Normalized Latent Measure Factor Models

Mario Beraha, Jim E. Griffin
Politecnico di Milano, University College London

Università Cattolica del Sacro Cuore, Milano

15 July 2022

Setup

Observation **naturally grouped** into g subpopulations

$$y_{11}, \ldots, y_{1n_1}, \ldots\ldots, y_{g1} \ldots, y_{gn_g}$$

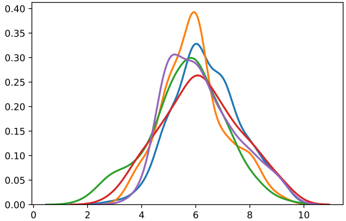
Setup

Observation **naturally grouped** into g subpopulations

$$y_{11}, \dots, y_{1n_1}, \dots\dots, y_{g1} \dots, y_{gn_g}$$

Invalsi Dataset

- Grades of a math test in Italian high schools, 40k students in $g > 1$k schools
- $4 \leq n_j \leq 140$



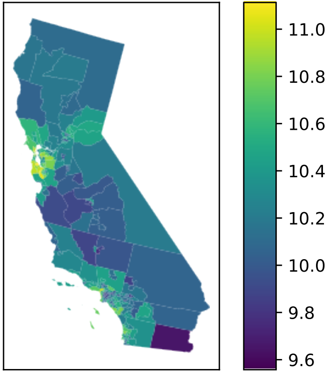
Setup

Observation **naturally grouped** into g subpopulations

$$y_{11}, \ldots, y_{1n_1}, \ldots\ldots, y_{g1} \ldots, y_{gn_g}$$

Income Data in California

- Fine spatial aggregation (PUMA level, 255 pumas)
- Take into account the spatial dependence



Focus and goals

- Focus # 1: **density modelling** in each group

$$y_{ji} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta), \qquad \widetilde{p}_1, \ldots, \widetilde{p}_g \sim Q$$

Focus and goals

- Focus # 1: **density modelling** in each group

$$y_{ji} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta), \qquad \widetilde{p}_1, \ldots, \widetilde{p}_g \sim Q$$

- Few observations per group: "*large g, small n_j setting*".
  $\Longrightarrow$ cannot inform **overparametrized** models

$$Q(\mathrm{d}\widetilde{p}_1, \ldots, \mathrm{d}\widetilde{p}_g) \neq \prod_{j=1}^{g} \mathrm{DP}(\mathrm{d}\widetilde{p}_j \mid \alpha G_0), \qquad Q(\mathrm{d}\widetilde{p}_1, \ldots, \mathrm{d}\widetilde{p}_g) \overset{?}{\sim} \mathrm{HDP}$$

Focus and goals

- Focus # 1: **density modelling** in each group

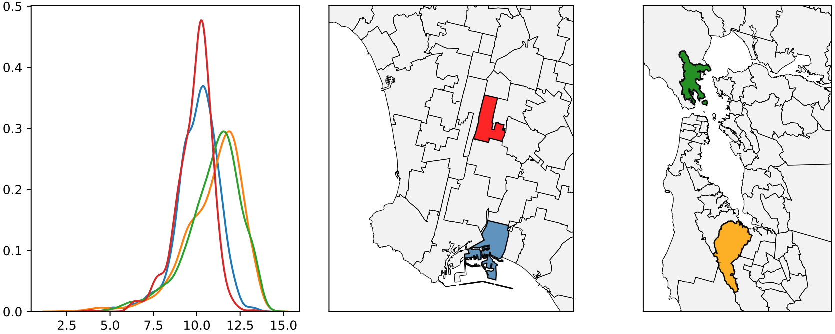
$$y_{ji} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta), \qquad \widetilde{p}_1, \ldots, \widetilde{p}_g \sim Q$$

- Few observations per group: "*large g, small n_j setting*".
$\Longrightarrow$ cannot inform **overparametrized** models

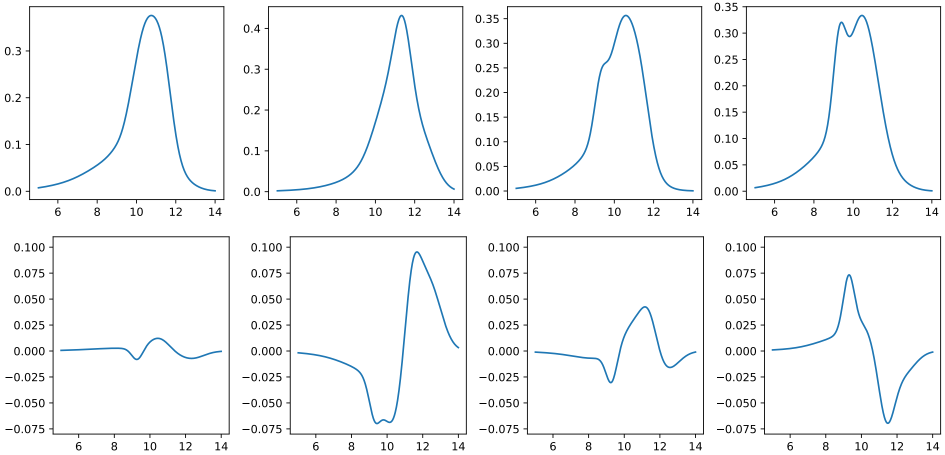
$$Q(\mathrm{d}\widetilde{p}_1, \ldots, \mathrm{d}\widetilde{p}_g) \neq \prod_{j=1}^{g} \mathrm{DP}(\mathrm{d}\widetilde{p}_j \mid \alpha G_0), \qquad Q(\mathrm{d}\widetilde{p}_1, \ldots, \mathrm{d}\widetilde{p}_g) \neq \mathrm{HDP}$$

- Focus # 2: **explore** and **explain** the difference across groups

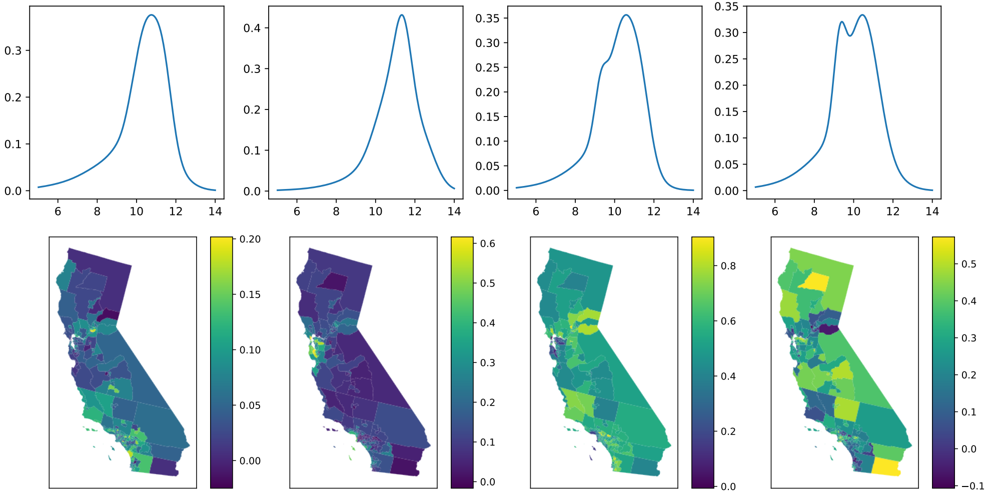
Practical Example: California Income Data



