SELECTIVE EDITING AS A STOCHASTIC OPTIMIZATION PROBLEM

Ignacio Arbués[1], Pedro Revilla, Soledad Saldaña

Instituto Nacional de Estadística

May 2011

[1]Instituto Nacional de Estadística, Castellana, 183. 28071, Madrid, Spain.
iarbues@ine.es

Structure

1. From a heuristic method to a formal one
2. The optimization problem
3. Case studies
4. Final Remarks

Heuristic method

Let $\tilde{x}_{ij}$ be the value of variable j reported by unit i.

We have a prediction $\hat{x}_{ij}$ (obtained using time series models), with variance ν_{ij}^2.

$$s_{ij} = \frac{|\tilde{x}_{ij} - \hat{x}_{ij}|}{\nu_{ij}} \tag{1}$$

Units with large s_{ij} are edited.

Program to formalization

We know that the method works, but this raises some questions:

- What does 'works' mean? (How we measure that?).
- Why does it work?

By giving formal answers to these questions we will arrive to a formal method and we will see the advantages of this.

What does 'works' mean? (I)

Observation error: $\tilde{x}_{ij} = x_{ij} + \varepsilon_{ij}$.

$$R_i = \begin{cases} 0 & \text{if } i \text{ is edited} \\ 1 & \text{if } i \text{ is not edited} \end{cases} \tag{2}$$

True (unknown) $X_j = \sum_i w_{ij} x_{ij}$ and edited $X_j^{ed} = \sum_i w_{ij}(x_{ij} + R_i \varepsilon_{ij})$ aggregates.

What does 'works' mean? (II)

We want (A) to limit the error of the aggregate:

$$\mathbb{E}[(X_j^{ed} - X_j)^2] = \mathbb{E}\Big[\Big(\sum_i w_{ij} R_i \varepsilon_{ij}\Big)^2\Big] \leq m_j \tag{3}$$

If we neglect the cross-products, we get a linear function $\sum_i w_{ij}^2 \varepsilon_{ij}^2 R_i$.

We want (B) to reduce the workload, that is to make $\sum_i R_i$ as large as possible.

Theoretical framework for selective editing

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.
- There is a $\sigma-$field $\mathcal{G} \subset \mathcal{F}$ representing the available information ($\tilde{x}_{ij}$, previous periods data, auxiliary information, ...).

Definition

A selection strategy (SS) is a $\mathcal{G}-$measurable random vector $R = (R_1, \ldots, R_N)$, such that $R_i \in [0, 1]$.

SE as a stochastic optimization problem

Problem (P)

$$\begin{array}{ll} \max_R & \mathbb{E}[\sum_i R_i] \\ R \in S(\mathcal{G}) & \mathbb{E}[\left(\sum_i w_{ij} R_i \varepsilon_{ij}\right)^2] \leq m_j, j = 1, \ldots, q \end{array}$$

If we neglect again the cross-products, we get a linear version of the problem. In vector notation:

Problem (P_L)

$$\begin{array}{ll} \max_R & \mathbb{E}[\mathbf{1}^T R] \\ R \in S(\mathcal{G}) & \mathbb{E}[DR] \leq m \end{array}$$

Step 1: Total Expectation Law

I., Arbués, M. González and P. Revilla, *Optimization*, (2010).

We apply the TEL and the fact that R is $\mathcal{G}-$measurable.

Problem (P_L)

$$\begin{array}{ll} \max_R & \mathbb{E}[\mathbf{1}^T R] \\ R \in S(\mathcal{G}) & \mathbb{E}[\Delta R] \leq m \end{array} \quad \textit{where } \Delta = \mathbb{E}[D|\mathcal{G}].$$

(Δ contains the conditional 2nd-order moments of the errors)

Step 2: Duality

Under some assumptionsthe original problem is equivalent to

Problem (D_L)

$$\min_{\lambda \geq 0} \max_{R} \quad \mathbb{E}[\mathbf{1}^T R - \lambda^T (\Delta R - m)].$$

This change is advantageous when the unconstrained maximization with respect to R is much easier than the original problem.

