Dynamic generalized structural equation modeling, with application to the effect of pollution on health

Dani Gamerman

Departamento de Métodos Estatísticos - IM

Universidade Federal do Rio de Janeiro

IMPS 2019 - Santiago, 16 July 2019

Based on work with...

Luigi Ippoliti

Pasquale Valentini

Content

- Introduction
  (Data)
- Model
  (Hierarchical levels)
  (Computation)
- Results
- Conclusions

1. Introduction

This talk is mainly concentrated on regression (effect of X on Y)

Hope: discuss a variety of issues associated with an environment of plenty of data (information)

Illustrated with examples:

1) effect of pollution on health

2) effect of anxiety on performance

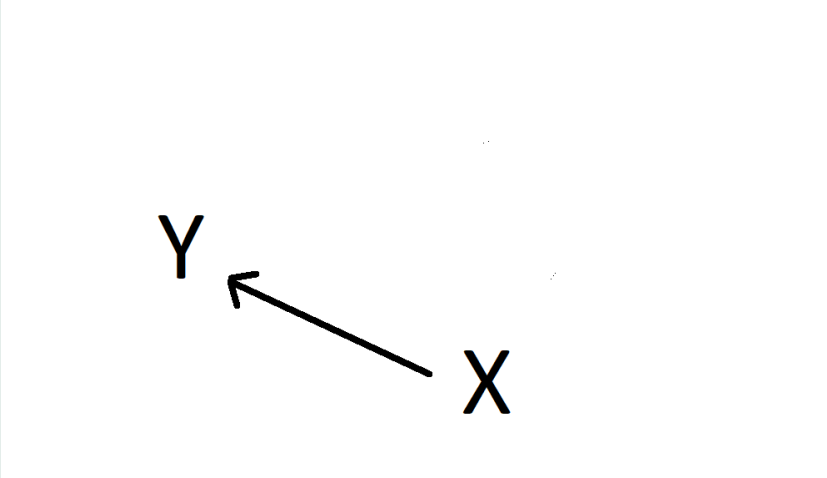
There are similarities and dissimilarites between examples

Issue #1: Lots of data

Plenty of observational units on both Examples 1 and 2

Plenty of responses for each observational unit

One can regress directly tons of data on tons of data



This is possible but unwise

It is best to stop and think before pushing buttons

Some of the possible problems:

Observational responses not best to represent intended response

Typically required response is some latent, underlying trait

Anxiety is unrecovered through observed responses to different stimuli

Health is uncovered through a variety of observed outcomes

Will overparametrize model thus removing significance

Many covariates are correlated

More parameters than observations

Removes focus and obliterates understanding

Too much information from the model will obscure the answer

Ideal is to have a few pointers to address the core of the effect

Some simplification/reduction must be performed... but with care

Aggregation over observational units is one such possibility

Example 2: aggregate over pollutant levels for each pollutant or aggregate over different pollutants

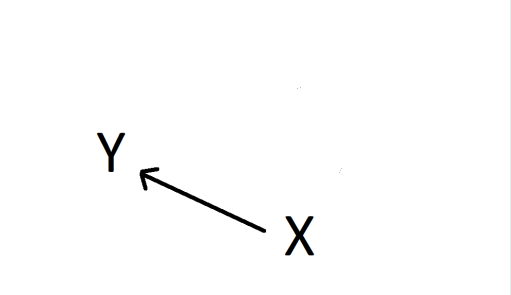
It may overlook relevant data features

It may oversmooth

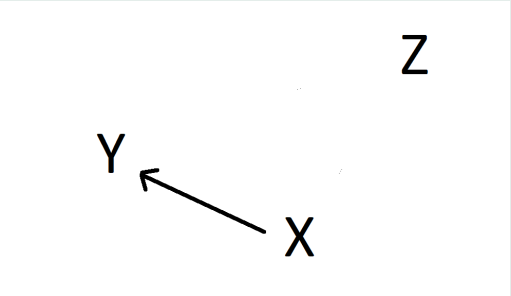
It seems best to use reduction driven by the data

Issue #2: Confounders

Ideal world:



Real world:



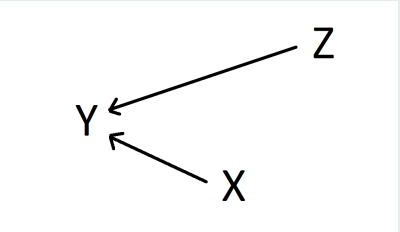
Sometimes confounder Z is at least as important as X to explain Y

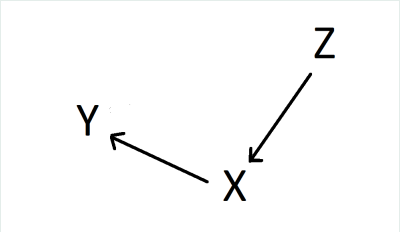
Example 1: Z = ability affects performance as much as anxiety

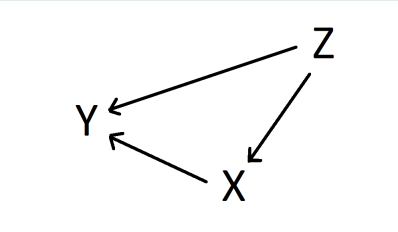
Example 2: Z = climate affects health as much as pollutions

These confounders MUST be included in the model

Many options available:







Confounders could be measured or unmeasured

Issue #3: Other sources of information

Importance of other sources is related to data pattern

1) Time

Are effects concomitant or delayed?

What is the temporal pattern of the effects?

Very important in Example 2 (pollution $\rightarrow$ health)

2) Space

Are variables geographically oriented?

If yes, they are likely to be spatially correlated

Relevant for Example 2 (pollution $\rightarrow$ health)

3) Hierarchy

Are there further data structures affecting outcomes?

Are there sub-groups (schools, ...) affecting effect of anxiety?

Relevant for Example 1 (anxiety $\rightarrow$ performance)

Knowledge of the relevant sub-group characteristics $\rightarrow$ confounders

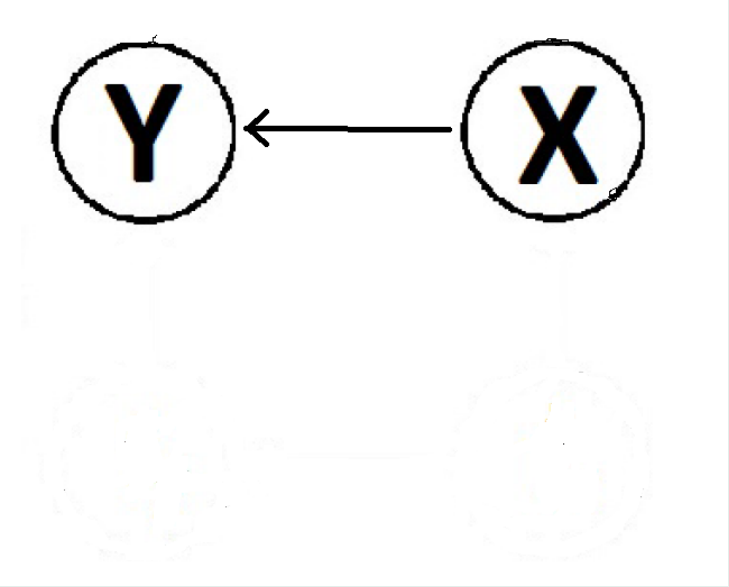
Otherwise, it is best to acknowledge the presence of the charateristics

Hierarchy can be elaborate (schools within cities within states ...)

