2 = \beta_{20\cdot 1} + \beta_{21\cdot 1}\mu_1$$

$$\Rightarrow \hat{\mu}_2 = \hat{\beta}_{20\cdot 1} + \hat{\beta}_{21\cdot 1}\hat{\mu}_1 = \overline{y}_2 + \hat{\beta}_{21\cdot 1}(\hat{\mu}_1 - \overline{y}_1)$$

which is known as the *regression estimate* of the mean

$$\hat{\sigma}_{22} = \hat{\sigma}_{22\cdot 1} + \hat{\beta}^2_{22\cdot 1}\hat{\sigma}_{11} = s_{22} + \hat{\beta}^2_{21\cdot 1}(\hat{\sigma}_{11} - s_{11})$$

$$\hat{\sigma}_{12} = \hat{\beta}_{21\cdot 1}\hat{\sigma}_{11} = s_{12}(\hat{\sigma}_{11} / s_{11})$$

Standard errors

- Large sample standard errors for estimates of ϕ follow from complete-data distribution theory

$$Var(\hat{\phi}-\phi)=\begin{bmatrix} Var(\hat{\phi}_1-\phi_1) & 0 \\ 0 & Var(\hat{\phi}_2-\phi_2) \end{bmatrix}$$

Large sample standard errors for estimates of functions of ϕ follow using standard ML transformation theory, that is:

$$Var\left(\theta(\hat{\phi})-\theta(\phi)\right)=\left(\frac{\partial\theta}{\partial\phi}\right)^T Var(\hat{\phi}-\phi)\left(\frac{\partial\theta}{\partial\phi}\right)$$

Note that these are large-sample expressions, but take into account the missing data (unlike single imputation)

Bayes Inference

- An attractive alternative to ML inference is to draw parameters from Bayes posterior distribution
  - Estimate = posterior mean
  - Standard error = posterior standard deviation
  - Standard errors are often easier than ML!

Bivariate Monotone Data

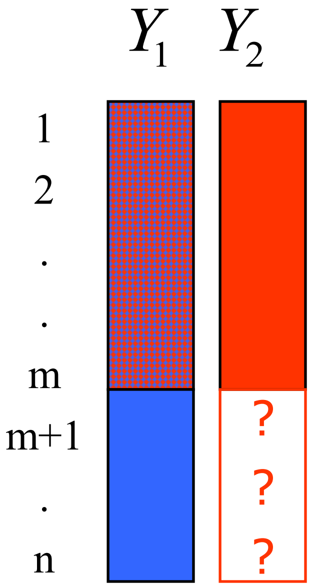
Bayes by *factoring likelihood*

$$f(y_1,y_2|\theta) = f\left(y_1|\phi_1(\theta)\right) \times f\left(\left(y_2|y_1,\phi_2(\theta)\right)\right.$$

Prior: $\pi(\phi_1,\phi_2) = \pi_1(\phi_1) \times \pi_2(\phi_2)$

$\Rightarrow$ Posterior:

$$p(\phi\,|\,Y_{obs}) \propto p(\phi_1|y_{11},\ldots,y_{n1}) \times p\left(\phi_2\,|\,(y_1,\ldots,y_r)\right)$$



$$p(\phi_1|y_{11},\ldots,y_{n1}) \propto \pi_1(\phi_1) \times f\left(y_1|\phi_1(\theta)\right)$$

$$p\left(\phi_2\,|\,(y_1,\ldots,y_r)\right) \propto \pi_2(\phi_2) \times f\left(\left(y_2|y_1,\phi_2(\theta)\right)\right.$$

Complete-data Posteriors

- In normal example with reference priors:

$$\pi_1(\mu_1, \log\sigma_{11})\ \pi_2(\beta_{20\cdot1}, \beta_{21\cdot1}, \log\sigma_{22\cdot1}) \propto const$$

standard complete-data calculations yield:

$\sigma_{11}|data \sim n\hat{\sigma}_{11} / \chi^2_{n-1}$ scaled inverse chi-squared dn