Step 3: Interchangeability Principle

By the IP, under some assumptions,

Problem (D_L)

$$\min_{\lambda \geq 0} \max_R \quad \mathbb{E}[\mathbf{1}^T R - \lambda^T (\Delta R - m)]$$

is equivalent to

Problem (D_L)

$$\min_{\lambda \geq 0} \quad \mathbb{E}\Big[\max_R \left\{ \mathbf{1}^T R - \lambda^T (\Delta R - m) \right\} \Big].$$

We have now a simple deterministic problem inside $\mathbb{E}\Big[\ldots \Big]$.

Step 4: Sample Average Approximation (SAA)

We can replace the expectation by sample average and minimize.

Problem $(\hat{D}_L)$

$$\min_{\lambda \geq 0} \quad \frac{1}{M} \sum_{\ell=1}^{M} \left[\max_R \left\{ \mathbf{1}^T R - \lambda^T (\Delta^{(\ell)} R - m) \right\} \right]$$

The number of multipliers is typically moderate, so the minimization with respect to λ is feasible by numerical methods.

Why does the heuristic method work?

When there is only one variable: edit the units with the largest Δ_i.

This reminds of the heuristic method: has $s_i = |\tilde{x}_i - \hat{x}_i|/\nu_i$ something to do with Δ_i?

$$\tilde{x}_i - \hat{x}_i = \tilde{x}_i - x_i + x_i - \hat{x}_i = \varepsilon_i + \xi_i \tag{4}$$

There is an observation error or not. Which one explains better the difference?

Conditional moments (I)

We make explicit assumptions on the behavior of the observation ε_i and prediction ξ_i errors.

Observation model:

1. $\varepsilon_i = \eta_i e_i$ where e_i is a Bernoulli that takes the values 1 and 0 with probabilities p and $1 - p$.
2. η_i is a normally-distributed, zero-mean variable with variance σ^2.

Prediction model:

1. ξ_i is a normally-distributed, zero-mean variable with variance ν_i^2.

Observation and prediction errors are independent.

Conditional moments (II)

Proposition

If we put $u_i = \hat{x}_i - \tilde{x}_i$,

$$E[\varepsilon_i|\mathcal{G}] = \frac{\sigma^2}{\sigma^2+\nu_i^2} u_i \zeta_i \tag{5}$$

$$E[\varepsilon_i^2|\mathcal{G}] = \left[\frac{\sigma^2\nu_i^2}{\sigma^2+\nu_i^2} + \left(\frac{\sigma^2}{\sigma^2+\nu_i^2}\right)^2 u_i^2\right]\zeta_i,$$

where

$$\zeta_i = \frac{1}{1+\frac{1-p}{p}\left(\frac{\nu_i^2}{\sigma^2+\nu_i^2}\right)^{-1/2}\exp\{-\frac{u_i^2\sigma^2}{2\nu_i^2(\sigma^2+\nu_i^2)}\}}. \tag{6}$$

