Statistics for High Dimensional Data (and CompStat Lab) a.a. 2022/2023 (1st edition)

Prof. Francesco Finazzi francesco.finazzi@unibg.it
Prof. Alessandro Fassò alessandro.fasso@unibg.it

Lesson 3

Spatial model with latent variable

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \alpha w(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \qquad (2)$$

- $w(\boldsymbol{s}) \sim GP\left(0, \rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \boldsymbol{\theta})\right)$
- $\rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \boldsymbol{\theta}) = corr\left(w(\boldsymbol{s}), w(\boldsymbol{s}')\right)$
- $\varepsilon(\boldsymbol{s}) \sim N(0, \sigma_\varepsilon^2)$
- The unknown parameter set is $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2, \boldsymbol{\theta}\}$
- How to estimate Ψ from data?

Maximum likelihook estimate

- We rely on MLE to estimate the model parameter vector
- Because MLE has good properties and…
- Because we can use the EM algorithm (if needed)
- Likelihood function of (2) is:

$$L(\Psi; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X}) = L(\Psi; \boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X})L(\Psi; \boldsymbol{w})$$

- $\boldsymbol{y} = (y_1, \dots, y_n)'$ is the vector of observations at n spatial locations
- $\boldsymbol{w} = (w_1, \dots, w_n)'$ is the vector of latent variables at n spatial locations
- $\boldsymbol{X}$ is the $n \times p$ design matrix

Likelihood decomposition

$$
\begin{aligned}
L(\Psi; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X}) &= L(\Psi; \boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X}) L(\Psi; \boldsymbol{w}) \\
&= L(\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2; \boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X}) L(\boldsymbol{\theta}; \boldsymbol{w})
\end{aligned}
$$

- Each likelihood term depends on a subset of Ψ.
- $L(\Psi; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X})$ is the complete-data likelihood (which assumes $\boldsymbol{w}$ to be known)
- $L(\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2; \boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X})$ and $L(\boldsymbol{\theta}; \boldsymbol{w})$ are densities of n –variate normal distributions

Log-likelihood function

- As usual we prefer to work with $\log(L_\Psi)$
- $-2\log L_\Psi$ is given by:

$$log|\Sigma_\varepsilon| + \boldsymbol{e}'\Sigma_\varepsilon^{-1}\boldsymbol{e} + log\ |\Sigma_w| + \boldsymbol{w}'\Sigma_w^{-1}\boldsymbol{w}$$

where

- $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\beta} - \alpha\boldsymbol{w}$
- $\Sigma_\varepsilon = \sigma_\varepsilon^2\boldsymbol{I}_n$, with $\boldsymbol{I}_n$ the identity matrix of dimension n
- Σ_w is the $n \times n$ correlation matrix (e.g., $exp(-D/\theta)$, with D the distance matrix)

ML estimate

- MLE is given by

$$\hat{\Psi} = argmin_{\boldsymbol{\beta},\alpha,\sigma_\varepsilon^2,\boldsymbol{\theta}} \; log|\Sigma_\varepsilon| + \boldsymbol{e}'\Sigma_\varepsilon^{-1}\boldsymbol{e} + log\;|\Sigma_w| + \boldsymbol{w}'\Sigma_w^{-1}\boldsymbol{w}$$

- Argmin because we are considering $-2\log(L_\Psi)$
- Unfortunately minimizing $-2\log(L_\Psi)$ is not feasible, plus $\boldsymbol{w}$ is latent and not observed
- We must rely on the EM algorithm

EM algorithm

- The EM is an iterative algorithm for MLE
- First iteration starts with initial values $\hat{\Psi}^{\langle 0 \rangle}$ (usually given by OLS and method of moments)

- E-step

$$Q\left(\Psi, \hat{\Psi}^{\langle m \rangle}\right) = E_{\hat{\Psi}^{\langle m \rangle}}(-2\log \mathrm{L}(\Psi; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{X})|\boldsymbol{y})$$

- M-step

$$\hat{\Psi}^{\langle m+1 \rangle} = argmax_{\Psi}\, Q\left(\Psi, \hat{\Psi}^{\langle m \rangle}\right)$$

EM algorithm, E-step

- E-step

$$E_{\hat{\Psi}\langle m\rangle}(-2\log\mathrm{L}(\Psi;\boldsymbol{y},\boldsymbol{w},\boldsymbol{X})|\boldsymbol{y})$$
$$= tr\left[\Sigma_{\varepsilon}^{-1}\left(E(\boldsymbol{e}|\boldsymbol{y})E(\boldsymbol{e}|\boldsymbol{y})' + Var(\boldsymbol{e}|\boldsymbol{y})\right)\right]$$
$$+ tr\left[\Sigma_{w}^{-1}\left(E(\boldsymbol{w}|\boldsymbol{y})E(\boldsymbol{w}|\boldsymbol{y})' + Var(\boldsymbol{w}|\boldsymbol{y})\right)\right]$$