Practical Example: California Income Data



Practical Example: California Income Data



Latent Factor Models

$$large\ g,\ small\ n_j \quad \approx \quad large\ p,\ small\ n$$

Latent Factor Models

$$\textit{large } g, \textit{ small } n_j \quad \approx \quad \textit{large } p, \textit{ small } n$$

$x_1, \ldots, x_n \in \mathbb{R}^p$, $p \gg n$.

$$x_i \mid \Lambda, \eta_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda\eta_i, \Sigma)$$

Latent Factor Models

$$\textit{large } g, \textit{ small } n_j \quad \approx \quad \textit{large } p, \textit{ small } n$$

$x_1, \ldots, x_n \in \mathbb{R}^p$, $p \gg n$.

$$x_i \mid \Lambda, \eta_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda \eta_i, \Sigma)$$
$$\eta_1, \ldots, \eta_n \overset{\text{iid}}{\sim} \mathcal{N}_H(0, I), \quad H \ll p$$
$$\Lambda \sim \pi(\Lambda) \qquad\qquad \Sigma \sim \pi(\Sigma)$$

Latent Factor Models

$$\textit{large } g, \textit{ small } n_j \quad \approx \quad \textit{large } p, \textit{ small } n$$

$x_1, \ldots, x_n \in \mathbb{R}^p$, $p \gg n$.

$$x_i \mid \Lambda, \eta_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda \eta_i, \Sigma)$$

- **Parsimonious** model
- **Interpretability** of latent factors (*main variability*) and loadings

Latent Factor Models

$$\textit{large } g, \textit{ small } n_j \quad \approx \quad \textit{large } p, \textit{ small } n$$

$x_1, \ldots, x_n \in \mathbb{R}^p$, $p \gg n$.

$$x_i \mid \Lambda, \eta_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda\eta_i, \Sigma)$$

- **Parsimonious** model
- **Interpretability** of latent factors (*main variability*) and loadings

$$x_{ij} = \sum_{h=1}^{H} \lambda_{jh}\eta_{ih} + \varepsilon_{ij}, \; j = 1, \ldots, p$$

Latent Factor Models

$$\textit{large } g, \textit{ small } n_j \quad \approx \quad \textit{large } p, \textit{ small } n$$

$x_1, \ldots, x_n \in \mathbb{R}^p$, $p \gg n$.

$$x_i \mid \Lambda, \eta_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda\eta_i, \Sigma)$$

- **Parsimonious** model
- **Interpretability** of latent factors (*main variability*) and loadings

$$x_{ij} = \sum_{h=1}^{H} \lambda_{jh}\eta_{ih} + \varepsilon_{ij}, \; j = 1, \ldots, p \quad \Longrightarrow \quad \widetilde{p}_j = \sum_{h=1}^{H} \lambda_{jh} p_h^*, \; j = 1, \ldots, g$$

A Normalized Random Measure Approach

$$y_{ji} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta), \quad i = 1, \ldots, n_j$$

Avoid overly constrained parameters by setting

$$\widetilde{p}_j = \frac{\widetilde{\mu}_j}{\widetilde{\mu}_j(\Theta)}, \qquad \widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh}\mu_h^*$$

A Normalized Random Measure Approach

$$y_{ji} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta), \quad i = 1, \ldots, n_j$$

Avoid overly constrained parameters by setting

$$\widetilde{p}_j = \frac{\widetilde{\mu}_j}{\widetilde{\mu}_j(\Theta)}, \qquad \widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh}\mu_h^*$$

- λ_{jh}'s must be positive
- $\mu_1^*, \ldots, \mu_H^*$ a collection of completely random measures

A Normalized Random Measure Approach

$$\widetilde{p}_j = \frac{\widetilde{\mu}_j}{\widetilde{\mu}_j(\Theta)}, \qquad \widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh} \mu_h^*$$

Connections with normalized **additive processes**

- Lijoi et al. (2014): $g = 2$, $H = 3$, $\lambda_1 = (1, 1, 0)$, $\lambda_2 = (1, 0, 1)$
  $\Longrightarrow$ focus on flexible sharing of information
- Griffin et al. (2013): $g > 2$, $H = 2^g$ (typically), $\lambda_{jh} \sim \text{Bern}(\rho)$
  $\Longrightarrow$ focus on detecting the presence of differences
- No way of extracting "characteristic traits" across populations

Prior Modelling – the μ_h^*'s

Most natural choice

$$\mu_h^* = \sum_{k \geq 1} J_{hk} \delta_{\theta_{hk}} \overset{\text{iid}}{\sim} \mathrm{CRM}(\nu_h; \Theta)$$

Prior Modelling – the μ_h^*'s

Most natural choice

$$\mu_h^* = \sum_{k \geq 1} J_{hk} \delta_{\theta_{hk}} \overset{\text{iid}}{\sim} \text{CRM}(\nu_h; \Theta)$$

Still too many parameters! No need for measure-specific atoms

Prior Modelling – the μ_h^*'s

Most natural choice

$$\mu_h^* = \sum_{k \geq 1} J_{hk} \delta_{\theta_{hk}} \stackrel{\text{iid}}{\sim} \text{CRM}(\nu_h; \Theta)$$

Still too many parameters! No need for measure-specific atoms

Compound Random Measures (CoRMs; Griffin and Leisen, 2017)

$$\mu_h^* = \sum_{k \geq 1} m_{hk} J_k \delta_{\theta_k^*}$$

Prior Modelling – the μ_h^*'s

Most natural choice

$$\mu_h^* = \sum_{k \geq 1} J_{hk} \delta_{\theta_{hk}} \stackrel{\text{iid}}{\sim} \mathrm{CRM}(\nu_h; \Theta)$$

Still too many parameters! No need for measure-specific atoms

Compound Random Measures (CoRMs; Griffin and Leisen, 2017)

$$\mu_h^* = \sum_{k \geq 1} m_{hk} J_k \delta_{\theta_k^*}$$

Default choice:

$$m_{hk} \stackrel{\text{iid}}{\sim} \mathrm{Gamma}(\phi, 1)$$

Prior Modelling – the μ_h^*'s

Most natural choice

$$\mu_h^* = \sum_{k \geq 1} J_{hk} \delta_{\theta_{hk}} \stackrel{\text{iid}}{\sim} \text{CRM}(\nu_h; \Theta)$$

Still too many parameters! No need for measure-specific atoms

Compound Random Measures (CoRMs; Griffin and Leisen, 2017)

$$\mu_h^* = \sum_{k \geq 1} m_{hk} J_k \delta_{\theta_k^*}$$

Default choice:

$$m_{hk} \stackrel{\text{iid}}{\sim} \text{Gamma}(\phi, 1)$$

Marginally, each μ_h^* is a Gamma process with base measure $\alpha(\mathrm{d}\theta)$.