Outlook of the talk

Data reduction is vital

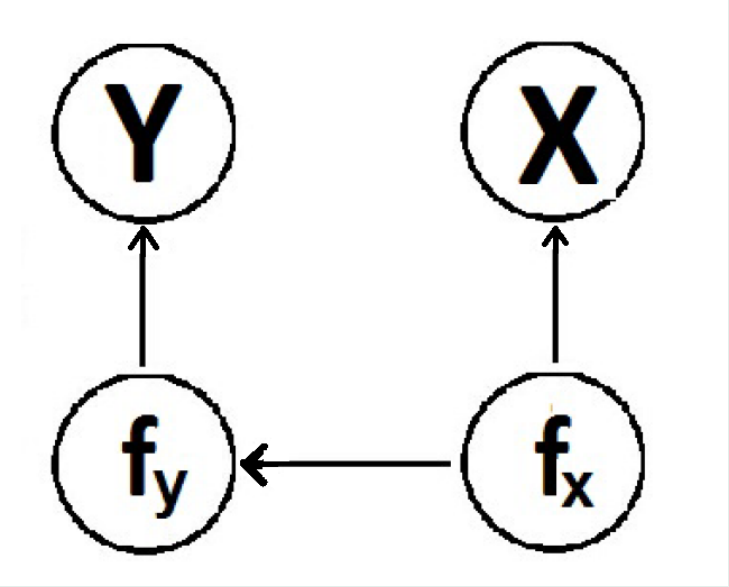
Main tool for reduction: factor analysis



Outlook of the talk

Data reduction is vital

Main tool for reduction: factor analysis



Crucial advantange: dimension $(f_X, f_Y) \ll$ dimension (X, Y)

Outlook of the talk (continued)

Set-up for factor analysis: Structural equation modeling (SEM)

Non-normal data: generalized SEM

Temporal pattern: dynamic SEM

Spatial component also incorporated (through factor loadings)

Next, data on pollution & health is presented

Important to understand observed patterns

After that, model is presented

Study of the association between **urban pollution and hospital admissions** in Lombardia and Piemonte (Italy) in 2011

Response (collected daily):

Y **hospital admissions counts** for respiratory and cardiovascular diseases ($n_y = 2$) at $N_y = 28$ districts for elderly population (age > 65)

Nature of variables: count data ($\rightarrow$ highly non-normal)

Spatial resolution: areal data

Other relevant variables (also collected daily):

X average pollutant concentration levels

Pollutants	CO	NO_2	PM_{10}	O_3
N_{x_j}	94	168	96	94

Nature of variables: asymmetric data ($\rightarrow$ also non-normal)

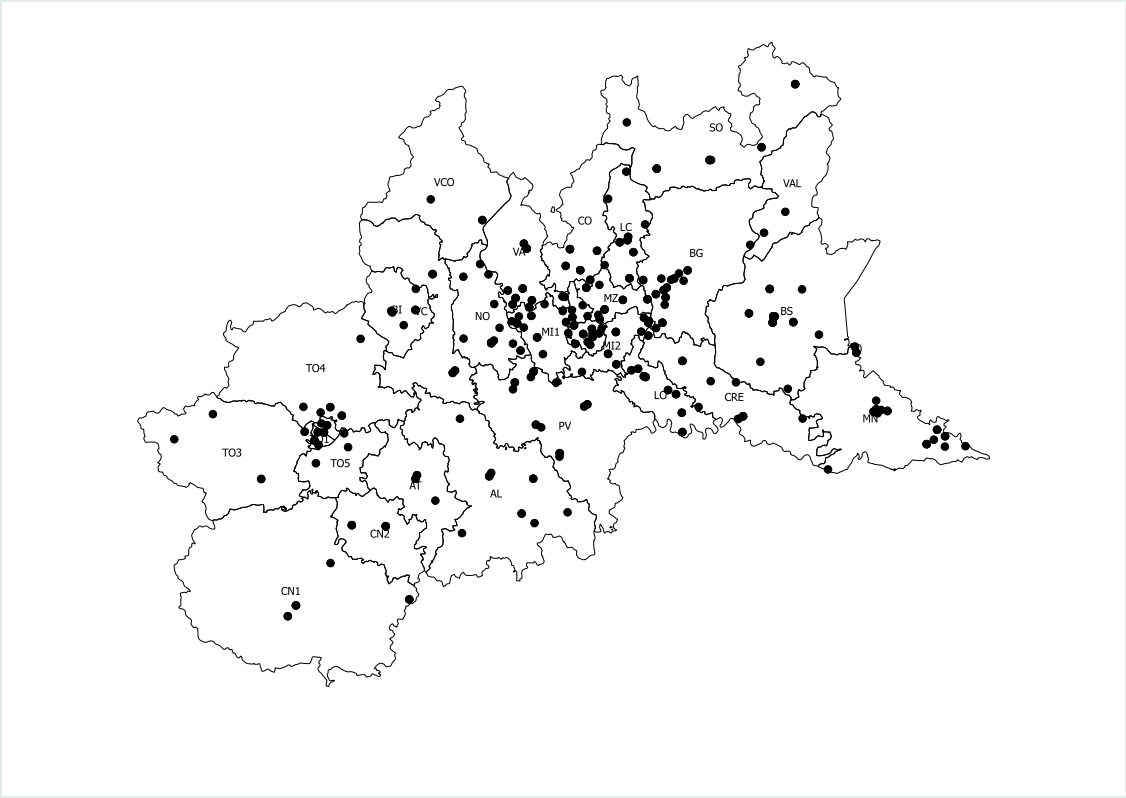
Spatial resolution: point data

O_3 removed from the analyses due to lack of significance

Z **temperature and humidity measurements**

collected on 271 sites and summarized through the first PC

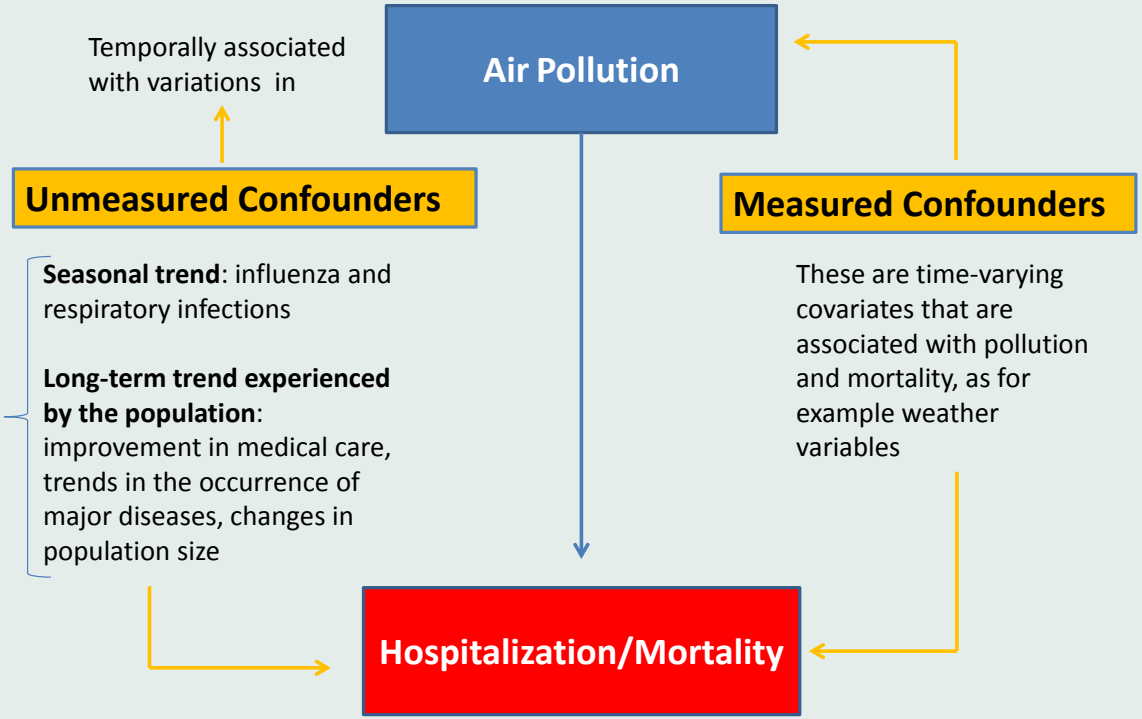
Data feature # 1: map of the region of study



- spatial element is highly relevant
- measurements for Y and X are spatially misaligned!

