Cross-Sector Summer Research in Residence at NISS

Nell Sedransk,
National Institute of Statistical Sciences
22 April 2010

Impetus for NISS-NASS Program

- NASS
  - Critical, complex problems
    - Sophisticated (but practical) problem solutions
    - Research requiring varied, specialized technical expertise
    - Immediate implementation
  - Limited statistical research base within agency
    - Postdoctoral training in agriculture survey context
    - Embedded graduate students as potential employees
    - Liaison to statistics research faculty
- NISS
  - Connected to academia (University Affiliates)
  - Active NISS postdoctoral fellows program

Paradigm

NASS	Academia
Nonstandard Problems	Specialized Technical Expertise
Data	Research Manpower
Contextual Experience	Research Mentoring Experience

TEAM

NISS

In residence at NISS   Summer 2009   NASS-Academic Team

September 2009 - Postdoc at NASS - May 2010

In residence at NISS   Summer 2010   NASS-Academic Team

September 2010 - Postdoc at NASS - May 2011

CONFERENCE

Three Teams

TEAM	FACULTY	POSTDOC	STUDENT	NASS
I	S.Ghosh	M. Robbins	J. Habiger	K White
	B. Goodwin			D. Miller
II	L. Young	H. Sang	K. Lopiano	D. Abreu
	P. Arroway			A. Lamas
III	B. Nandram	J. Wang	C. Toto	E. Anderson
	S. Holan			W. Barboza

Three Survey-based Problems

- I: ARMS (Agriculture Resources Management Survey) – both NASS and ERS (Economic Research Service)
  - Microdata analysis
- II: June Area Survey of Small Farms & (5-year) Census of Agriculture
  - Coherent estimation of number of small farms
- III: AYS (Agricultural Yield Survey) & DAS (December Agricultural Survey) & OYS (Objective Yield Survey)
  - Prediction with variance estimates

Common Threads

- Multiple Data Sources
  - Different sampling frames
  - Different sample designs
  - Different sources of variation
  - Different sources of bias
- Imputation
- Macro to Micro
  - Estimation of totals – multiplicative factor
  - Estimation for small areas, “disagreggation”
  - Analysis of covariation and microdata analysis
- Technology and opportunities
  - Access to multiple sources including covariates
  - Advances in software – to replace expert opinion

ARMS: Imputation for Item Nonresponse

- ARMS: Comprehensive survey
  - 100s of items with 10s of required items
  - $\Rightarrow$ high rate of item nonresponse
- Conditional mean imputation*
  - Classification by 3 factors: $$, farm type, region
  - Disrupts joint distribution structure
    - Covariance structure
  - Disrupts marginal distribution structure
    - Skewed distribution for much economic data
  - Underestimates variances
    - For tested factors: underestimates std dev by up to 50%

*: with restrictions: donor pool size > 10; extreme values excluded from pool

Objective: Preserve Data Structure

- Goals
  - Analysis of microdata
    - Example: relationship of two highly skewed variables
  - Variance estimation
- Imputation Approaches
  - MCMC
  - EM
  - Data augmentation
    - Good representation of joint distribution
      - Allows random draws from joint distribution
      - If parametric, permits transformation

        (e.g., log transformation of data

        $\Rightarrow \Rightarrow$ skew-normal distribution)

Joint Distribution Construction

- Sequential procedure
  - Transform data to use (skew)normal theory
    - Continuous economic data – log transformation
    - Discrete and mixed data – see paper*
  - Fit sequentially expanded subsets of data
    - Initiate with maximal set of variables & maximal set of complete observations
    - Expand set of observations: Impute by random draw from posterior distribution of missing data *given* observed data
    - Recompute posterior distribution
    - Iterate
  - Apply inverse transformation to imputed data values

See Schafer (19970, Little & Rubin (2002), Robbins (2009)

Method Performance: ARMS Data

- Commodity payments & Farm income
  - Highly skewed distributions
  - Separate models
  - Random item response deletion
- Results
  - Improved estimated distribution tails
  - Improved variance estimates
  - Good covariance estimates
- Next Step – Method robustness
  - Missing at random from simple pattern

AYS, DAS &OYS: Composite Prediction

- Forecasting: from planting to harvest
- Current practice
  - Expert panel review of data, ancillary information
- Objectives
  - Estimates (predictions) with stated precision
  - Variance quantified by source

Paradigm

			Planting					Harvest
Year 1	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒
Year 2	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒
Year 3	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒

Modeling Goals

- Hierarchical Bayes Model
  - Prediction with quantified variance
  - Multiple repeated surveys
  - Model for complex structure
  - Priors for parameters
  - Model comparisons
  - Forecast comparisons – actual data

Structure: Survey Level (time series)

Planting Harvest

Year 1
Survey #1 farmer opinion sample ⇒ ⇒ ⇒ ⇒
Survey #2 farmer opinion large sample
Survey #3 measurement sample subsample ⇒ ⇒ ⇒ ⇒

Year 2
Survey #1 farmer opinion sample ⇒ ⇒ ⇒ ⇒
Survey #2 farmer opinion large sample
Survey #3 measurement sample subsample ⇒ ⇒ ⇒ ⇒

Year 3
Survey #1 farmer opinion sample ⇒ ⇒ ⇒ ⇒
Survey #2 farmer opinion large sample
Survey #3 measurement sample subsample ⇒ ⇒ ⇒ ⇒

Process Model

Structure: Historical Series

			Planting					Harvest
Year 1	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒
Year 2	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒
Year 3	Survey #1	farmer opinion	sample	⇒	⇒	⇒	⇒	
	Survey #2	farmer opinion						large sample
	Survey #3	measurement	sample	subsample	⇒	⇒	⇒	⇒

Hierarchical Model Approach

- Stage 1: Data Model
  - {Survey Data | True yield Θ_d}
- Stage 2: Process Model
  - { True yield | Φ_p }
- Stage 3: Parameter Model
  - {Θ_d, Φ_p }
- Posterior for process & parameters | Survey data
  - {True yield, Θ_d, Φ_p} |
    $\propto$ {Data$_{\#3}$|True yield, Θ_d} {Data$_{\#1}$|True yield, Θ_d}
    {Data$_{\#2}$|True yield, Θ_d} {True yield| Φ_p} {Θ_d, Φ_p }

Hierarchical Model

- Data Model {Survey Data | True yield Θ_d}
  - [$Data_{\#1,}$ $Data_{\#2}$] AR(1)
  - $Data_{\#3}$ AR(1)
  - Conditionally independent
- Survey Biases
  - Bias parameters {$B_{\#1}$, $B_{\#2}$}
  - Independent forecasting errors
- Latent Process Model
  - Regression
    - Location/Region specific factor values
    - Weather
    - Crop progress
    - Interactions
- Prior Distributions

Model Performance

- Example:NASS survey data for corn yield
- Survey biases
  - Non-ignorable
  - Consistent across years
  - AR (1) – good fit to data
  - Survey #2 “close” to True Yield
- Bayesian Hierarchical Model
  - Outperforms other composite estimators