A fresh take on the model

When $\mu_1^*, \ldots, \mu_H^* \sim \mathrm{CoRM}$, we have

$$\widetilde{\mu}_j = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k}$$

A fresh take on the model

When $\mu_1^*, \ldots, \mu_H^* \sim \text{CoRM}$, we have

$$\widetilde{\mu}_j = \sum_{k \geq 1} \Gamma_{jk} J_k \delta_{\theta_k}$$

- Essentially, forcing a parsimonious factorization of the matrix $\Gamma \approx \Lambda M$
- Connections to Nonnegative Matrix Factorization and Independent Component Analysis!

Prior Modelling – the matrix Λ, Invalsi dataset

- No additional group-specific information

Prior Modelling – the matrix Λ, Invalsi dataset

- No additional group-specific information
- Very large number of groups $\Longrightarrow$ impractical to choose H via model selection

Prior Modelling – the matrix Λ, Invalsi dataset

- No additional group-specific information
- Very large number of groups $\Longrightarrow$ impractical to choose H via model selection
- Multiplicative Gamma Process (Bhattacharya and Dunson, 2011) for Λ:

$$\lambda_{jh} = (\phi_{jh}\tau_h)^{-1}, \qquad \tau_h = \prod_{j=1}^{h} \theta_j,$$

$$\theta_1 \sim \mathrm{Ga}(a_1, 1), \quad \theta_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Ga}(a_2, 1), \quad \phi_{jh} \overset{\text{iid}}{\sim} \mathrm{Ga}(\nu/2, \nu/2)$$

Prior Modelling – the matrix Λ, Invalsi dataset

- No additional group-specific information
- Very large number of groups $\Longrightarrow$ impractical to choose H via model selection
- Multiplicative Gamma Process (Bhattacharya and Dunson, 2011) for Λ:

$$\lambda_{jh} = (\phi_{jh}\tau_h)^{-1}, \qquad \tau_h = \prod_{j=1}^{h} \theta_j,$$

$$\theta_1 \sim \mathrm{Ga}(a_1, 1), \quad \theta_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Ga}(a_2, 1), \quad \phi_{jh} \overset{\text{iid}}{\sim} \mathrm{Ga}(\nu/2, \nu/2)$$

- Learn H through adaptive Gibbs sampling in the first MCMC iterations

Prior Modelling – the matrix Λ, California Income

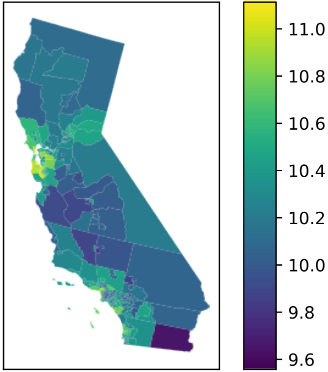
Known neighboring relation between areas $i \sim j$.

- log-Gaussian Markov Random Field prior for Λ (by column)

$$\log \boldsymbol{\lambda}^h \overset{\text{iid}}{\sim} \mathcal{N}_H \left(\mu, (\tau(F - \rho G))^{-1}\right)$$

$$G_{ij} = 1 \text{ if and only if } i \sim j,\ F_{ii} = \sum_j G_{ij}.$$

- Poor mixing when τ is random
- Poor scalability when ρ is random



Prior Modelling – the matrix Λ, California Income

Known neighboring relation between areas $i \sim j$.

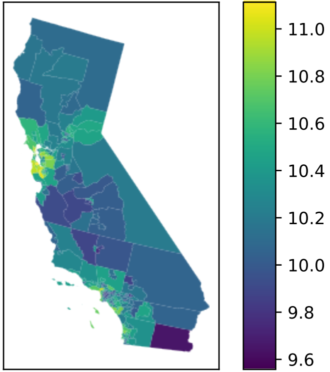
- log-Gaussian Markov Random Field prior for Λ (by column)

$$\log \boldsymbol{\lambda}^h \overset{\text{iid}}{\sim} \mathcal{N}_H \left(\mu, (\tau(F - \rho G))^{-1} \right)$$

$$G_{ij} = 1 \text{ if and only if } i \sim j,\ F_{ii} = \sum_j G_{ij}.$$

- Poor mixing when τ is random
- Poor scalability when ρ is random

Choose H via model selection.



Posterior Inference

- No close form expression of the marginal distribution or the posterior
- MCMC via slice sampling or **a priori truncation**
- Sampling of Λ and M via Hamiltonian Monte Carlo

Posterior Inference

- No close form expression of the marginal distribution or the posterior
- MCMC via slice sampling or **a priori truncation**
- Sampling of Λ and M via Hamiltonian Monte Carlo

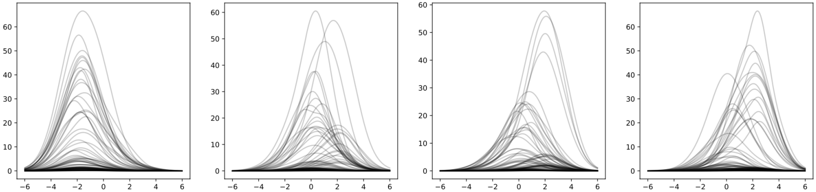
We want to **interpret**

- $\displaystyle\int_\Theta f(\cdot \mid \theta)\mu^*_h(\mathrm{d}\theta)$ (possibly after normalization) as the H *common traits*
- The matrix Λ in light of the common traits

Non-indentifiability

Recall
$$\widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh} \mu_h^* = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k^*}$$

Posterior samples from $\int_\Theta f(\cdot \mid \theta) \mu_h^*(\mathrm{d}\theta)$ in a simulated example

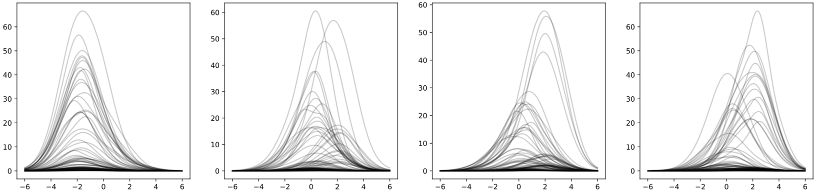


Non-indentifiability

Recall

$$\widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh}\mu_h^* = \sum_{k\geq 1}(\Lambda M)_{jk} J_k \delta_{\theta_k^*}$$

