

# Inverse problem for interacting models in social sciences

Cecilia Vernia

<sup>1</sup>Dipartimento di Matematica Pura ed Applicata  
Università degli studi di Modena e Reggio Emilia

Eindhoven, 18 May, 2011

## Collaboration

Joint work with

- Pierluigi Contucci, Micaela Fedele (University of Bologna)
- Vincenzo Coscia (University of Ferrara)
- Cristian Giardiná (University of Modena and Reggio Emilia)
- Raffaella Burioni, Elena Agliari (University of Parma)

# Topics Outline

- 1 Introduction
- 2 Discrete Choice Theory (DCT)
  - Description
  - Model
  - Inverse Problem
- 3 Interacting models
  - Description
- 4 Inverse Problem
  - One Homogeneous population
  - Multi-species Curie-Weiss Model
  - Data Test
  - Work in progress

# Mathematics and social sciences

- Description of social phenomena on mathematical grounds
- Mathematical framework for studying collective human behavior (by integrating econometric tools with mathematical techniques derived from statistical mechanics)
- Problem: evaluate how people respond to screening invitations
- Estimate the relative importance of different factors in making a choice (peer-to-peer influences, cultural heritage, public information campaigns,...)

Starting points: Discrete Choice Theory (DCT) of Mc Fadden,  
socio-physical approach (Serge Galam)

## Contribution of Mc Fadden

- A celebrated example of a predictive theory in the sense of the hard sciences within social and economic sciences;
- Purpose: description of people's behavior
- It is an econometric technique to infer people's preferences based on empirical data
- The decision maker is assumed to make choices that maximize its own benefit.
- The model does not account for peer-pressure or herding effects

The performance of discrete choice model is close to optimal for the analysis of many phenomena where peer influence is not a major factor in an individual's decision

# Results of Mc Fadden's study

DCT predicts how many customers would have used the new transportation system (BART) in San Francisco

**Table 1. Prediction Success Table, Journey-to-Work**  
 (Pre-BART Model and Post-BART Choices)

Cell Counts	Predicted Choices				
Actual Choices	Auto Alone	Carpool	Bus	BART	Total
Auto Alone	255.1	79.1	28.5	15.2	378
Carpool	74.7	37.7	15.7	8.9	137
Bus	12.8	16.5	42.9	4.7	77
BART	9.8	11.1	6.9	11.2	39
<b>Total</b>	<b>352.4</b>	<b>144.5</b>	<b>94.0</b>	<b>40.0</b>	<b>631</b>

Predicted Share	55.8%	22.9%	14.9%	6.3%
(Std. Error)	(11.4%)	(10.7%)	(3.7%)	(2.5%)
Actual Share	59.9%	21.7%	12.2%	6.2%

(Source: McFadden 2001)

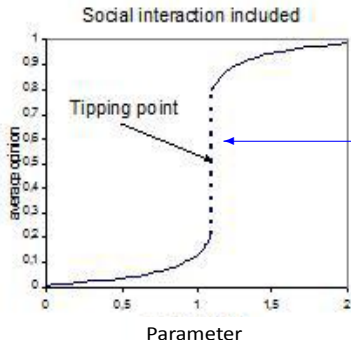
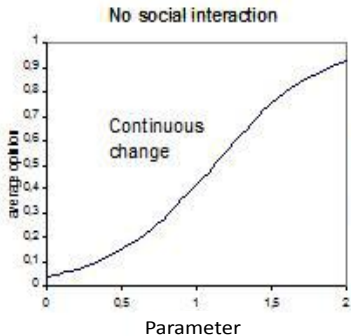
98%  
 agreement

## Reasons for DCT successes

- Method to quantify how potential customers weigh different factors in making social-economic decisions (regarding transportation: we tend to choose according to principles of quickness, low cost, comfort...)
- Good method for situations in which individuals can directly experience the costs and benefits of different choices (peer pressure and herding effects are small).
- From a theoretical physics perspective, DCT corresponds to a mixture of non interacting perfect gases, which is a solvable mathematical problem.
- DCT fails in situations in which choices are basically driven by peer-interaction effects.

# Tipping Point

interactions between individuals cause predicted behavior to depend discontinuously on parameter values.



riots,  
bank runs,  
market crashes,  
wars,...