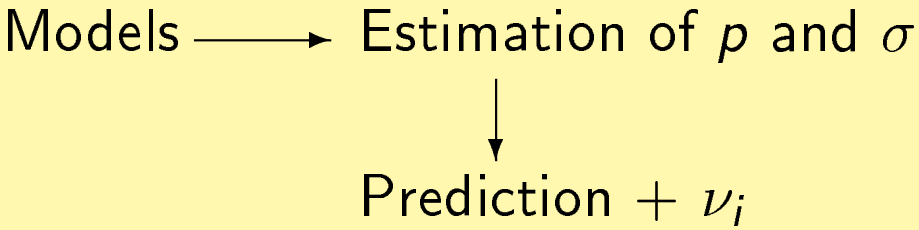
Implementation scheme

Models

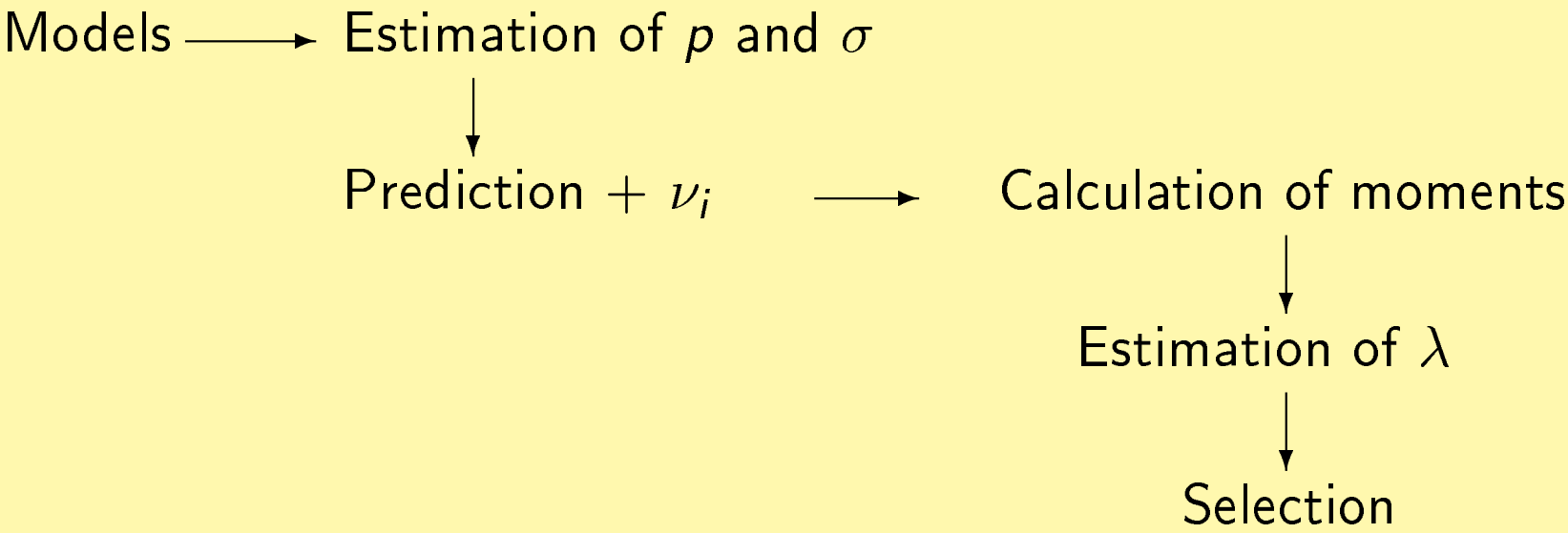
Implementation scheme

Models $\longrightarrow$ Estimation of p and σ

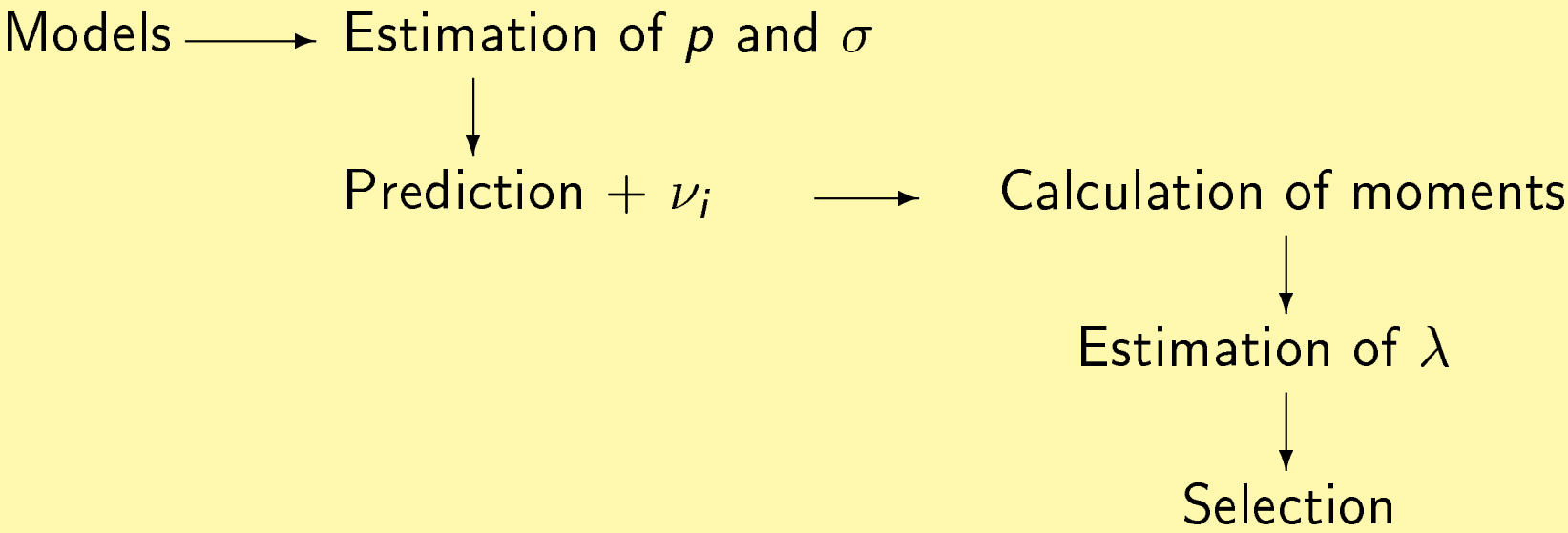
Implementation scheme