Posterior samples from $\int_\Theta f(\cdot \mid \theta)\mu_h^*(\mathrm{d}\theta)$ in a simulated example



Sources of non-identifiability

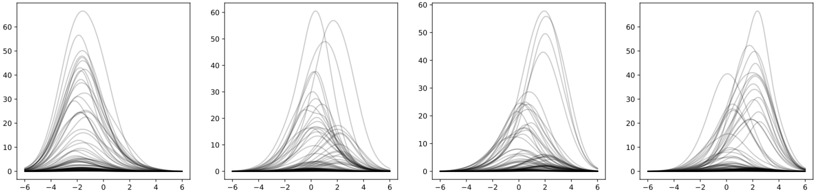
- For all invertible Q, $\Lambda' = \Lambda Q^{-1}$, $M' = QM$
- Label switching of the latent measures

Non-indentifiability

Recall

$$\widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh} \mu_h^* = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k^*}$$

Posterior samples from $\int_\Theta f(\cdot \mid \theta) \mu_h^*(\mathrm{d}\theta)$ in a simulated example



Sources of non-identifiability

- For all invertible Q, $\Lambda' = \Lambda Q^{-1}$, $M' = QM$
- Label switching of the latent measures

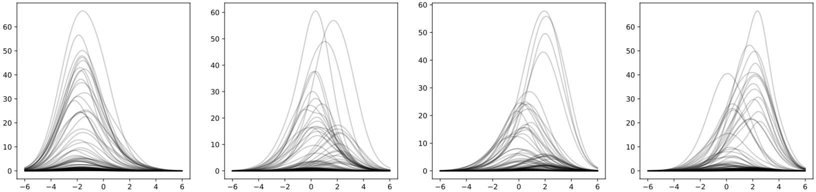
We follow Poworoznek et al. (2021) and propose a post-processing algorithm

Non-indentifiability

Recall

$$\widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh} \mu_h^* = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k^*}$$

Posterior samples from $\int_\Theta f(\cdot \mid \theta) \mu_h^*(\mathrm{d}\theta)$ in a simulated example



Sources of non-identifiability

- For all invertible Q, $\Lambda' = \Lambda Q^{-1}$, $M' = QM \implies$ find optimal Q
- Label switching of the latent measures $\implies$ optimal matching to a template

We follow Poworoznek et al. (2021) and propose a post-processing algorithm

Learning an optimal Q – What to look for?

We optimize an "interpretability criterion". Let $p'_j(y) = \sum_k (QM)_{jk} J_k f(y \mid \theta_k)$

$$Q^* = \operatorname{argmin} \mathcal{L}(Q; \mu^*, f) = \sum_{j,\ell} \langle p'_j, p'_\ell \rangle$$

- $\langle f, g \rangle$ is the L_2 inner product
- low values of $\mathcal{L}$ correspond to densities with little overlap

Learning an optimal Q – Where to look for?

Constraints

$$QM \geq 0, \qquad \Lambda Q^{-1} \geq 0$$

Learning an optimal Q – Where to look for?

Constraints

$$QM \geq 0, \qquad \Lambda Q^{-1} \geq 0$$

Q^{-1} must exist!

Learning an optimal Q – Where to look for?

Constraints

$$QM \geq 0, \qquad \Lambda Q^{-1} \geq 0$$

Q^{-1} must exist!

We don't want too extreme values in Q

Learning an optimal Q – Where to look for?

Constraints

$$QM \geq 0, \qquad \Lambda Q^{-1} \geq 0$$

Q^{-1} must exist!

We don't want too extreme values in Q

$$Q^* = \underset{Q \in SL(H)}{\operatorname{argmin}} \mathcal{L}(Q; \mu^*, f)$$

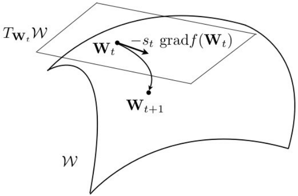
SL is the special linear group: matrices with determinants equal to 1

Learning an optimal Q – How to look for it?

SL is a Riemannian manifold. We can design a gradient descent algorithm that stays always inside SL.

Basic version (Riemannian gradient descent):

$$Q_{n+1} = Q_n \exp\left(-h\partial_Q \mathcal{L}\right), \qquad \partial_Q \mathcal{L} = (\nabla \mathcal{L})^\top$$

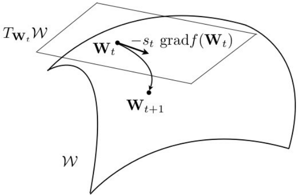


Learning an optimal Q – How to look for it?

SL is a Riemannian manifold. We can design a gradient descent algorithm that stays always inside SL.

Basic version (Riemannian gradient descent):

$$Q_{n+1} = Q_n \exp\left(-h\partial_Q\mathcal{L}\right), \qquad \partial_Q\mathcal{L} = (\nabla\mathcal{L})^\top$$



Deal with $QM \geq 0$, $\Lambda Q^{-1} \geq 0$ using an augmented Lagrangian multiplier method

$$\mathcal{L}_\rho(Q,\gamma) = \mathcal{L}(Q; M, J, \theta) + \frac{\rho}{2}\sum_j \max\left\{0, \frac{\gamma_j}{\rho}c_j(Q)\right\}$$

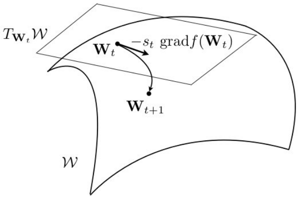
Learning an optimal Q – How to look for it?

SL is a Riemannian manifold. We can design a gradient descent algorithm that stays always inside SL.

Basic version (Riemannian gradient descent):

$$Q_{n+1} = Q_n \exp\left(-h\partial_Q \mathcal{L}\right), \qquad \partial_Q \mathcal{L} = (\nabla \mathcal{L})^\top$$

Deal with $QM \geq 0$, $\Lambda Q^{-1} \geq 0$ using an augmented Lagrangian multiplier method



$$\mathcal{L}_\rho(Q,\gamma) = \mathcal{L}(Q; M, J, \theta) + \frac{\rho}{2}\sum_j \max\left\{0, \frac{\gamma_j}{\rho} c_j(Q)\right\}$$

Alternate optimizing w.r.t. Q and w.r.t. γ_j, ρ.

Solving the Label-Switching

As in Poworoznek et al. (2021), we choose a template $\hat{\mu}_1, \ldots, \hat{\mu}_H$ and align all samples to it.