## Model of McFadden

Standard DCT



non interacting spin system in  
statistical mechanics

Hamiltonian or Cost function

$$H_N(\sigma) = - \sum_{i=1}^N h_i \sigma_i$$

$$\sigma_i = \begin{cases} +1 & \text{if } i \text{ says YES} \\ -1 & \text{if } i \text{ says NO} \end{cases}$$

$h_i$  external vector field,  $i = 1, \dots, N$

## Parametrization of the field $h_i$

Assign to each person  $i$  a vector of socio-economic attributes

$$a_i = \{a_i^{(1)}, a_i^{(2)}, \dots, a_i^{(k)}\}$$

deduced from surveys, polls, censuses,....

$$a_i^{(1)} = \begin{cases} 1 & \text{for } i \text{ Male} \\ 0 & \text{for } i \text{ Female} \end{cases}, \quad a_i^{(2)} = \begin{cases} 1 & \text{for } i \text{ Employee} \\ 0 & \text{for } i \text{ Self-employed} \end{cases}, \dots$$


$$h_i = \sum_{j=1}^k h^{(j)} a_i^{(j)} + h^{(0)}$$

model's parameters  $(h^{(0)}, h^{(1)}, \dots, h^{(k)})$  independent of individual  $i$



partition of the population into  $n = 2^k$  disjoint groups  $P_\ell$ ,  
 $\ell = 1, \dots, n$

# The joint distribution

Indeterminacy  set of random equilibrium states

## Boltzmann-Gibbs distribution

$$P\{\boldsymbol{\sigma}\} = \frac{e^{-H_N(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma} \in \Omega_N} e^{-H_N(\boldsymbol{\sigma})}}$$

where  $\Omega_N = \{-1, 1\}^N$  and  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)$  is the spin (choices) configuration

$\langle \cdot \rangle_{BG}$  expectation value with respect to the Boltzmann-Gibbs measure

## Solution of the model

magnetization of a set  $P_\ell$  of spins = average opinion of a group  $P_\ell$

$$m_\ell(\boldsymbol{\sigma}) = \frac{1}{|P_\ell|} \sum_{i \in P_\ell} \sigma_i$$

In the thermodynamic limit:

$$\lim_{N \rightarrow \infty} \langle m_\ell(\boldsymbol{\sigma}) \rangle_{BG} = m_\ell(\mathbf{h}) \quad \ell = 1, \dots, n$$

where  $(m_1(\mathbf{h}), \dots, m_n(\mathbf{h}))$  is a solution of the system

$$\begin{cases} m_1(\mathbf{h}) = \tanh(h_1) \\ \vdots \\ m_n(\mathbf{h}) = \tanh(h_n). \end{cases}$$

# Inverse Problem

## Parameter Estimation

Starting from empirical data calculate the model parameter  $\mathbf{h}$

$$h_\ell = \tanh^{-1}(m_\ell) \quad \ell = 1, \dots, n$$

$m_\ell$  can be estimated from empirical data

The parameters  $h_\ell$  tell us the relative importance of the various socio-economic factors in people's decision making



use statistics in order to find the value of the  $h_\ell$ 's for which our distribution best fits the real data

# Strength and Weakness of DCT

- DCT has been used successfully for the last thirty years
- DCT can only be applied when the functional relation between the people's attributes and the population's behavior is a smooth one.

but...

behavior at a collective level can be marked by sudden jumps

From statistical mechanics point of view: the reason for these abrupt changes may be looked for in **phase transitions** caused by the **interactions** between individuals, not covered by standard DCT

# Interacting Model

## Hamiltonian or Cost function

$$H_N(\sigma) = - \sum_{i,j=1}^N J_{ij} \sigma_i \sigma_j - \sum_{i=1}^N h_i \sigma_i$$

H favors

- the agreement of people's choices  $\sigma_i$  with some external influence  $h_i$
- the agreement between couples  $i$  and  $j$  of people who have positive interaction coefficient  $J_{ij}$

# $n$ interacting populations

$$\begin{matrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{matrix} \left\{ \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} \right.$$

$$\begin{matrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{matrix} \left\{ \begin{pmatrix} \overbrace{J_{11}}^{N_1} & \overbrace{J_{12}}^{N_2} & \dots & \overbrace{J_{1n}}^{N_n} \\ J_{12} & J_{22} & \dots & J_{2n} \\ \vdots & \vdots & & \vdots \\ J_{1n} & J_{2n} & \dots & J_{nn} \end{pmatrix} \right.$$