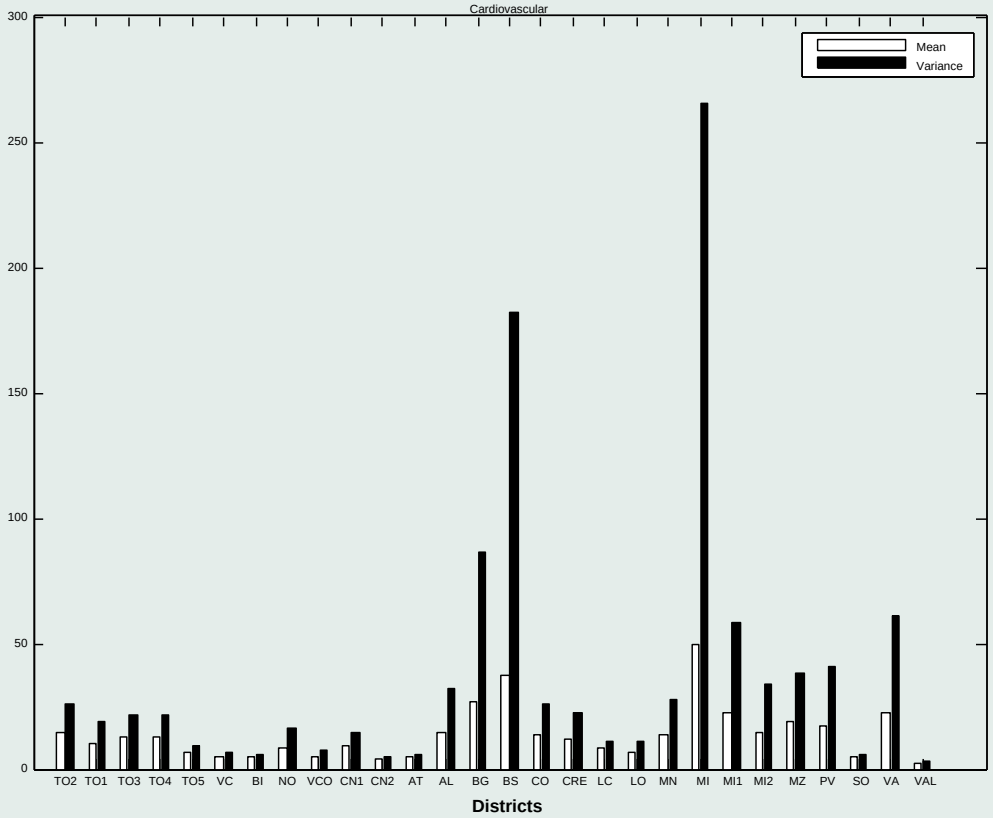
Data feature # 2: time/weather confounding

Measured and unmeasured confounders exist

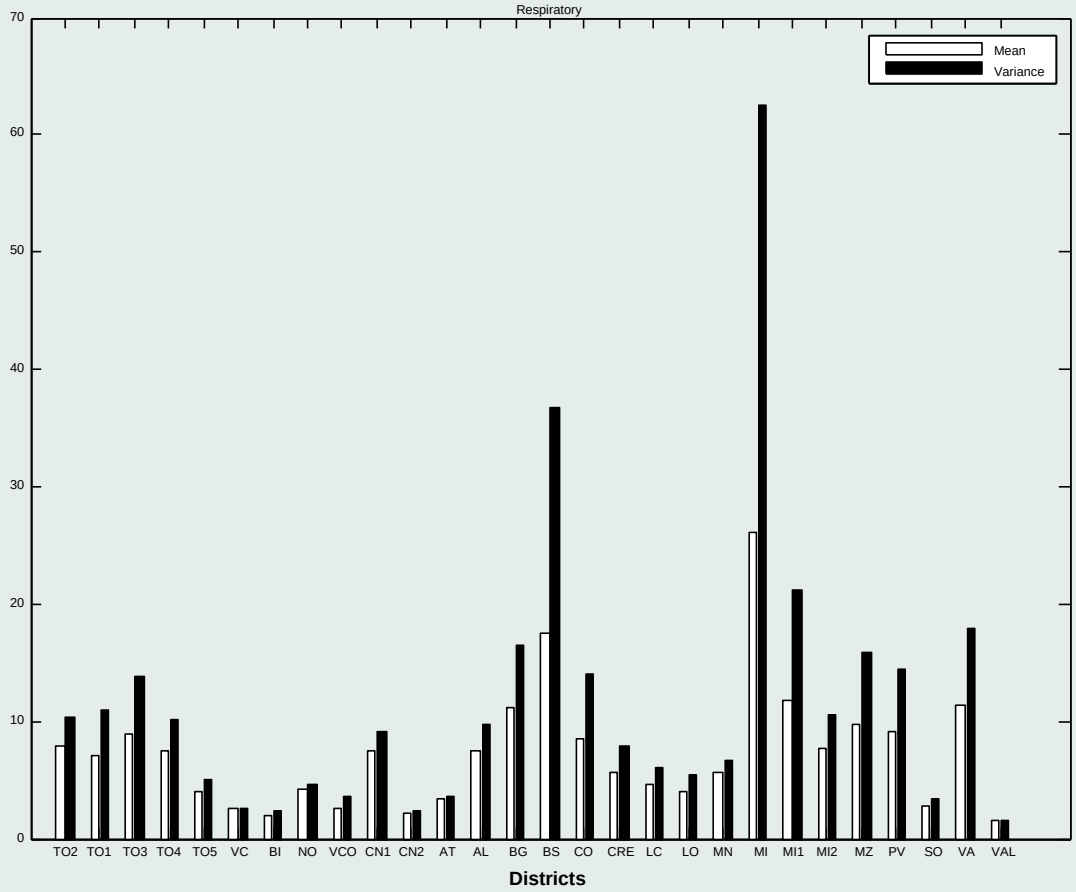


Data feature # 3: data overdispersion

Counts of **cardiovascular diseases** per district



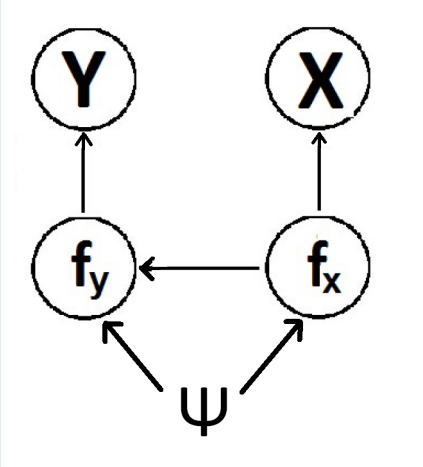
Counts of **respiratory diseases** per district



2. Model

Model is built hierarchically:

- observational equations for X and Y given f_X and f_Y
- latent level specification for f_X and f_Y given Ψ
- hyperparameter specification Ψ



Observational level

$$Y_k(s,t) \mid \eta_{y_k}(s,t), \sigma^2_{y_k} \overset{ind}{\sim} F_y(\eta_{y_k}(s,t), \sigma^2_{y_k})$$

$$X_j(u,t) \mid \eta_{x_j}(u,t), \sigma^2_{x_j} \overset{ind}{\sim} F_x(\eta_{x_j}(u,t), \sigma^2_{x_j}),$$

Link functions complete the specification

$$g_y\big[\eta_{y_k}(\mathbf{s},t)\big] \;=\; \mu_{y_k}(s,t) + \sum_{i=1}^{m} h_{y_k,i}(s)\; f_{y,i}(t)$$

$$g_x\big[\eta_{x_j}(u,t)\big] \;=\; \mu_{x_j}(u,t) + \sum_{i=1}^{r} h_{x_j,i}(u)\; f_{x,i}(t)$$

- $\mu_{y_k}(s,t)$ and $\mu_{x_j}(u,t)$ are mean terms, including (lagged) effects of Z
- $h_{y_k,i}(s)$ and $h_{x_j,i}(u)$ are factor loadings of variables Y_k and X_j
- $f_{y,i}(t)$ and $f_{x,i}(t)$ are corresponding common factors.