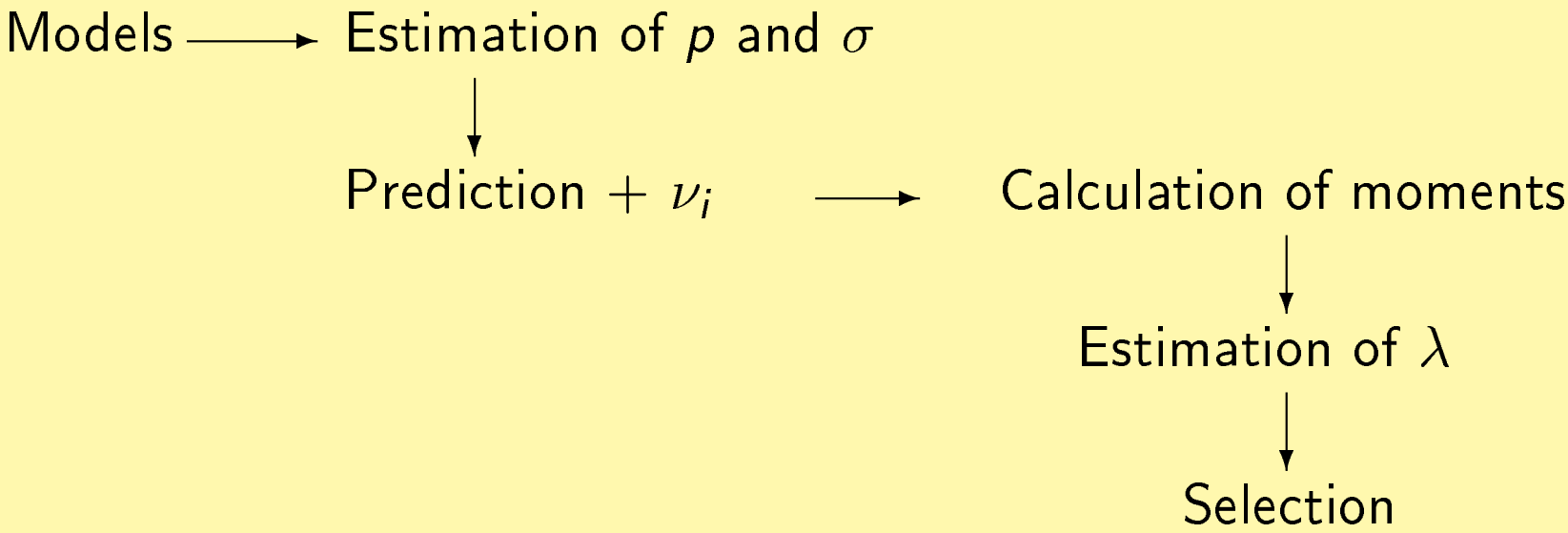
Implementation scheme



Implementation scheme



Implementation scheme



Observation model + Prediction model $\rightarrow$ Selection

Case 1: short term indicator

New Orders/Turnover Survey.