$n$  disjoint groups

$P_1, \dots, P_n, |P_\ell| = N_\ell,$

where  $\sum_{\ell=1}^n N_\ell = N,$

relative group size

$\alpha_\ell = N_\ell/N$

$J_{\ell s}$  is a matrix with constant elements  $J_{\ell s}.$

$J_{\ell s}$  tunes the interaction between an individual of the group  $P_\ell$  and one of the group  $P_s$



# Multi-species Curie-Weiss Model

Average opinion of a group  $P_\ell$

$$m_\ell(\boldsymbol{\sigma}) = \frac{1}{N_\ell} \sum_{i \in P_\ell} \sigma_i$$

The Hamiltonian becomes

$$H_N(\boldsymbol{\sigma}) = -N \left( \frac{1}{2} \sum_{\ell, s=1}^n \alpha_\ell \alpha_s J_{\ell s} m_\ell(\boldsymbol{\sigma}) m_s(\boldsymbol{\sigma}) + \sum_{\ell=1}^n \alpha_\ell h_\ell m_\ell(\boldsymbol{\sigma}) \right)$$

and the distribution

$$P_{N, \mathbf{J}, \mathbf{h}}\{\boldsymbol{\sigma}\} = \frac{\exp(-H_N(\boldsymbol{\sigma}))}{\sum_{\boldsymbol{\sigma} \in \Omega_N} \exp(-H_N(\boldsymbol{\sigma}))}$$

## Solution of the model

In the thermodynamic limit:

$$\lim_{N \rightarrow \infty} \langle m_\ell(\boldsymbol{\sigma}) \rangle_{BG} = m_\ell(\mathbf{J}, \mathbf{h}), \quad \ell = 1, \dots, n$$

where  $(m_1(\mathbf{J}, \mathbf{h}), \dots, m_n(\mathbf{J}, \mathbf{h}))$  is a solution of the  $n$  mean field equations (Contucci, Gallo, Ghirlanda):

$$\begin{cases} m_1(\mathbf{J}, \mathbf{h}) &= \tanh(\sum_{s=1}^n \alpha_s J_{1s} m_s(\mathbf{J}, \mathbf{h}) + h_1) \\ m_2(\mathbf{J}, \mathbf{h}) &= \tanh(\sum_{s=1}^n \alpha_s J_{2s} m_s(\mathbf{J}, \mathbf{h}) + h_2) \\ \vdots & \\ m_n(\mathbf{J}, \mathbf{h}) &= \tanh(\sum_{s=1}^n \alpha_s J_{ns} m_s(\mathbf{J}, \mathbf{h}) + h_n) \end{cases}$$

# One Homogeneous population (Curie-Weiss model)

## Hamiltonian or Cost function

$$H_N(\boldsymbol{\sigma}) = -\frac{J}{2N} \sum_{i,j=1}^N \sigma_i \sigma_j - h \sum_{i=1}^N \sigma_i$$

In the thermodynamic limit:

$$\lim_{N \rightarrow \infty} \langle m_N(\boldsymbol{\sigma}) \rangle_{BG} = m(J, h),$$

where  $m(J, h)$  is a solution of the mean field equation

$$m(J, h) = \tanh(Jm(J, h) + h)$$

Differentiating with respect to  $h$

$$\lim_{N \rightarrow \infty} \frac{\partial}{\partial h} \langle m_N(\boldsymbol{\sigma}) \rangle_{BG} = \chi$$

## One Homogeneous population: inverse problem

$$m = \tanh(Jm + h)$$



$$\chi = \frac{\partial m}{\partial h} = \frac{1 - m^2}{1 - J(1 - m^2)}$$

$$J = \frac{1}{1 - m^2} - \frac{1}{\chi},$$

$$h = \tanh^{-1}(m) - Jm.$$

Observing that

$$\frac{\partial}{\partial h} \langle m_N(\boldsymbol{\sigma}) \rangle_{BG} = N \left( \langle m_N^2(\boldsymbol{\sigma}) \rangle_{BG} - \langle m_N(\boldsymbol{\sigma}) \rangle_{BG}^2 \right) \simeq \chi,$$

to solve the inverse problem we need to estimate  $\langle m_N(\boldsymbol{\sigma}) \rangle_{BG}$  and  $\langle m_N^2(\boldsymbol{\sigma}) \rangle_{BG}$

# Inverse problem

## Parameter Estimation

Starting from empirical data calculate the model parameter  $J$   $h$

Suppose we have a sample of size  $M$  (i.e.  $M$  independent realizations)

$$\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(M)}$$

of the random vector  $\sigma = (\sigma_1, \dots, \sigma_N)$  from data.