- $E(\boldsymbol{e}|\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{X\beta} - \alpha E(\boldsymbol{w}|\boldsymbol{y})$
- $Var(\boldsymbol{e}|\boldsymbol{y}) = Var(\boldsymbol{y} - \boldsymbol{X\beta} - \alpha\boldsymbol{w}|\boldsymbol{y}) = \alpha^2 Var(\boldsymbol{w}|\boldsymbol{y})$

EM algorithm, E-step

- $E(\boldsymbol{w}|\boldsymbol{y}) = Cov(\boldsymbol{w},\boldsymbol{y})Var(\boldsymbol{y})^{-1}[\boldsymbol{y} - \boldsymbol{X\beta}]$ (see multivariate normal)
- $Var(\boldsymbol{w}|\boldsymbol{y}) = \Sigma_w - Cov(\boldsymbol{w},\boldsymbol{y})Var(\boldsymbol{y})^{-1}Cov(\boldsymbol{w},\boldsymbol{y})'$
- $Var(\boldsymbol{y}) = Var(\boldsymbol{X\beta} + \alpha\boldsymbol{w} + \boldsymbol{\varepsilon}) = Var(\alpha\boldsymbol{w} + \boldsymbol{\varepsilon}) = \alpha^2 Var(\boldsymbol{w}) + Var(\boldsymbol{\varepsilon}) + 2\mathrm{Cov}(\boldsymbol{w},\boldsymbol{\varepsilon})$
- $Var(\boldsymbol{w}) = \Sigma_w$
- $Var(\boldsymbol{\varepsilon}) = \Sigma_\varepsilon = \sigma_\varepsilon^2 \boldsymbol{I}_n$
- $2\mathrm{Cov}(\boldsymbol{w},\boldsymbol{\varepsilon}) = \boldsymbol{0}$ (from model assumptions)
- $Cov(\boldsymbol{w},\boldsymbol{y}) = Cov(\boldsymbol{w},\boldsymbol{X\beta} + \alpha\boldsymbol{w} + \boldsymbol{\varepsilon}) = Cov(\boldsymbol{w},\alpha\boldsymbol{w}) = \alpha Cov(\boldsymbol{w},\boldsymbol{w}) = \alpha Var(\boldsymbol{w}) = \alpha\Sigma_w$

EM algorithm, M-step

$$\widehat{\Psi}^{\langle m+1\rangle} = argmax_{\Psi}\, Q\left(\Psi, \widehat{\Psi}^{\langle m\rangle}\right)$$

$$\frac{dQ\left(\Psi, \widehat{\Psi}^{\langle m\rangle}\right)}{d\Psi} = 0$$

$$\alpha^{\langle m+1\rangle} = \frac{tr\left[\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{\langle m\rangle}\right)E(\boldsymbol{w}|\boldsymbol{y})'\right]}{tr[E(\boldsymbol{w}|\boldsymbol{y})E(\boldsymbol{w}|\boldsymbol{y})' + Var(\boldsymbol{w}|\boldsymbol{y})]}$$

$$\boldsymbol{\beta}^{\langle m+1\rangle} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left[\boldsymbol{X}'\left(\boldsymbol{y} - \alpha^{\langle m+1\rangle}E(\boldsymbol{w}|\boldsymbol{y})\right)\right]$$

EM algorithm, M-step

$$\sigma_{\varepsilon}^{2\langle m+1\rangle} = \frac{1}{n} tr[E(\boldsymbol{e}|\boldsymbol{y})E(\boldsymbol{e}|\boldsymbol{y})' + Var(\boldsymbol{e}|\boldsymbol{y})]$$

$$\boldsymbol{\theta}^{\langle m+1\rangle} = argmin_{\boldsymbol{\theta}} \; log\,|\Sigma_w^{-1}(\boldsymbol{\theta})| + tr[\Sigma_w^{-1}(\boldsymbol{\theta})(\widehat{\boldsymbol{w}}\widehat{\boldsymbol{w}}')]$$

Where $\widehat{\boldsymbol{w}} = E_{\boldsymbol{\theta}^{\langle m\rangle}}(\boldsymbol{w}|\boldsymbol{y}) = E(\boldsymbol{w}|\boldsymbol{y})$