- $N \approx 13.500$.
- Main variables New orders, Turnover.
- Monthly survey.

Model

Available data: previous periods data from the same survey.

Model:

(i) Variables x_{ijt} and $x_{i'j't'}$ are independent if $(i,j) \neq (i',j')$.

(ii) Processes $\{x_{ijt}\}_t$ satisfy a model among

a) $(1-B)y_{ijt} = a_t$
b) $(1-B^{12})y_{ijt} = a_t$
c) $(1-B^{12})(1-B)y_{ijt} = a_t$

where $y = \log(x+c)$ y a_t is a Gaussian white noise.

Results: Error bounds in the quadratic version

m	e_1/m	e_2/m	n	m	e_1/m	e_2/m	n
0.0250	1.86	2.50	578.4	0.1742	1.24	0.92	77.8
0.0304	1.46	2.02	605.9	0.2116	0.99	0.76	59.8
0.0369	1.42	1.82	401.9	0.2569	0.92	0.67	35.4
0.0448	0.95	1.35	477.3	0.3120	0.71	0.56	27.7
0.0544	0.96	1.17	390.5	0.3788	0.64	0.66	28.0
0.0660	1.41	1.28	278.8	0.4600	0.71	0.59	13.3
0.0801	1.31	1.30	283.4	0.5585	0.56	0.57	6.8
0.0973	0.99	0.84	225.8	0.6782	0.46	0.47	6.8
0.1182	1.32	1.22	140.3	0.8235	0.39	0.38	2.9
0.1435	1.30	0.95	93.1	1.0000	0.32	0.46	1.6

Table: Error bounds in the quadratic version (M=6).

Case 2: cross-sectional data

Agricultural census, $q = 186$ variables. Experiment with one NUTS-3 unit.

Observation models:

(1) As in the previous case.

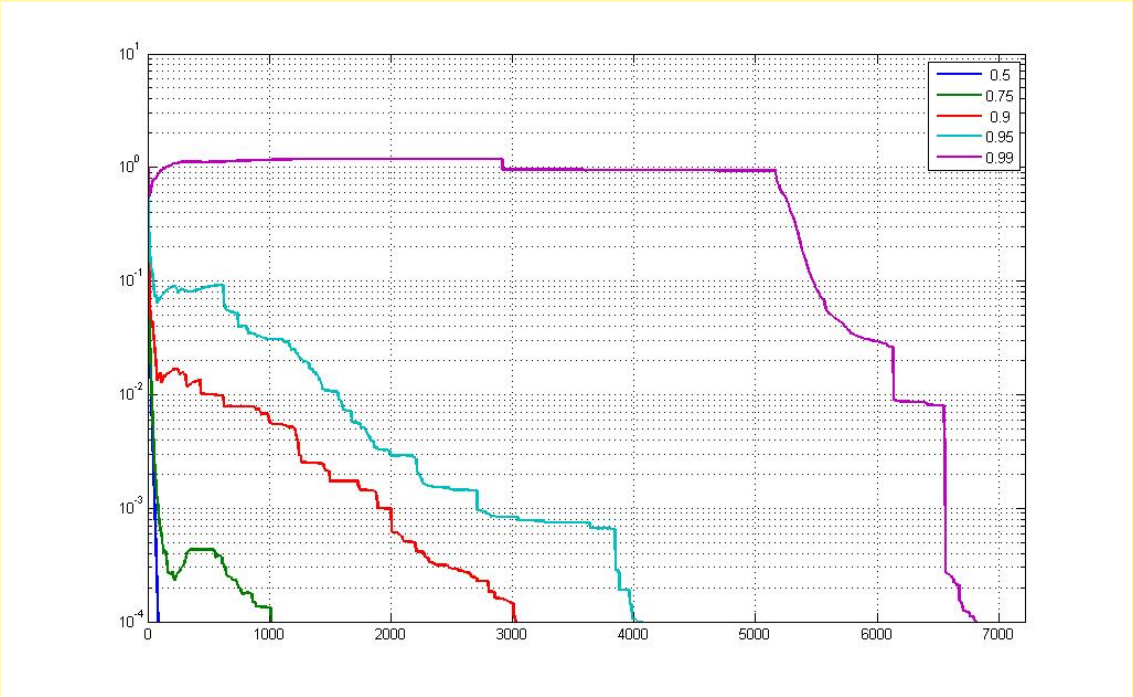
(2) Mixture of normal-distributed errors and a distribution concentrated at zero.