Solving the Label-Switching

As in Poworoznek et al. (2021), we choose a template $\hat{\mu}_1, \ldots, \hat{\mu}_H$ and align all samples to it.

$$\inf_{P \in \mathrm{Perm}_H} \sum_{h,k=1}^{H} d(\hat{\mu}_h, \mu_k^{(j)}) P_{hk}$$

Solving the Label-Switching

As in Poworoznek et al. (2021), we choose a template $\hat{\mu}_1, \dots, \hat{\mu}_H$ and align all samples to it.

$$\inf_{P \in \mathrm{Perm}_H} \sum_{h,k=1}^{H} d(\hat{\mu}_h, \mu_k^{(j)}) P_{hk}$$

Efficiently solved through Optimal Transport (cf. Birkhoff's theorem)

Solving the Label-Switching

As in Poworoznek et al. (2021), we choose a template $\hat{\mu}_1, \ldots, \hat{\mu}_H$ and align all samples to it.

$$\inf_{P \in \mathrm{Perm}_H} \sum_{h,k=1}^{H} d(\hat{\mu}_h, \mu_k^{(j)}) P_{hk}$$

Efficiently solved through Optimal Transport (cf. Birkhoff's theorem)

$$d(\hat{\mu}_h, \mu'_j) = \left\| \hat{\mu}_h(\Theta)^{-1} \int_\Theta f(y \mid \theta) \hat{\mu}_h(\mathrm{d}\theta) - \mu'_j(\Theta)^{-1} \int_\Theta f(y \mid \theta) \mu'_j(\mathrm{d}\theta) \right\|$$

Solving the Label-Switching

As in Poworoznek et al. (2021), we choose a template $\hat{\mu}_1, \ldots, \hat{\mu}_H$ and align all samples to it.

$$\inf_{P \in \mathrm{Perm}_H} \sum_{h,k=1}^{H} d(\hat{\mu}_h, \mu_k^{(j)}) P_{hk}$$

Efficiently solved through Optimal Transport (cf. Birkhoff's theorem)

$$d(\hat{\mu}_h, \mu'_j) = \left\| \hat{\mu}_h(\Theta)^{-1} \int_\Theta f(y \mid \theta) \hat{\mu}_h(\mathrm{d}\theta) - \mu'_j(\Theta)^{-1} \int_\Theta f(y \mid \theta) \mu'_j(\mathrm{d}\theta) \right\|$$

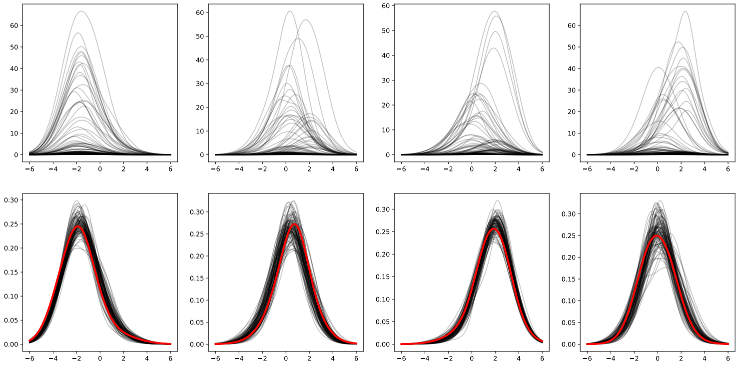
or

$$d(\hat{\mu}, \mu')^2 = \inf_{T \in \Gamma(\hat{\mu}, \mu')} \sum_{h,k=1}^{K} W_2^2(f(\cdot \mid \hat{\theta}_h), f(\cdot \mid \theta'_k)) T_{hk}$$

A Simulated Example

100 groups from

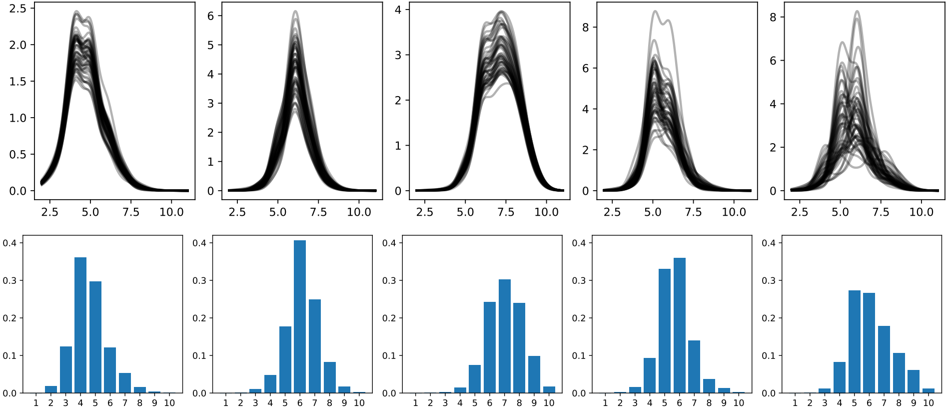
$$y_{j,i} \overset{\text{iid}}{\sim} w_{j1}\mathcal{N}(-2,2) + w_{j2}\mathcal{N}(0,2) + w_{j1}\mathcal{N}(2,2), \qquad \boldsymbol{w}_j \overset{\text{iid}}{\sim} \text{Dirichlet}(1,1,1)$$



The Invalsi Dataset

Grades of a math test in Italian high schools, 40k students in $g > 1$k schools.

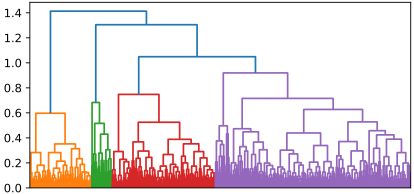
The latent measures

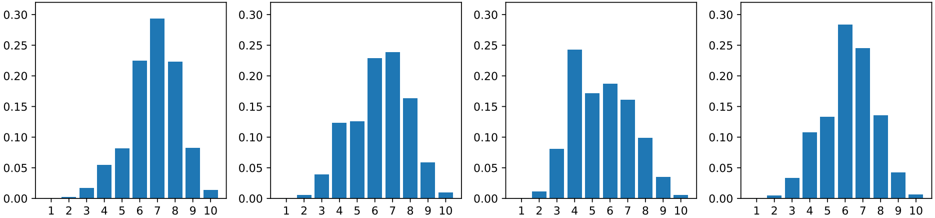


The Invalsi Dataset

Grades of a math test in Italian high schools, 40k students in $g > 1$k schools.