To estimate  $m$  and  $\chi$  in the probability distribution from the sample we use (maximum likelihood)

$$\langle m_N(\sigma) \rangle_{BG} \simeq \frac{1}{M} \sum_{\ell=1}^M m_N(\sigma^{(\ell)}) \quad \langle m_N^2(\sigma) \rangle_{BG} \simeq \frac{1}{M} \sum_{\ell=1}^M m_N^2(\sigma^{(\ell)})$$

## One Homogeneous population: inverse problem

Using the estimators

$$\tilde{m} \text{ for } \langle m_N(\boldsymbol{\sigma}) \rangle_{BG} \text{ and } \tilde{\chi} \text{ for } N \left( \langle m_N^2(\boldsymbol{\sigma}) \rangle_{BG} - \langle m_N(\boldsymbol{\sigma}) \rangle_{BG}^2 \right)$$

we have

$$J = \frac{1}{1 - \tilde{m}^2} - \frac{1}{\tilde{\chi}}.$$

and

$$h = \tanh^{-1}(\tilde{m}) - J \tilde{m}$$

# Multi-species Curie-Weiss Model

$$H_N(\boldsymbol{\sigma}) = -N \left( \frac{1}{2} \sum_{\ell, s=1}^n \alpha_\ell \alpha_s J_{\ell s} m_\ell(\boldsymbol{\sigma}) m_s(\boldsymbol{\sigma}) + \sum_{\ell=1}^n \alpha_\ell h_\ell m_\ell(\boldsymbol{\sigma}) \right)$$

$$\lim_{N \rightarrow \infty} \langle m_\ell(\boldsymbol{\sigma}) \rangle_{BG} = m_\ell, \quad \lim_{N \rightarrow \infty} \frac{\partial}{\partial h_s} \langle m_\ell(\boldsymbol{\sigma}) \rangle_{BG} = \chi_{\ell s} \quad \ell, s = 1, \dots, n.$$

$$\mathbf{J} = (\mathbf{P}^{-1} - \boldsymbol{\chi}^{-1}) \mathbf{D}_\alpha^{-1}$$

where  $\mathbf{P} = \text{diag}\{1 - m_1^2, \dots, 1 - m_n^2\}$ ,  $\mathbf{D}_\alpha = \text{diag}\{\alpha_1, \dots, \alpha_n\}$ ,  $\boldsymbol{\chi}$  is the susceptibility matrix, whose elements are

$$\chi_{\ell s} = N_s \left( \langle m_\ell(\boldsymbol{\sigma}) m_s(\boldsymbol{\sigma}) \rangle_{BG} - \langle m_\ell(\boldsymbol{\sigma}) \rangle_{BG} \langle m_s(\boldsymbol{\sigma}) \rangle_{BG} \right)$$

$$h_\ell = \tanh^{-1}(m_\ell) - \sum_{s=1}^n \alpha_s J_{\ell s} m_s \quad \ell = 1, \dots, n.$$

## Data Test

- statistical data available since 2003, produced by Italian national health system in a screening campaign for the prevention of cervical cancer (district of Parma)
- women aged 25-65 invited to take test every three years
- data available: age, place of residence, date of first and successive invitations. health information,.....

This large set of empirical data is well suited for the study of the relative factors involved in the individual choice (role of the imitation behavior and of the external influence)



## Perspectives

- predictions about how to improve public information campaigns in order to maximize the number of women taking the test
- will it be more effective:
  - to pay off to engage "opinion leaders" (hubs)?
  - to target those demographic groups which responded less to previous campaigns?

By predicting changes in behavior following changes in campaign strategy, it would allow to maximize the return on investment for given budgets

## Open problems

how to choose:

- the suitable partition of the individuals into subgroups;
  - the individuals in any group should share some characteristics (e.g. age, place of residence...)
- a proper sampling procedure;
  - splitting the groups into suitable sub-sample capable to reproduce the independence of the elements in the sample
- the parametrization of the interactions  $J_{ij}$  and of the fields  $h_i$ ;
  - for  $h_i$ : the same parametrization used in DCT
  - for  $J_{ij}$ : let individuals interact according to their similarity ("similarity-attraction" Byrne '97, Grant '93, Michinov '02)

$$J_{ij} = \sum_{\ell=1}^n \beta_{\ell} a_i^{(\ell)} a_j^{(\ell)} + \beta_0$$

# Thank you!