Prediction models:

(1) $y = X'\beta + \xi$ with $\xi \sim N(0, \nu^2)$.

(2) $y = (1 - Z)(X'\beta + \xi)$ with $\xi \sim N(0, \nu^2), Z \sim B(\mathrm{LOGIT}(X'\gamma))$.

Results



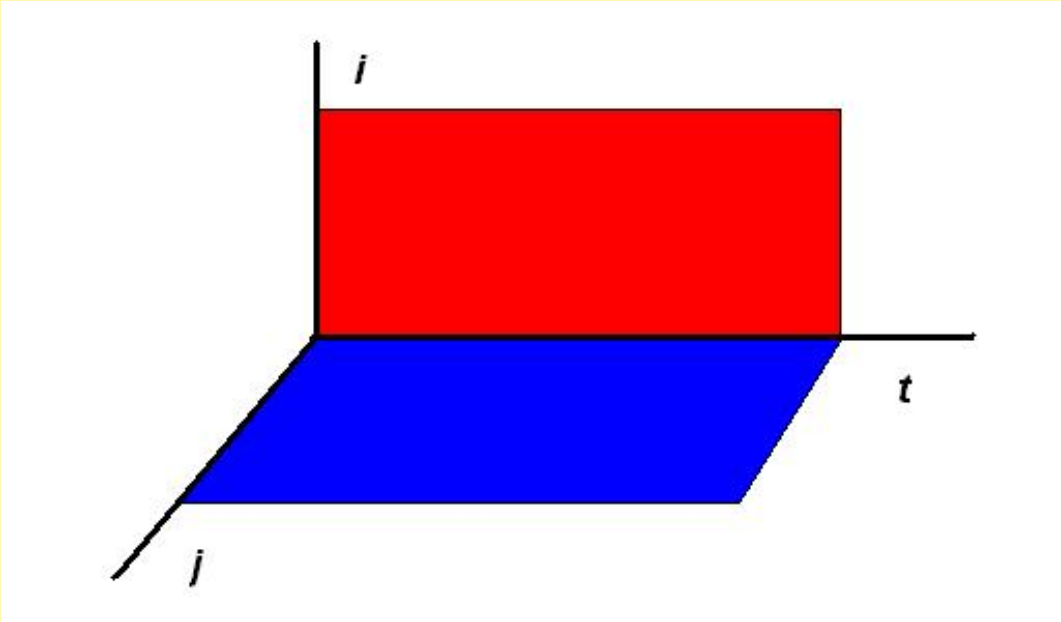
The semicontinuous models (2) do not outperform the linear ones.

Advantages of the formal method

- It uses all the information in an efficient way.
- The underlying assumptions have been made explicit and can be checked by the standard methods of statistical inference.
- The theoretical framework provides a guide to adapt the method to new situations.
  - Only prediction and observation models have to be adapted.
  - Well-known statistical tools exist to build models for the data.

Current and future developments

- Relax the normality assumption: semicontinuous models.
- Integrate all dimensions.



- Imputation.

The Optimal Selection Strategy of the Linear problem.

To maximize $\mathbf{1}^T R - \lambda^T(\Delta^{(\ell)} R - m)$ with respect to R is easy.

$$R_i = \begin{cases} 1 & \text{if} \quad \lambda^T \Delta_i < 1 \\ 0 & \text{if} \quad \lambda^T \Delta_i > 1 \end{cases} \tag{7}$$

where $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{iq})^T$.

(back)