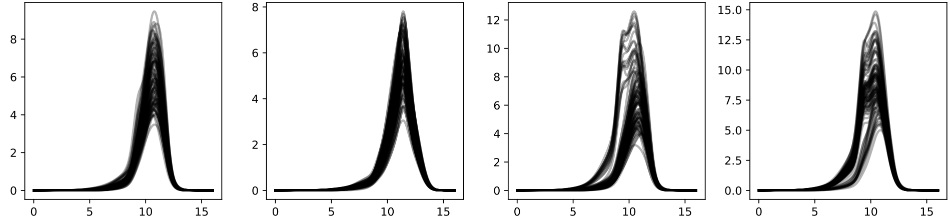
Clustering Λ's rows





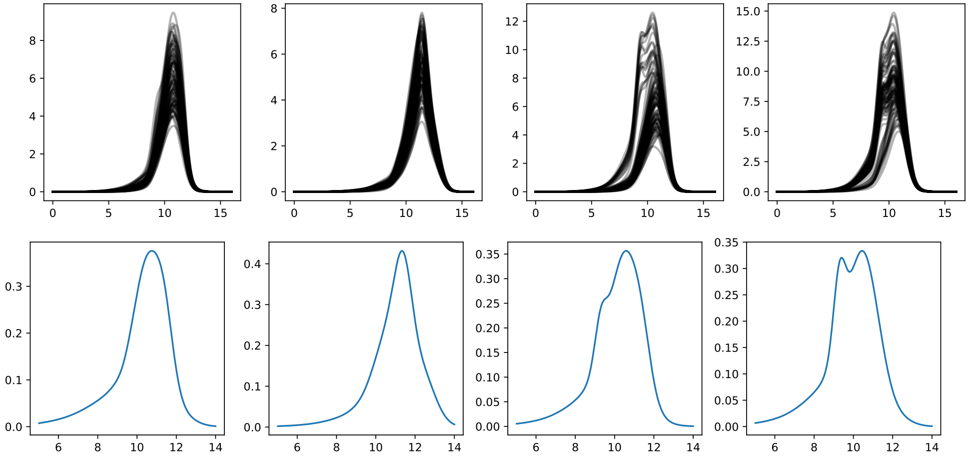
US Income Data

Personal income in $g > 250$ PUMAs in California



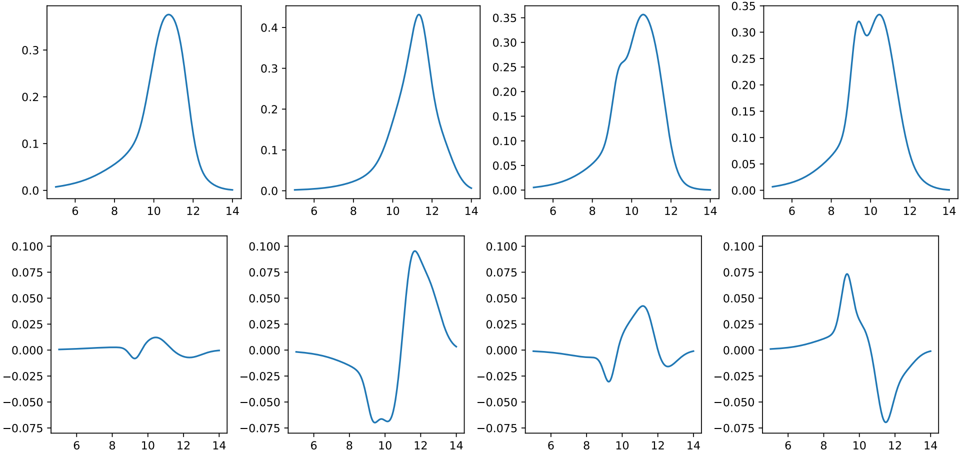
US Income Data

Personal income in $g > 250$ PUMAs in California



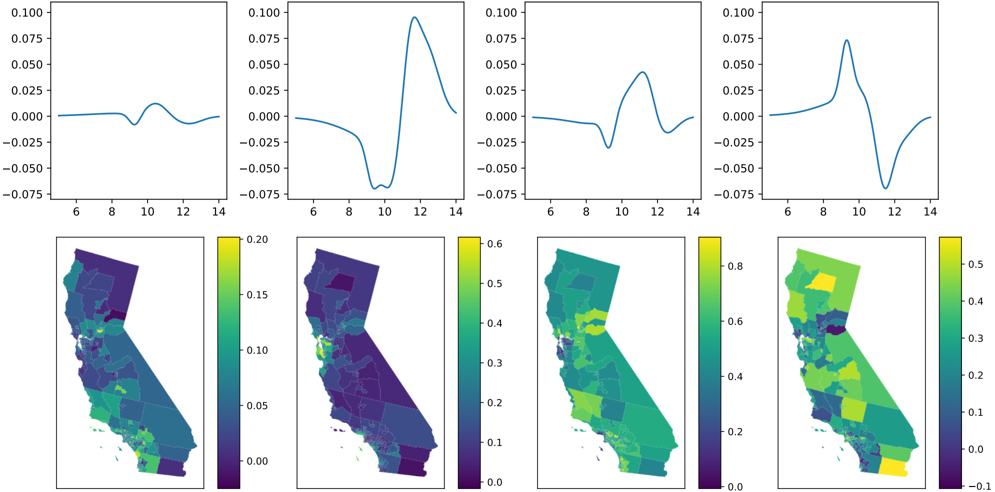
US Income Data

Personal income in $g > 250$ PUMAs in California



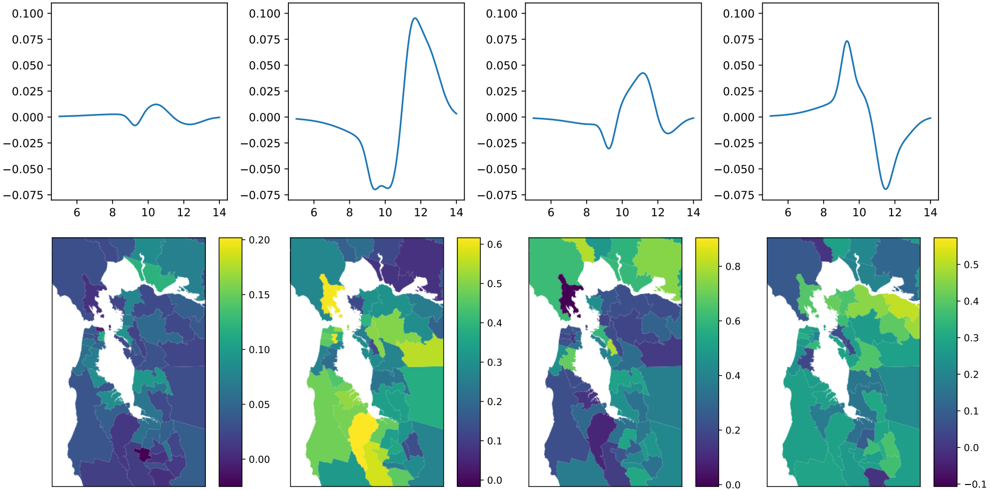
US Income Data

Personal income in $g > 250$ PUMAs in California



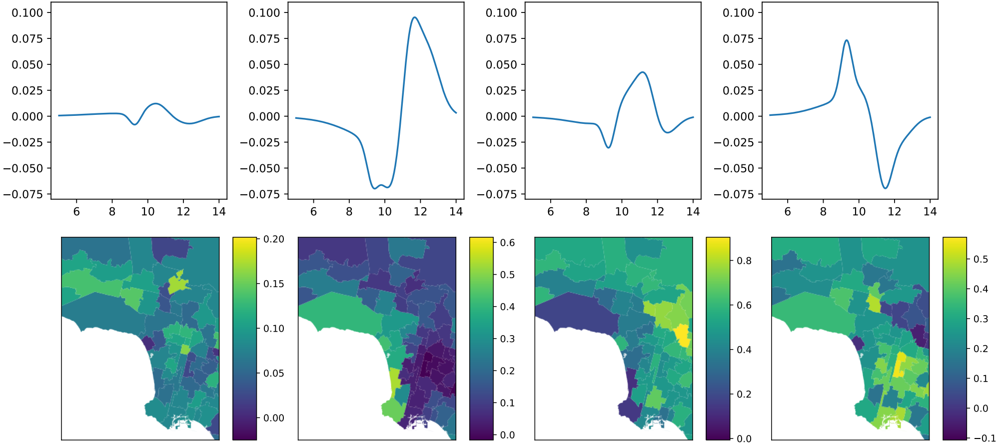
US Income Data

Personal income in $g > 250$ PUMAs in California



US Income Data

Personal income in $g > 250$ PUMAs in California



Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings
- Interpretability through post-processing

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to “big data” settings
- Interpretability through post-processing
  - Latent measures $\longrightarrow$ **latent common traits** across populations

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings
- Interpretability through post-processing
  - Latent measures $\longrightarrow$ **latent common traits** across populations
  - $\Lambda \longrightarrow$ to explore the variability

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings
- Interpretability through post-processing
  - Latent measures $\longrightarrow$ **latent common traits** across populations
  - $\Lambda \longrightarrow$ to explore the variability
- Trade-off between analytical tractability and algorithmic efficiency / scalability

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings
- Interpretability through post-processing
  - Latent measures $\longrightarrow$ **latent common traits** across populations
  - $\Lambda \longrightarrow$ to explore the variability
- Trade-off between analytical tractability and algorithmic efficiency / scalability
  - Posterior (of the μ_h^*'s) is not a CoRM. No marginal distribution in close form.

Conclusions

Normalized Latent Random Measures

- A framework for exploring difference in distribution across groups of data
- Scalable to "big data" settings
- Interpretability through post-processing
  - Latent measures $\longrightarrow$ **latent common traits** across populations
  - $\Lambda \longrightarrow$ to explore the variability
- Trade-off between analytical tractability and algorithmic efficiency / scalability
  - Posterior (of the μ_h^*'s) is not a CoRM. No marginal distribution in close form.