$\mu_1|\sigma_{11}, data \sim N(\hat{\mu}_1, \sigma_{11} / n)$ normal (unconditionally μ_1 is T)

$$\sigma_{22\cdot1} \,|\, data \sim rs_{22\cdot1} / \chi^2_{r-2}$$

$$\beta_{21\cdot1} \,|\, \sigma_{22\cdot1}, data \sim N(\hat{\beta}_{21\cdot1}, \sigma_{22\cdot1} / (rs_{11}))$$

$$\beta_{20\cdot1} \,|\, \beta_{21\cdot1}, \sigma_{22\cdot1}, data) \sim N(\bar{y}_2 - \beta_{21\cdot1}\bar{y}_1, \sigma_{22\cdot1} / r)$$

Posteriors of other parameters

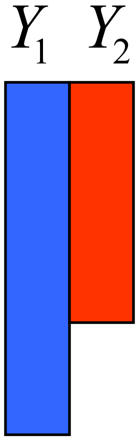
- Posteriors of other parameters can be simulated by expressing them as functions of ϕ and substituting draws $\phi^{(d)}$ from the posterior distribution of ϕ:

$$\mu_2 = \beta_{20\cdot 1} + \beta_{21\cdot 1}\mu_1$$

$$\mu_2^{(d)} = \beta_{20\cdot 1}^{(d)} + \beta_{21\cdot 1}^{(d)}\mu_1^{(d)}$$

$$\sigma_{22}^{(d)} = \sigma_{22\cdot 1}^{2(d)} + \beta_{21\cdot 1}^{2(d)}\sigma_{11}^{(d)}$$

$$\sigma_{12}^{(d)} = \beta_{21\cdot 1}^{(d)}\ \sigma_{11}^{(d)}$$



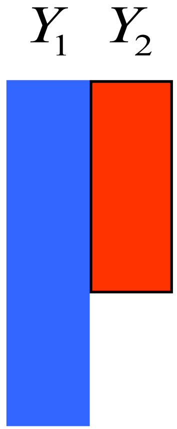
Idea is analogous to ML, with draws replacing ML estimates

S.e.'s are computationally easier, since they are estimated as sample s.d. of the drawn parameters, without the need to invert an information matrix

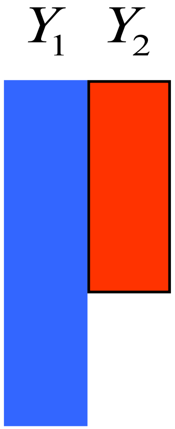
Bayes for Monotone Normal Data

- Simply replace ML estimates, e.g.:

$$\hat{\mu}_2 = \hat{\beta}_{20\cdot 1} + \hat{\beta}_{21\cdot 1}\hat{\mu}_1$$



Bayes for Monotone Normal Data

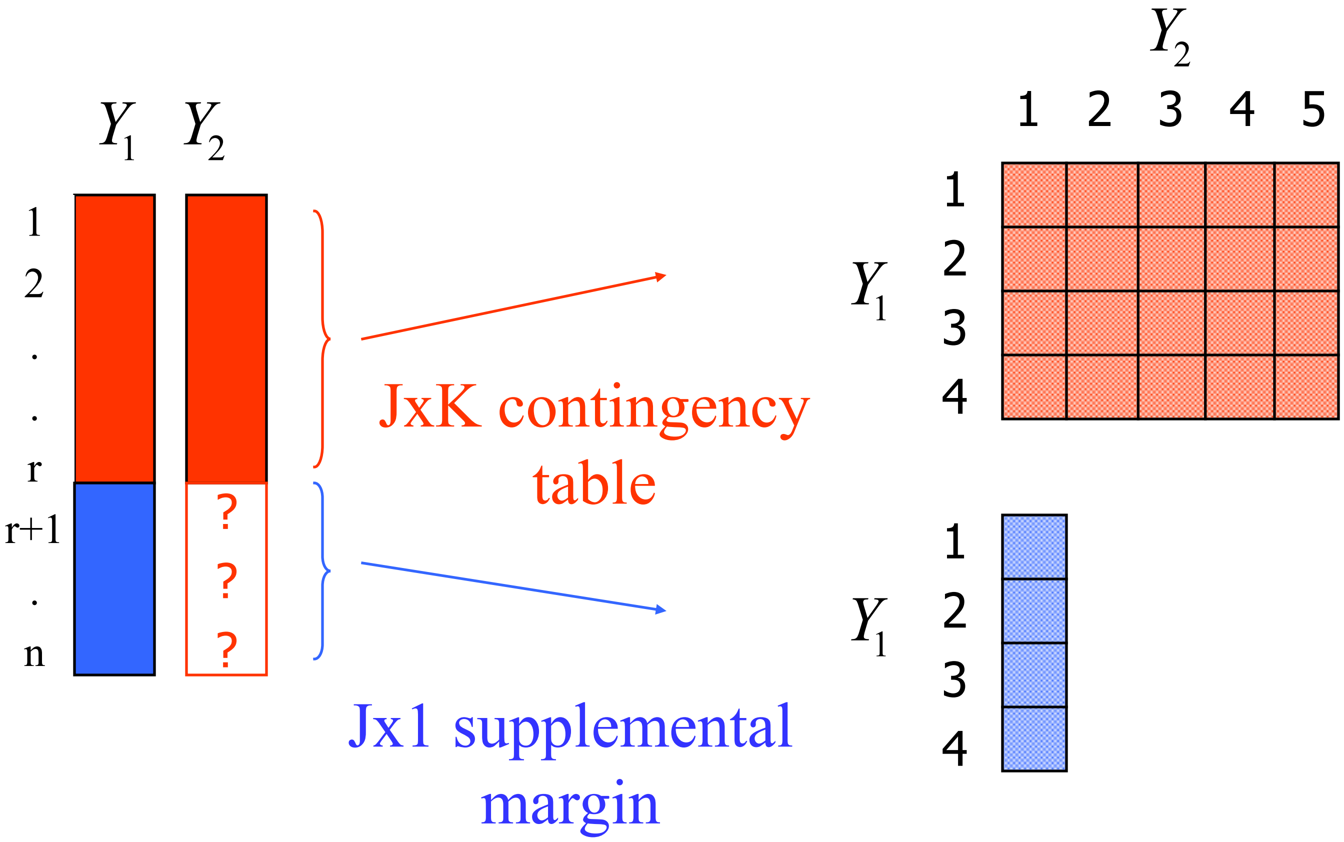


- …by draws (d) from posterior distribution,

$$\mu_2^{(d)} = \beta_{20\cdot 1}^{(d)} + \beta_{21\cdot 1}^{(d)} \mu_1^{(d)}$$

- With conjugate priors, $\beta_{20\cdot 1}^{(d)}, \beta_{21\cdot 1}^{(d)}, \mu_1^{(d)}$ are simple functions of ML estimates and chi-squared and standard normal deviates
  - (Little and Rubin 2002)
- Directly simulates posterior distribution
- Immediate estimates of uncertainty, which incorporate "Student T" corrections for estimating the variances; hence better frequentist properties than maximum likelihood, which is asymptotic

Categorical Data: Contingency Tables



JxK contingency table

Jx1 supplemental margin

Multinomial Probabilities

$$\Pr(Y_1 = j, Y_2 = k) = \theta_{jk}, j = 1,\ldots,J;\ k = 1,\ldots,K$$

$$\phi_1 = (\theta_{1+},\ldots,\theta_{J+})$$

$$\phi_2 = \left\{\phi_{jk} \equiv \theta_{jk} / \theta_{j+}\right\},\ j = 1,\ldots,J; k = 1,\ldots,K$$