Observational level (continued)

Equivalently, in matrix form

$$g_y\Big[\eta_y(t)\Big] = \mu_y(t) + H_y\ f_y(t) \quad (+ \quad \epsilon_y(t) \quad)$$

$$g_x\Big[\eta_x(t)\Big] = \mu_x(t) + H_x\ f_x(t) \quad (+ \quad \epsilon_x(t) \quad)$$

- the mean level terms, $\mu_y(t)$ and $\mu_x(t)$, are fixed effect components,
- $f_y(t) = (f_{y,1}(t), \ldots, f_{y,m}(t))'$ and $f_x(t) = (f_{x,1}(t), \ldots, f_{x,r}(t))'$ are factor vectors for which, potentially, $n_y \cdot N_y \gg m$ and $n_x \cdot N_x \gg r$
- H_y and H_x are matrices of factor loadings of dimensions $(n_y \cdot N_y) \times m$ and $(n_x \cdot N_x) \times r$,

Latent level

Factors must be temporally dependent to address dynamics of the study

VARX: a model for the temporal dynamics of the common factors

$$f_x(t) = \sum_{i=1}^{s} D_i f_x(t-i) + v_x(t)$$

$$f_y(t) = \sum_{i=1}^{p} B_i f_y(t-i) + \sum_{i=0}^{q} C_i f_x(t-i) + v_y(t)$$

- B_i $(m \times m)$, C_i $(m \times r)$ and D_i $(r \times r)$ are AR coefficient matrices modeling the temporal evolution of $f_y(t)$ and $f_x(t)$.
- $v_x(t) \overset{ind}{\sim} N(0, \Sigma_{v_x})$ and $v_y(t) \overset{ind}{\sim} N(0, \Sigma_{v_y})$.
- $\{\ H_x,\ H_x,\ \{B_i\},\ \{C_i\},\ \{D_i\}\ \} \subset\ \Psi$

Latent level (continued)

Factors with order (p, q, s) Markov evolution $\rightarrow$ augmented factors

$$f(t) = \left(f_y(t)' \ \ldots \ f_y(t-p+1)' \ \ f_x(t)' \ \ldots \ f_x(t-q+1)'\right)'$$

Their evolution given by the order 1 Markov equation

$$f(t) = \Gamma \ f(t-1) + \zeta(t), \quad \zeta(t) \sim N(0, \Lambda)$$

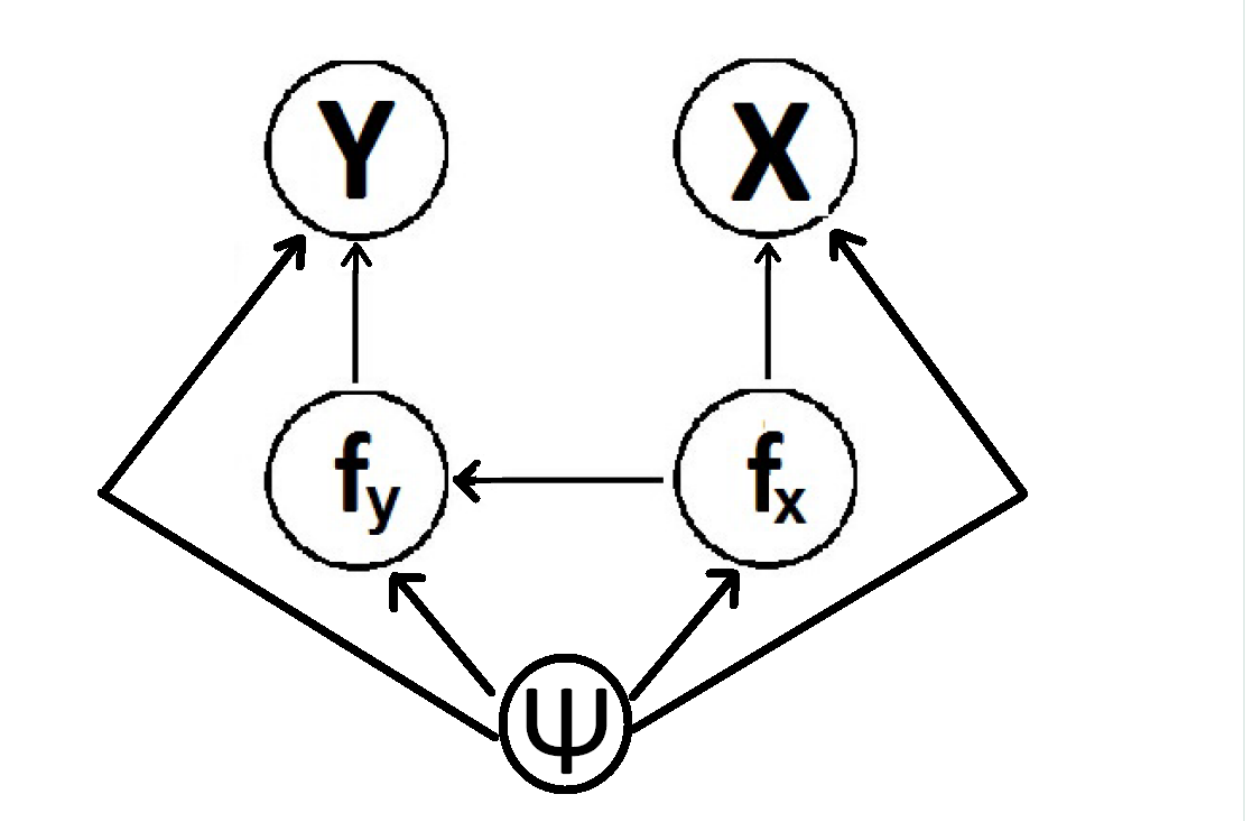
Γ is coefficient matrix with elements γ_{ij} (functions of $\{B_i\}, \{C_i\}, \{D_i\}$)

Λ is covariance matrix with elements λ_{ij} (functions of $\Sigma_{v_x}, \Sigma_{v_y}$).

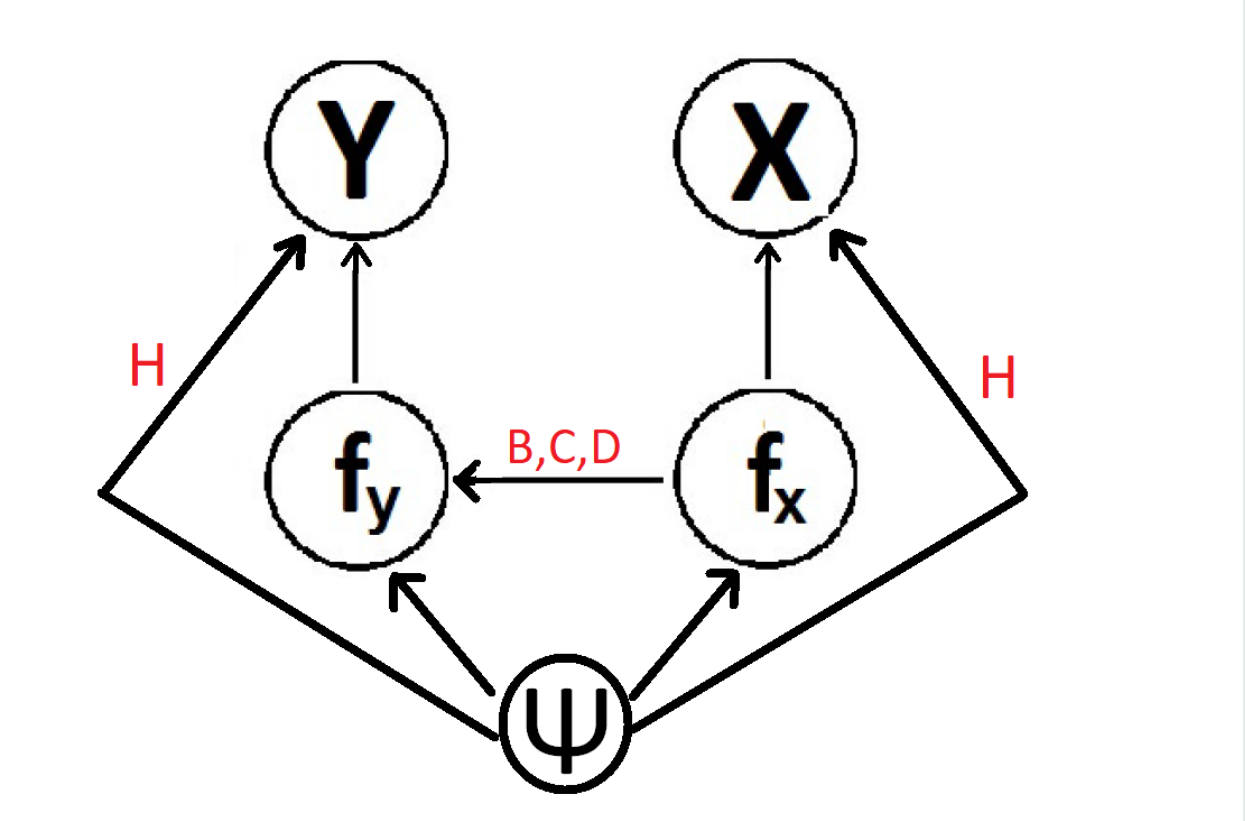
The prior for the latent process $f(t)$ is completed by $f(0) \sim N(a_0, \Sigma_{f0})$.