  - Prior elicitation carried out case-by-case

References

Beraha and Griffin. *Normalized Latent Measure Factor Models.* arXiv 2205.15654

França, Barp, Girolami, and Jordan. *Optimization on manifolds: A symplectic approach.* arXiv 2107.11231

Griffin and Leisen. *Compound Random Measures and their use in Bayesian nonparametrics.* JRSSB 2017

Griffin, Kolossiatis, and Steel. *Comparing distributions by using dependent normalized random-measure mixtures.* JRSSB 2013

Lijoi, Nipoti, and Pünster. *Bayesian inference with dependent normalized completely random measures.* Bernoulli 2014

Poworoznek, Ferrari, and Dunson. *Efficiently resolving rotational ambiguity in bayesian matrix sampling with matching.* arXiv:2107.13783

Teh, Jordan, Beal, and Blei. *Hierarchical Dirichlet Processes.* JASA (2006)

Theoretical results

Theorem

Let $(\mu_1^*, \dots, \mu_H^*)$ be a CoRM with i.i.d. scores. Denote with $\mathcal{L}_m(u) := \mathbb{E}[e^{-um}]$ the Laplace transform of the scores' distribution and with $\kappa_m(u, n) := \mathbb{E}[e^{-um} m^n]$. Then for all measurable $A \subset \Theta$

$$\mathbb{E}[\widetilde{p}_j(A)] =$$

$$\alpha(A) \sum_{h=1}^{H} \int \mathbb{E}\left[\lambda_{jh} \psi_\rho(u\lambda_{j1}, \dots, u\lambda_{jH}) \int_{\mathbb{R}_+} z \prod_{k \neq h} \mathcal{L}_m(u\lambda_{jk} z) \kappa_m(u\lambda_{jh} z, 1) \nu^*(\mathrm{d}z)\right] \mathrm{d}u$$

where ψ_ρ is the Laplace functional of $(\mu_1^*, \dots, \mu_H^*)$ (evaluated at the constant functions $u\lambda_{j1}, \dots, u\lambda_{jH}$).

Theoretical results, cont'd

Proposition

The following expression holds.

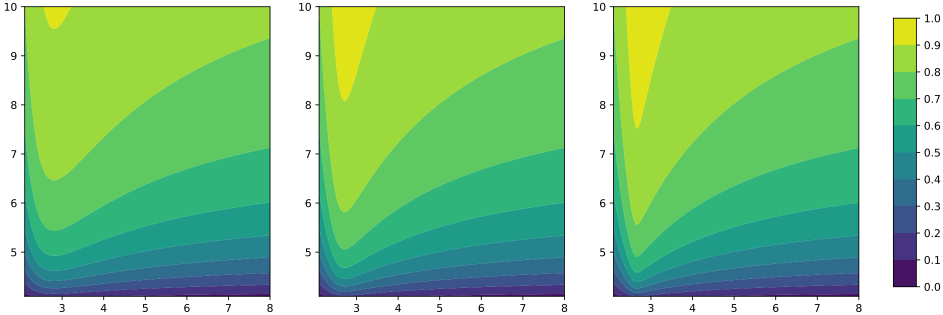
$$\mathrm{Cov}\left[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(B)\right] = \sum_{h,k} \mathbb{E}[\lambda_{jh}\lambda_{\ell k}]\mathrm{Cov}(\mu_h^*(A), \mu_k^*(B)) + \mathrm{Cov}(\lambda_{jh}, \lambda_{\ell k})\mathbb{E}[\mu_h^*(A)\mu_k^*(B)]$$

If the λ_{jh}'s have the same marginal distribution, the μ_h^*'s have the same marginal distribution, $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jH})$ and λ_ℓ (defined analogously) are independent, $\mathbb{E}[\lambda_{jh}\lambda_{\ell h}] = \kappa$, $\mathrm{Cov}(\lambda_{jh}, \lambda_{\ell h}) = \rho$ for all j, ℓ, h, then:

$$\mathrm{Cov}\left[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(B)\right] = \mathrm{Cov}(\mu_1^*(A), \mu_1^*(B))\kappa H + m_1^*(A)m_1^*(B)\rho H + \sum_{h \neq q} \bar{\lambda}_{11}^2 \mathrm{Cov}(\mu_h^*(A), \mu_k^*(B))$$