$$\theta_{jk} = \theta_{j+} \times \phi_{jk}$$

$$\hat{\theta}_{jk} = \hat{\theta}_{j+} \times \hat{\phi}_{jk}$$

$\hat{\theta}_{j+} = n_{j+} / n = Y_1$-margin proportions, all n cases

$\hat{\phi}_{jk} = r_{jk} / r_{j+} =$ row proportions from r complete cases

Example:2x2 table

		Y_2	
		1	2
Y_1	1	100	50
	2	75	75

complete cases

Y_1	1	30
	2	60

supplemental margin

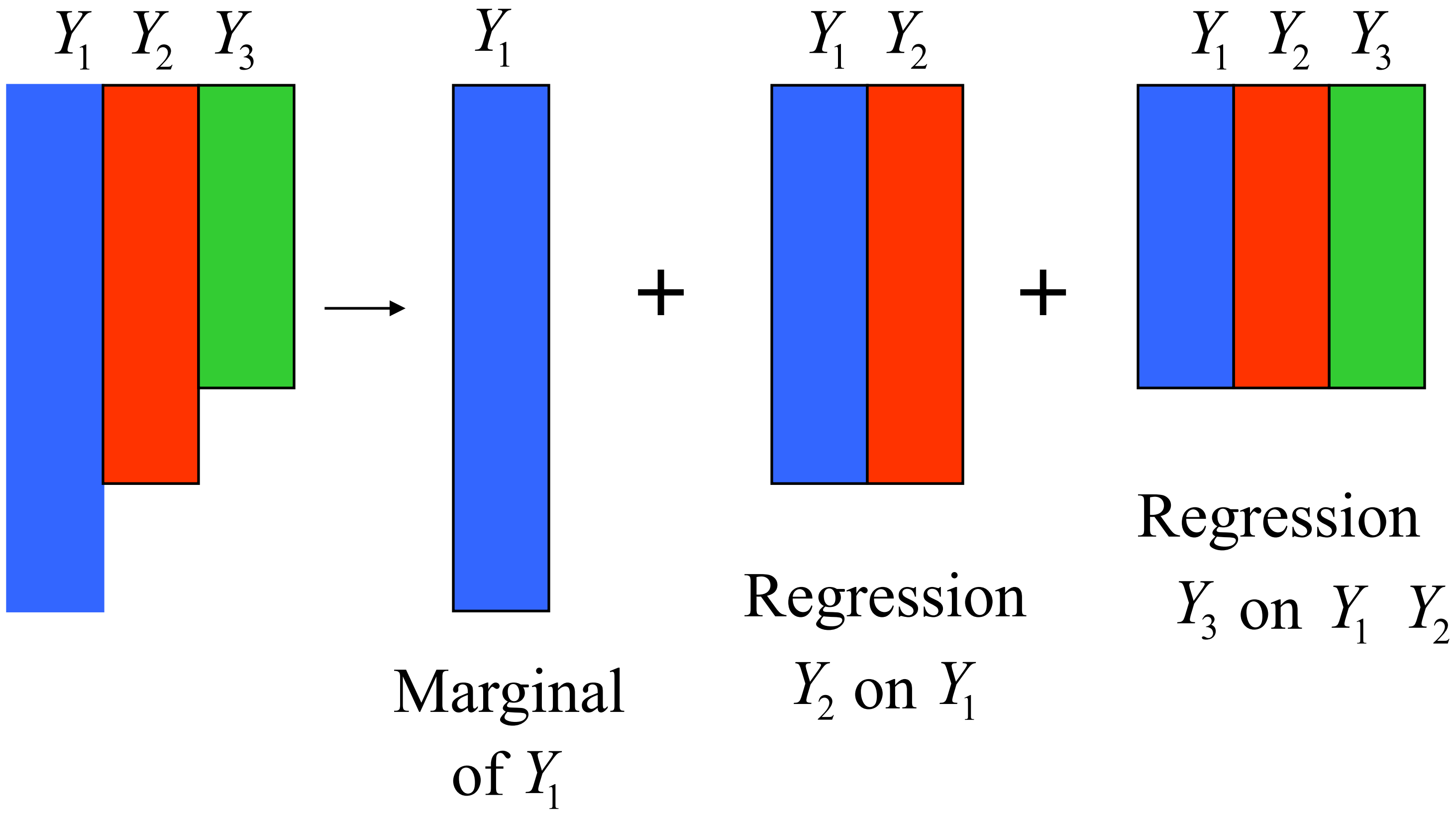
$$\hat{\phi}_1 = \begin{pmatrix} {}^{(100+50+30=180)}\!/_{390} \\ {}^{(75+75+60=210)}\!/_{390} \end{pmatrix} \qquad \hat{\phi}_2 = \begin{pmatrix} {}^{100}\!/_{150} & {}^{50}\!/_{150} \\ {}^{75}\!/_{150} & {}^{75}\!/_{150} \end{pmatrix}$$