Ample literature on inference for SSM (order 1 Markov evolution)

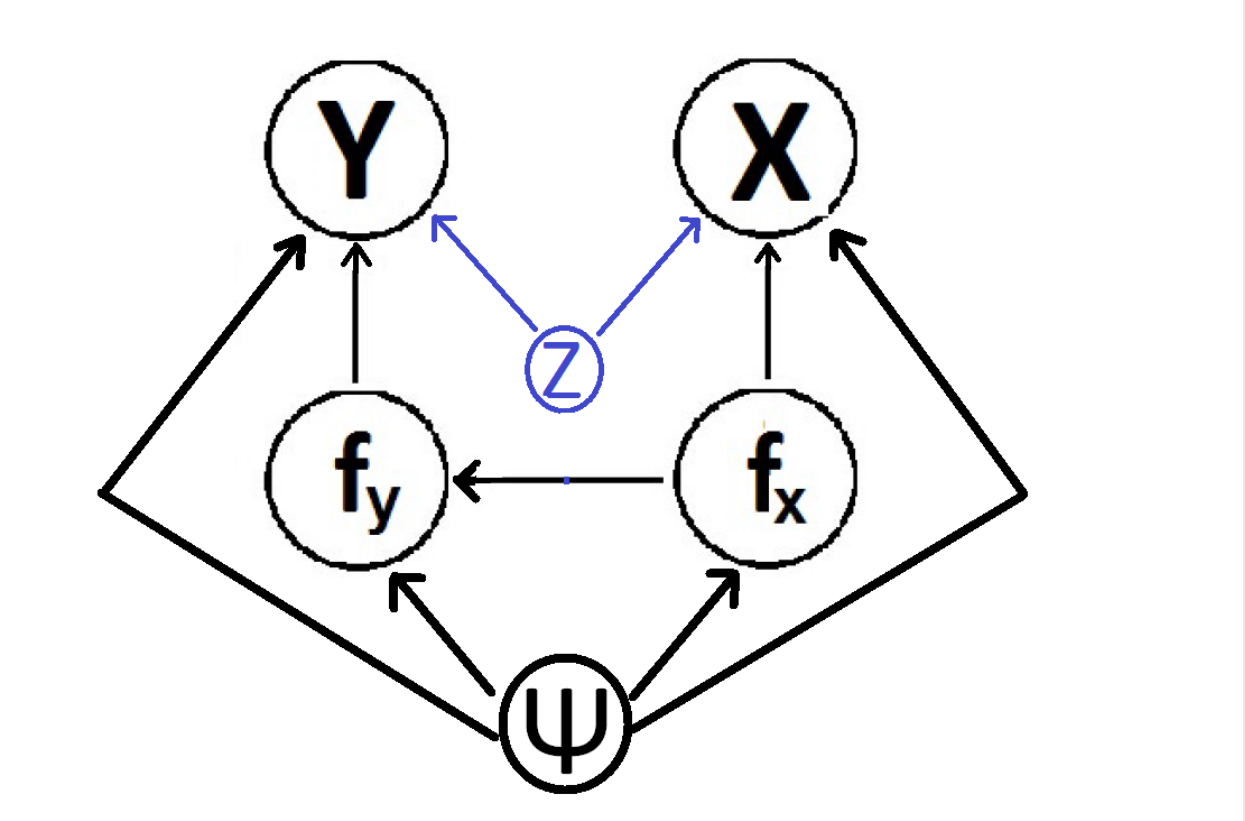
Model summary



Model summary with hyperparameter details



Model summary with confounders



Hyperparameters

Two blocks of hyperparameters deserve special attention:

1) AR coefficients γ_{ij}

2) factor loading matrices H_x and H_y

We will treat them separately

Hyperparameters (continued)

1) AR coefficients γ_{ij}

Factor analysis reduce dramatically dimension

But model still remains richly parametrized

Further reduction based on data would be welcomed

One possibility: spike and slab priors

usual prior is mixed with its version strongly concentrated on 0

Example: $\gamma \sim wN(0, \tau) + (1 - w)N(0, c\tau)$ with c very small

Encourages 0's for the AR coefficients

$\rightarrow$ only the really relevant coefficients survive

Hyperparameters (continued)

2) factor loading matrices H_x and H_y

A number of restrictions can be imposed on loading matrices

a) Deterministic constraints (identifiability)

$$H_y \quad = \quad \left(\begin{array}{c} \blacktriangleright \\ \blacksquare \end{array} \right) \Bigg\} n_y \cdot N_y$$

$$m$$

m factors represent N_y measurements of n_y responses

2) factor loading matrices H_x and H_y (continued)

b) Deterministic constraints (by model choice)

Example: each factor associated with a single pollutant

$$H_x = \begin{pmatrix} \blacksquare & 0 & 0 \\ 0 & \blacksquare & 0 \\ 0 & 0 & \blacksquare \end{pmatrix} \begin{matrix} CO \\ NO_2 \\ PM_{10} \end{matrix}$$

$$r_1 \quad r_2 \quad r_3$$

r_1 factors associated with CO, r_2 with NO_2 and r_3 with PM_{10}

c) Stochastic constraints (on factor loadings H_x and H_y)

Measurements in nearby locations must be summarized similarly

$\rightarrow$ loading vectors $h_{x,i},\ h_{y,j}$ (of $f_{x,i},\ f_{y,j}$) should be spatially structured

Areal data: $h_{y,j} \sim GMRF \Rightarrow E[h_{y,j}(u_k) \mid h_{y,j}(-u_k)] = \sum_{l \sim k} w_{kl} h_{y,j}(u_l)$

$\leftarrow$ depends on neighbourhood

Point data: $h_{x,i} \sim GRF \Rightarrow E[h_{y,j}(s_k) \mid h_{y,j}(-s_k)\}] = \sum_l w_{kl} h_{y,j}(s_l)$

$w_{k,l} = g(\mid s_k - s_l \mid)$ $\leftarrow$ depends on distances

Allow for interpolation $\rightarrow$ useful for factor identification

Factor loadings help "remove"space and drive the spatial aggregation

Side effect: solves spatial misalignment

2.1. Computation

Full model specification

$$\begin{aligned}
\text{Observation equation:} \quad (X(t), Y(t)) &\sim F_w(\eta(t), \sigma(t)) \\
\text{Link function:} \quad g(\eta(t)) &= \mu(t) + Hf(t) \\
\text{Evolution equation:} \quad f(t) &= \Gamma f(t-1) + \zeta(t) \\
\text{Hyperparameter prior:} \quad \Psi &\sim F_\Psi
\end{aligned}$$

Bayesian inference is performed

Intractability of posterior $\rightarrow$ MCMC (Gamerman & Lopes, 2006)

Sampling $f(t)$ jointly (or one at a time) is inefficient

Some options available for efficent MCMC sampling of state vectors:

DG (1998), Knorr-Held & Rue (2002), Lopes, DG & Salazar (2011)

2.2. Outputs from statistical analysis

1. Impulse response functions (IRF)

One of the main interests: effect of an increase in pollutants on health

IRF's perform this task by integrating effects over time

Our factor models mix the pollutants

Making it difficult to obtain IRF for an increase in a single pollutant

Connection of pollutants to health indicators come through factors

Having factors associated with a single pollutant enables it

Outputs from statistical analysis (continued)

2. Prediction

Predictions are readily obtained through predictive distributions

$$p(YT + k \mid Y(1:T), X(1:T+k))$$

or

$$p(YT + k \mid Y(1:T), X(1:T))$$

These distributions are easily sampled from with our SSM formulation

3. Results

Model equations: observational non-normality and overdispersion

$$\text{Response} \begin{cases} Y_k(s,t) \mid \eta_{y_k}(s,t), \sigma^2_{y_k} \overset{ind}{\sim} Poisson(\eta_{y_k}(s,t)) \\ X_j(u,t) \mid \eta_{x_j}(u,t), \sigma^2_{x_j} \overset{ind}{\sim} log-N(\eta_{x_j}(u,t), \sigma^2_{x_j}) \end{cases}$$

$$\text{Link} \begin{cases} log\left\{\frac{\eta_{y_k}(s,t)}{E_k(s,t)}\right\} = \mu_{y_k}(s,t) + \sum_{i=1}^{m} h_{y_k,i}(s)\ f_{y,i}(t) + \epsilon_k(s,t) \\ \eta_{x_j}(u,t) = \mu_{x_j}(u,t) + \sum_{i=1}^{r} h_{x_j,i}(u)\ f_{x,i}(t) \end{cases}$$

$$\text{VARX evolution} \begin{cases} f_x(t) = \sum_{i=1}^{s} D_i f_x(t-i) + v_x(t) \\ f_y(t) = \sum_{i=1}^{p} B_i f_y(t-i) + \sum_{i=0}^{q} C_i f_x(t-i) + v_y(t) \end{cases}$$