Prior Elicitation, Invalsi dataset

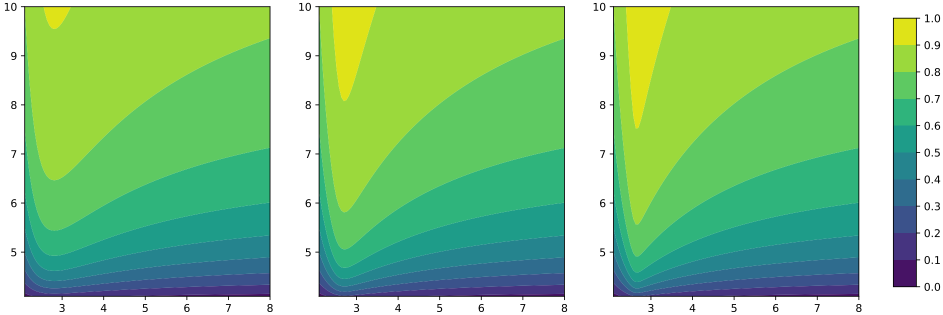
$\mathrm{Corr}(\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(A))$ for $H = 4, 8, 16$ when $a_1 = 2.5$ and $\phi = 2$.



x-axis: a_2, y-axis: ν.

Prior Elicitation, Invalsi dataset

$\mathrm{Corr}(\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(A))$ for $H = 4, 8, 16$ when $a_1 = 2.5$ and $\phi = 2$.

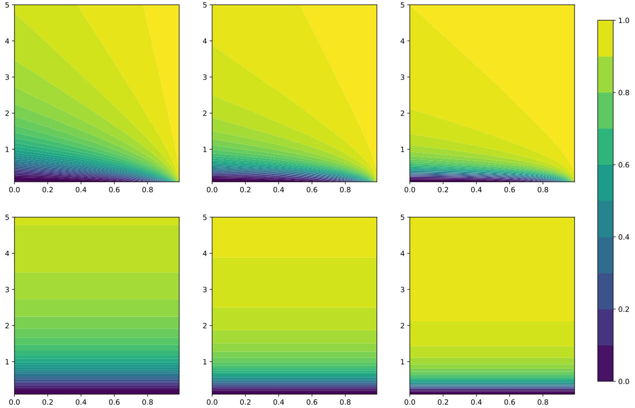


x-axis: a_2, y-axis: ν.

We set as default $a_2 = 3$ and $\nu = 5$.

Prior Elicitation, US Income

$\mathrm{Corr}(\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(A))$ for $H = 4, 8, 16$ for $j \sim \ell$ (first row) and j far apart from ℓ (second) row



x-axis: ρ, y-axis: τ.

Postprocessing Algorithm: RALM

Input: Starting point Q, initial values ρ, γ_j, target threshold ε^*, initial threshold ε

While $\varepsilon \leq \varepsilon^*$; $\|Q - Q'\| \leq \varepsilon$, **repeat**:

1. $Q = Q'$
2. solve $Q' = \arg\min_{Q} \mathcal{L}_\rho(Q, \gamma)$ for fixed ρ, γ with theshold ε
3. $\gamma_j = \gamma_j + \rho c_j(Q')$
4. $\rho = 0.9\rho \ \varepsilon = \max\{\varepsilon^*, 0.9\varepsilon\}$

Postprocessing Algorithm: Inner Optimization

Input: Starting point Q, P, momentum τ, stepsize s, threshold ε

While $\|Q - Q'\| \leq \varepsilon$, **repeat**:

1. $P = \tau\left(P - s\Pi_{\mathfrak{sl}(H)}(\partial_Q \mathcal{L}_\rho(Q,\gamma), Q)\right)$
2. $Q = Q\exp_m(\chi P)$, $\chi = \cosh(-\log\tau)$
3. $P = \tau\left(P - s\Pi_{\mathfrak{sl}(H)}(\partial_Q \mathcal{L}_\rho(Q,\gamma), Q)\right)$

Simulation on Area Referenced data

Data on a $\sqrt{g} \times \sqrt{g}$ regular lattice.

$$y_{j,i} \overset{\text{iid}}{\sim} w_{j1}\mathcal{N}(-5,1) + w_{j2}\mathcal{N}(0,1) + w_{j3}\mathcal{N}(5,1)$$

$$(w_{j1}, w_{j2}, w_{j3}) = \left(e^{\widetilde{w}_{j1}}, e^{\widetilde{w}_{j2}}, 1\right) / \left(1 + e^{\widetilde{w}_{j1}} + e^{\widetilde{w}_{j2}}\right)$$

where

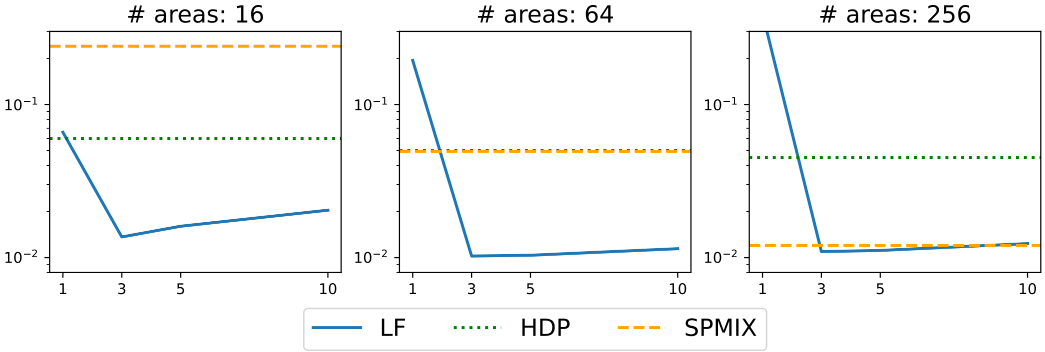
$$\widetilde{w}_{j1} = 3(x_j - \bar{x}) + 3(y_j - \bar{y}), \quad \widetilde{w}_{j2} = -3(x_j - \bar{x}) - 3(y_j - \bar{y})$$

and $(\bar{x}, \bar{y})$ denote the center of the lattice.

Simulation on Area Referenced data

Data on a $\sqrt{g} \times \sqrt{g}$ regular lattice.

$$y_{j,i} \overset{\text{iid}}{\sim} w_{j1}\mathcal{N}(-5,1) + w_{j2}\mathcal{N}(0,1) + w_{j3}\mathcal{N}(5,1)$$



Average Kullback–Leibler divergence (across all the groups)