$$\hat{\theta} = \begin{pmatrix} \left({}^{180}\!/_{390} \times {}^{100}\!/_{150}\right) & \left({}^{180}\!/_{390} \times {}^{50}\!/_{150}\right) \\ \left({}^{210}\!/_{390} \times {}^{75}\!/_{150}\right) & \left({}^{210}\!/_{390} \times {}^{75}\!/_{150}\right) \end{pmatrix} = \begin{pmatrix} {}^{12}\!/_{39} & {}^{6}\!/_{39} \\ {}^{7}\!/_{26} & {}^{7}\!/_{26} \end{pmatrix}$$

“Distribute supplemental margin into table using row probabilities from complete cases.”

Monotone Multivariate Normal Data

Regress current on more observed variables using available cases; e.g. 3 variables:



Marginal of Y_1

Regression Y_2 on Y_1

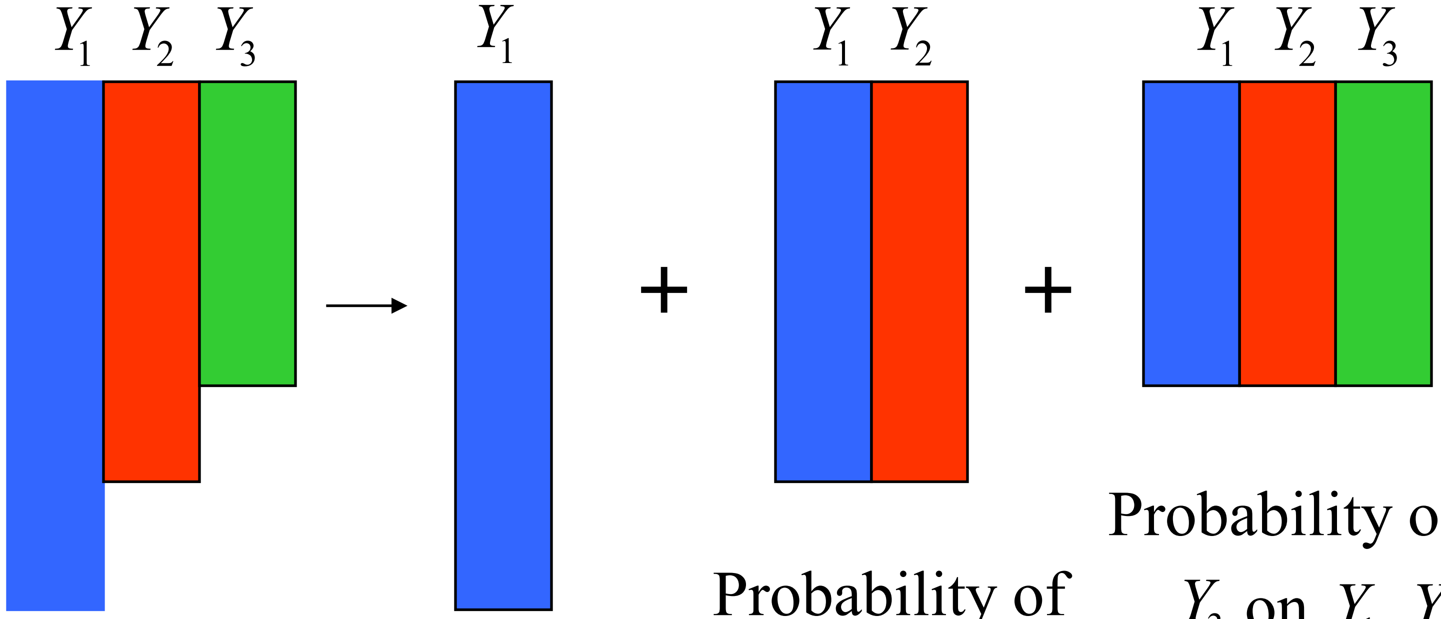
Regression Y_3 on Y_1 Y_2

Sweep operator is elegant way to do computations

Monotone Multinomial Data

Compute conditional probabilities using available cases;

e.g. 3 variables:

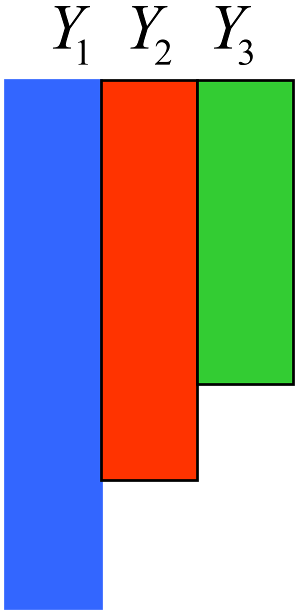


Marginal of Y_1

Probability of Y_2 on Y_1

Probability of Y_3 on Y_1 Y_2

Multiple imputation for monotone data



- Factoring joint distribution is also useful for multiple imputation. E.g. for the monotone pattern displayed:

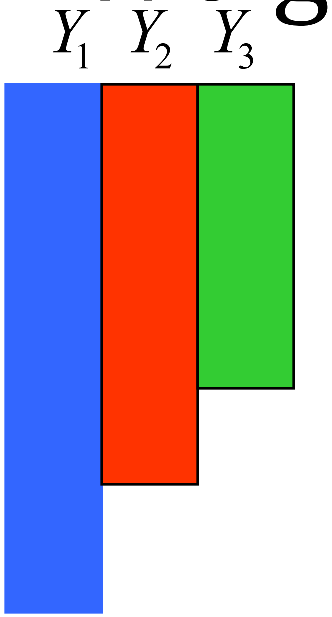
(a) Draw missing values of Y_2 from predictive distribution of Y_2 given Y_1

(b) Draw missing values of Y_3 from predictive distribution of Y_3 given Y_1 and Y_2 (observed or imputed from (a))

(c) Repeat steps (a) and (b) to create MI data sets

Extends in obvious way to more than three patterns.

Weights for monotone data



Factoring joint distribution of M is also useful for creating propensity weights. E.g. for the monotone pattern displayed:

(a) Propensity for M_2 from logistic regression of M_2 given Y_1

(b) Propensity for M_3 given M_2=0 from logistic regression of M_3 given Y_1 and Y_2

(c) $\Pr(M_3 = 0) = \Pr(M_3 = M_2 = 0)$

$= \Pr(M_3 = 0 \mid M_2 = 0) \times \Pr(M_2 = 0)$

Inverting estimated propensities, weight for (c) is product of weights from (a) and (b)

Extends in obvious way to more than three patterns.

Summary

- Factoring distribution of Y or M into conditional distributions is useful for monotone missing data

• Modeling Y using factored likelihood requires distinctness of parameters of distribution. E.g. it does not work if Y is bivariate normal with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix},$$

since parameters ϕ_1, ϕ_2 are no longer distinct

• In such cases, and when the pattern is non-monotone, iterative algorithms such as EM are needed for ML estimation. Consider these approaches next.