Model completed with hyperparameter specifications

VARX coefficients

Model: joint for both diseases, 1 health factor, 4 pollutant factors

Posterior mean estimate and inclusion probabilities

B_i's

B_1	B_2	B_3	B_4
0.08	0.01	0.06	-0.01
(0.52)	(0.30)	(0.45)	(0.26)

C_i's

lag 0	0.01 (0.25)	-0.01 (0.24)	-0.01 (0.29)	0.02 (0.29)	lag 1	0.04 (0.38)	0.01 (0.29)	**0.20 (0.86)**	**0.07 (0.48)**
lag 2	0.02 (0.33)	0.04 (0.39)	**-0.14 (0.70)**	0.00 (0.31)	lag 3	**0.14 (0.74)**	0.00 (0.297)	0.01 (0.32)	**-0.13 (0.71)**
lag 4	0.01 (0.30)	-0.00 (0.22)	0.01 (0.25)	-0.02 (0.31)					

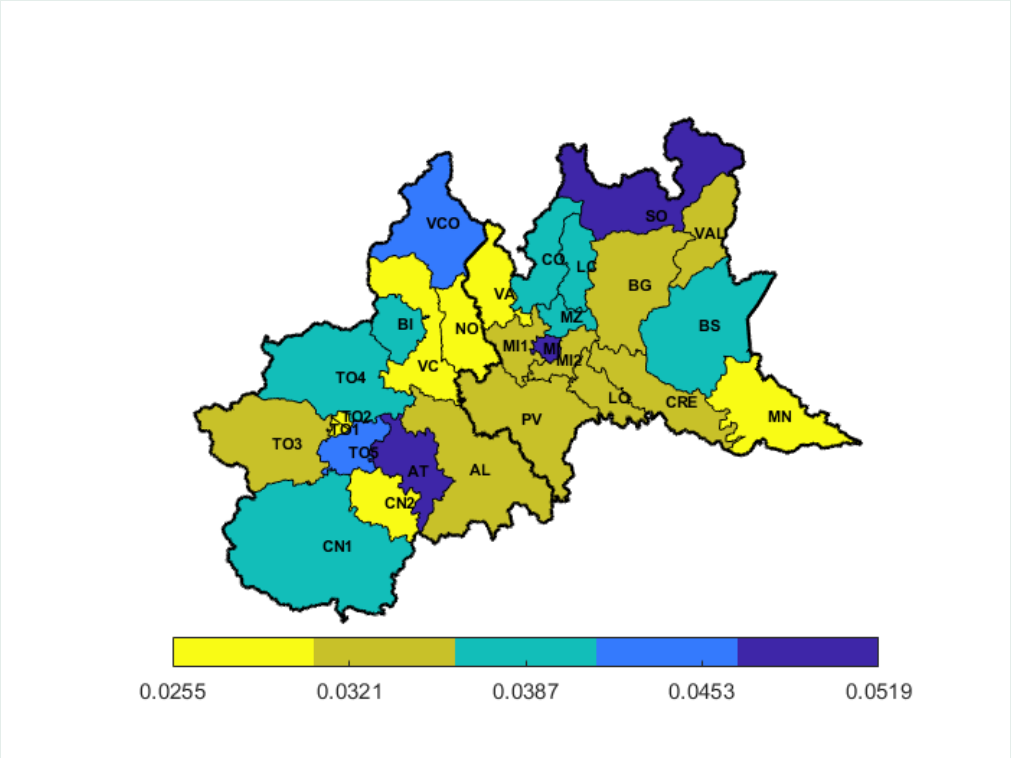
diag (D_i)'s

lag 1	**0.67 (1.00)**	**0.85 (1.00)**	**0.57 (1.00)**	**0.71 (1.00)**	lag 2	-0.03 (0.21)	**-0.22 (0.95)**	-0.01 (0.08)	0.00 (0.03)
lag 3	-0.01 (0.09)	0.00 (0.02)	-0.00 (0.02)	0.00 (0.03)	lag 4	0.00 (0.00)	0.00 (0.02)	0.00 (0.02)	0.02 (0.11)

Factor loadings

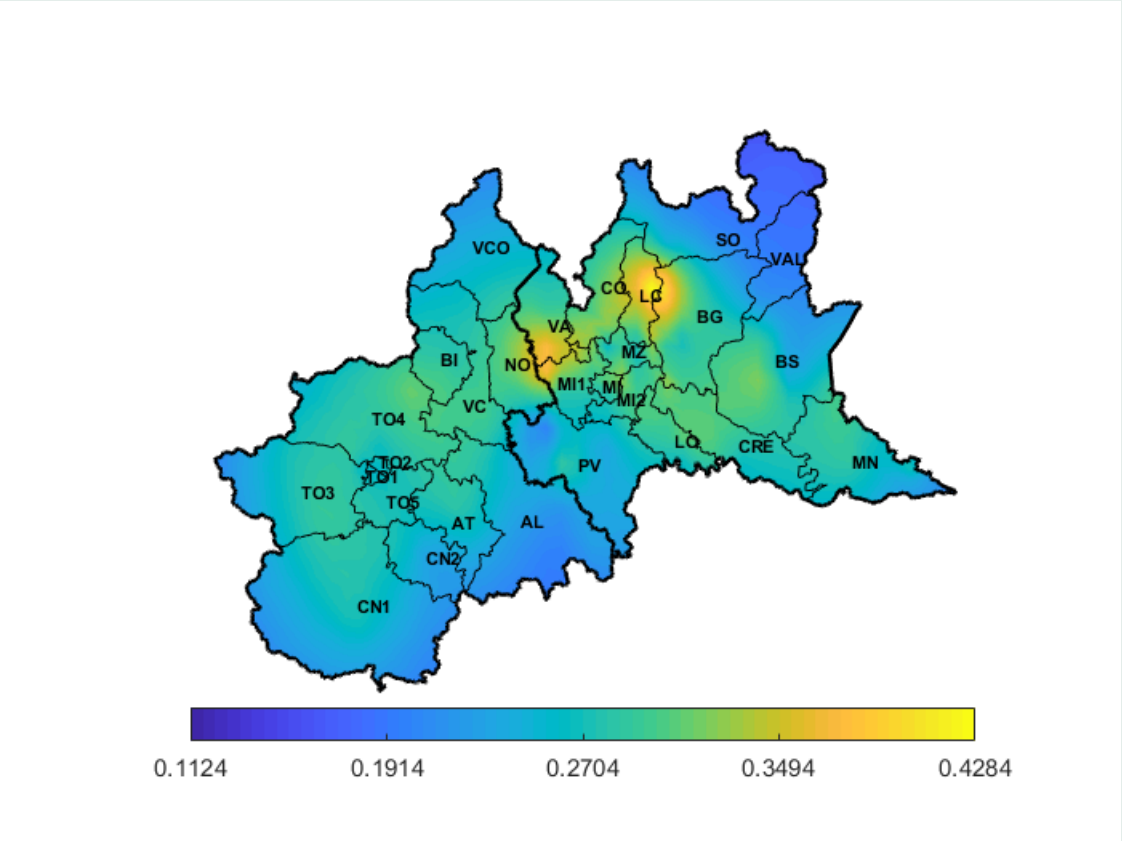
Model: a single factor for respiratory disease, a single factor for PM_{10}

1) Respiratory outcomes



Factor loadings (continued)

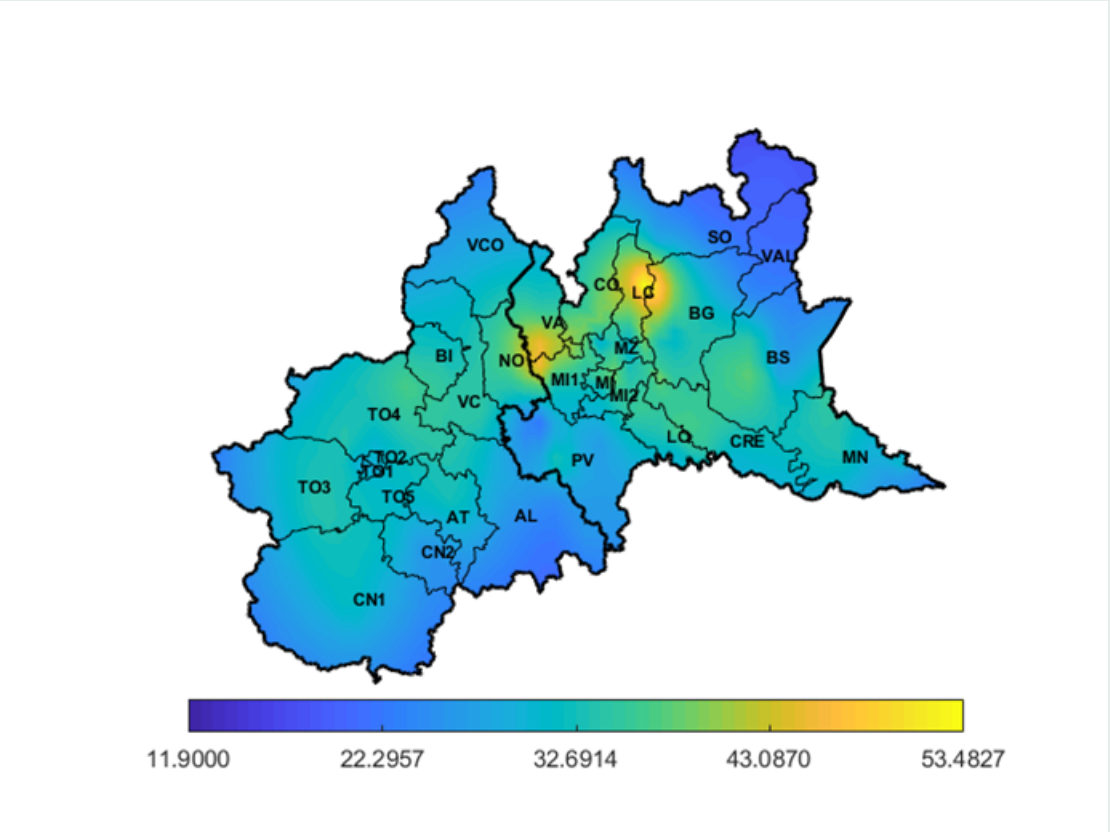
2) PM_{10} (interpolated from monitoring sites)



Metropolitan region of Milan drives this factor (traffic pollution?)

Factor loadings (continued)

2) PM_{10} (reproduced at outcome level)



mean % increase at outcome from increase of 1 unit at factor level

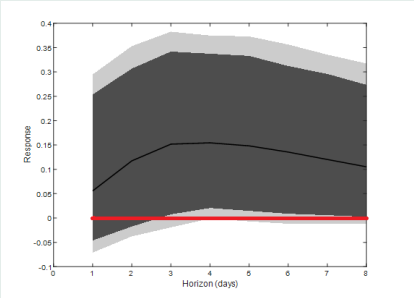
Impulse response functions (latent level)

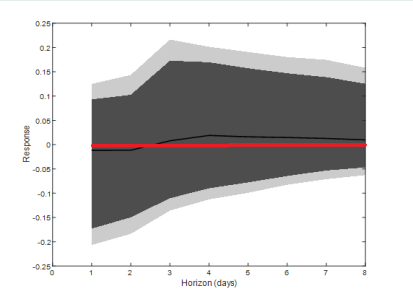
1) Respiratory disease

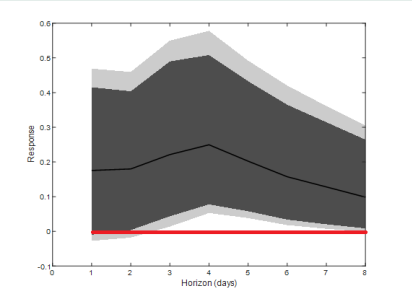
3 simple models were considered:

a single factor for the disease and a single factor for one pollutant

CO NO_2 PM_{10}







NO_2 seems not to have a relevant effect on health

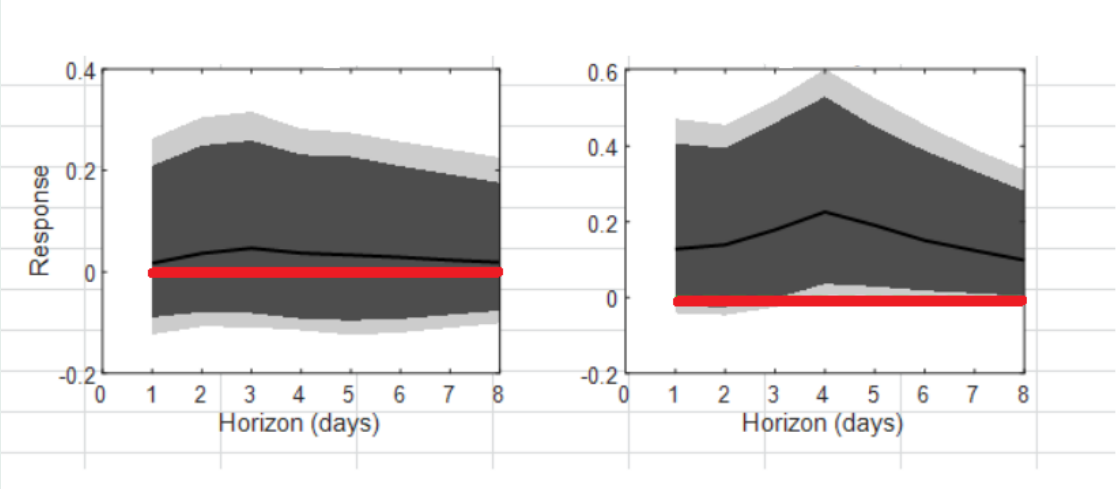
Impulse response functions (latent level)

1) Respiratory disease (continued)

1 model considered:

a single factor for the disease, 1 factor for CO and 1 factor for PM_{10}

CO PM_{10}



only PM_{10} seems to have a relevant effect on health

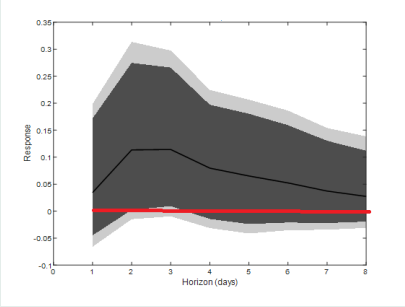
Impulse response functions (latent level)

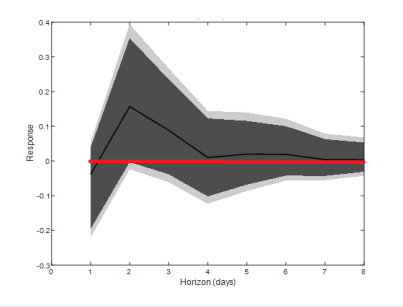
2) Cardiovascular disease

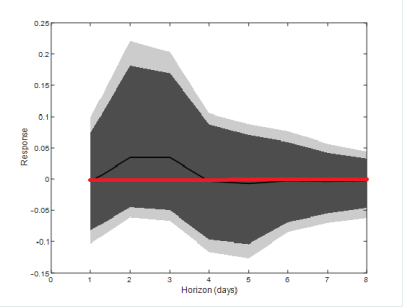
3 simple models were considered:

a single factor for the disease and a single factor for one pollutant

CO NO_2 PM_{10}







PM_{10} seems not to have a relevant effect on health

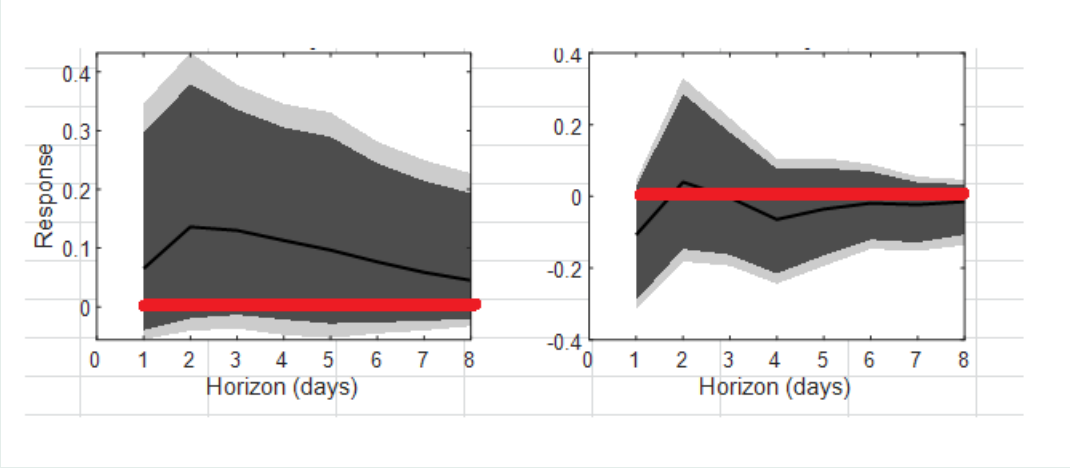
Impulse response functions (latent level)

2) Cardiovascular disease (continued)

1 model considered:

a single factor for the disease, 1 factor for CO and 1 factor for NO_2

CO NO_2



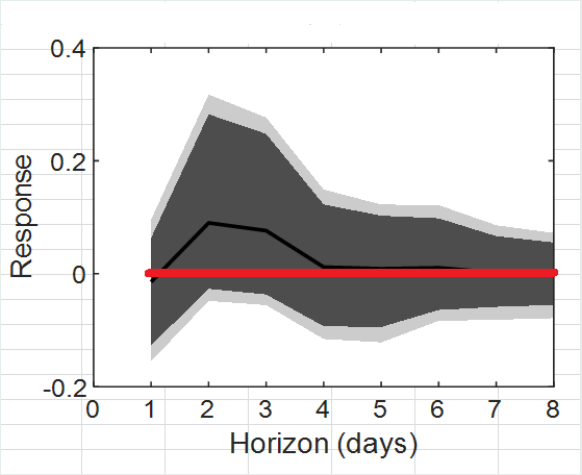
only CO seems to some effect on health

Impulse response functions (latent level)

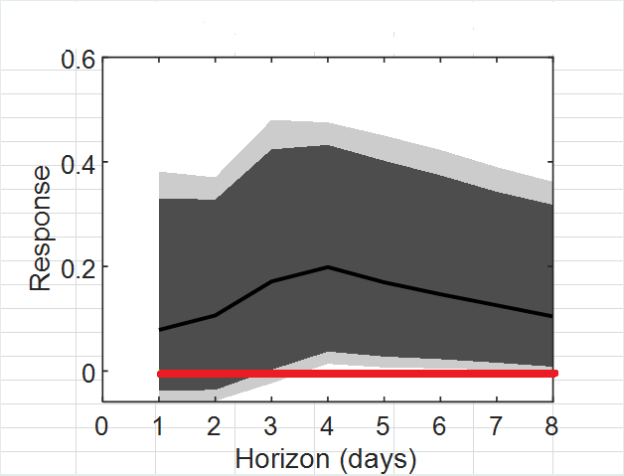
3) Both diseases jointly

Model: 1 factor for each disease, 1 factor for all pollutants

Cardiovascular



Respiratory



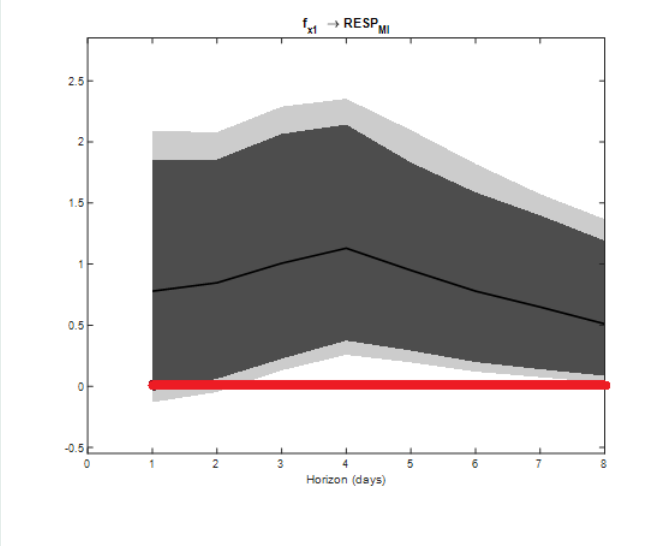
Pollutant factor much more relevant for respiratory disease

2nd factor for pollutants not relevant for both diseases

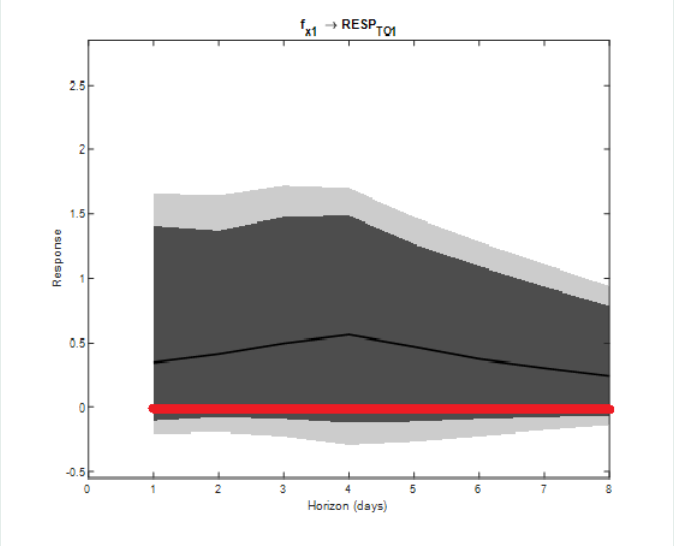
Impulse response functions (outcome level)

Model: a single factor for respiratory disease, a single factor for PM_{10}

Milan



Torino

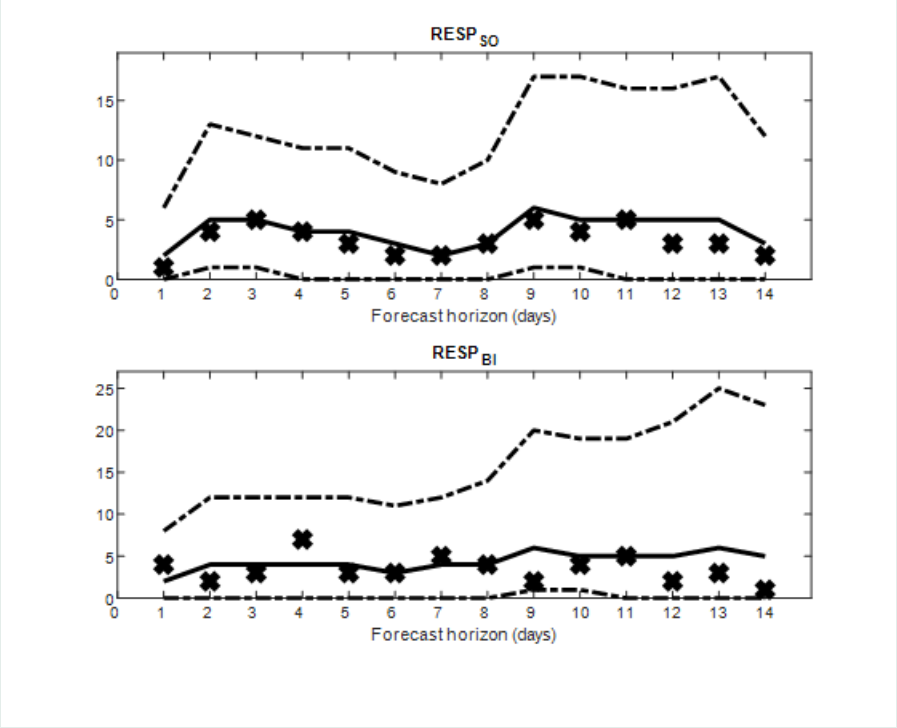


Pollution factor is relevant in Milan but not in Torino

Predictions

Model: a single factor for respiratory disease, a single factor for PM_{10}

Out-of-sample prediction for 2 weeks ahead



4. Conclusion

Generalized Dynamic SEM was proposed

- Latent approach avoids the curse of dimensionality
- Specifies the problem in terms of temporal effects
- Separates spatial from temporal effects
- Facilitates identification of spatial clusters
- Solves spatial misalignment
- Not limited to Gaussianity
- Easy to obtain predictions and IRFs

Thank you

dani@im.ufrj.br

www.statpop.com.br