Inside-out cross-covariance for spatial multivariate data

Michele Peruzzi
Assistant Professor of Biostatistics
University of Michigan–Ann Arbor

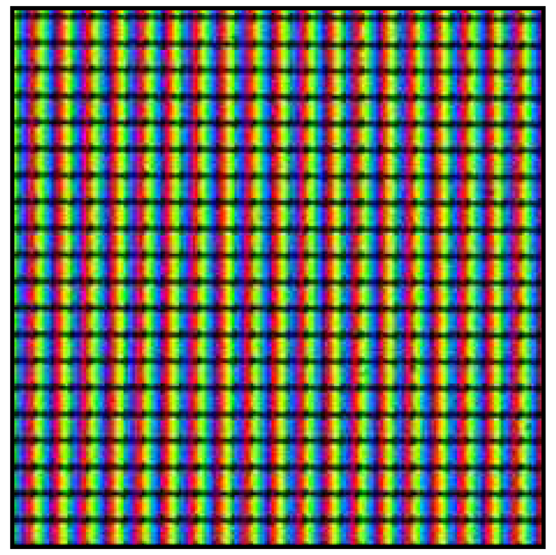
Spatial multivariate data

Features

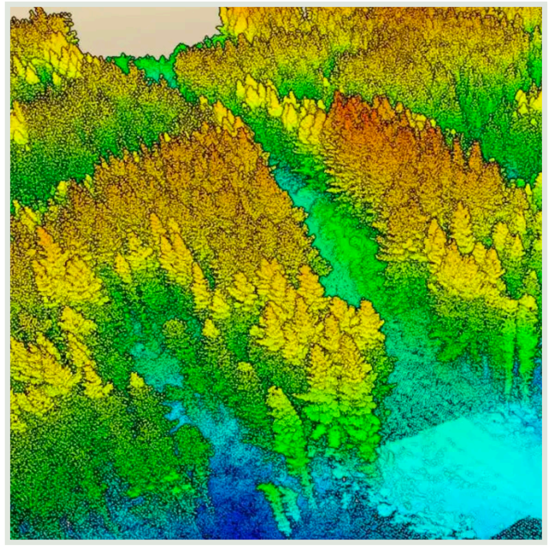
- Several variables observed over 2-D domain (earth)
- **Spatial** dependence
  - » "Near things are more related than distant things"
- **Cross-variable** dependence
  - » temperature, humidity
  - » industrial pollutants of water
  - » air quality

Sources

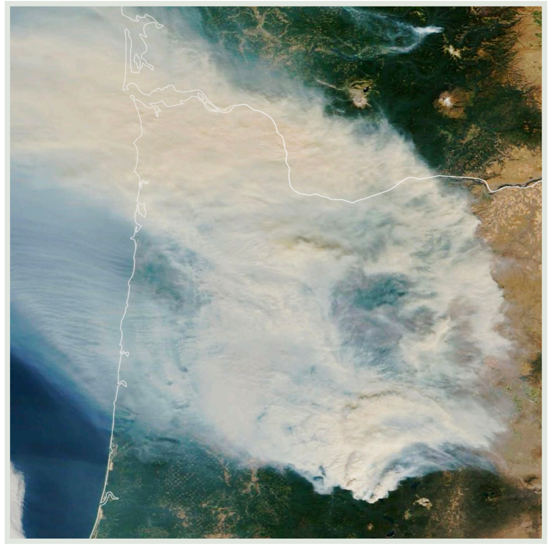
Imaging



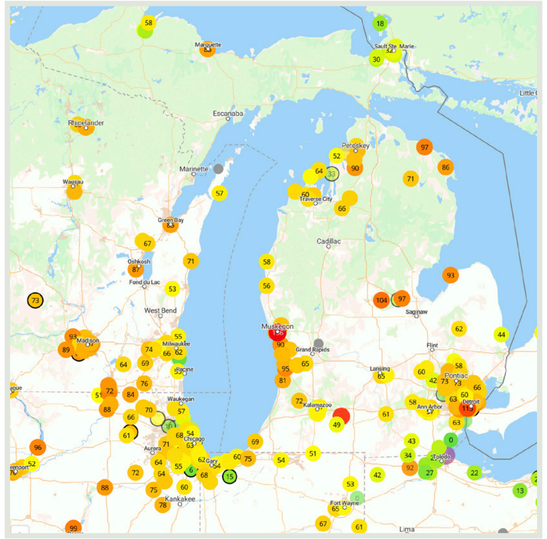
LiDAR



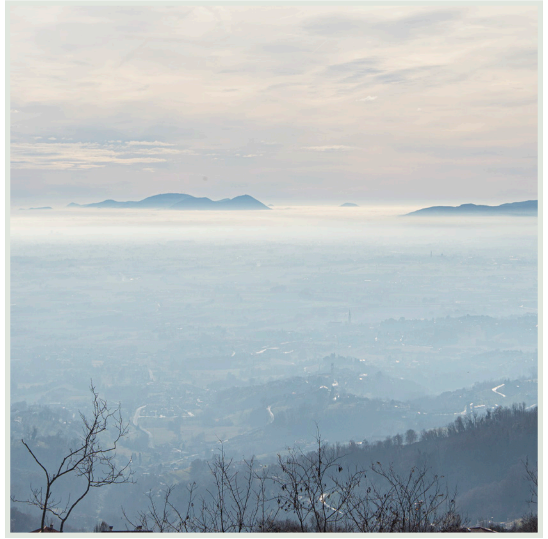
Satellite



Home sensors

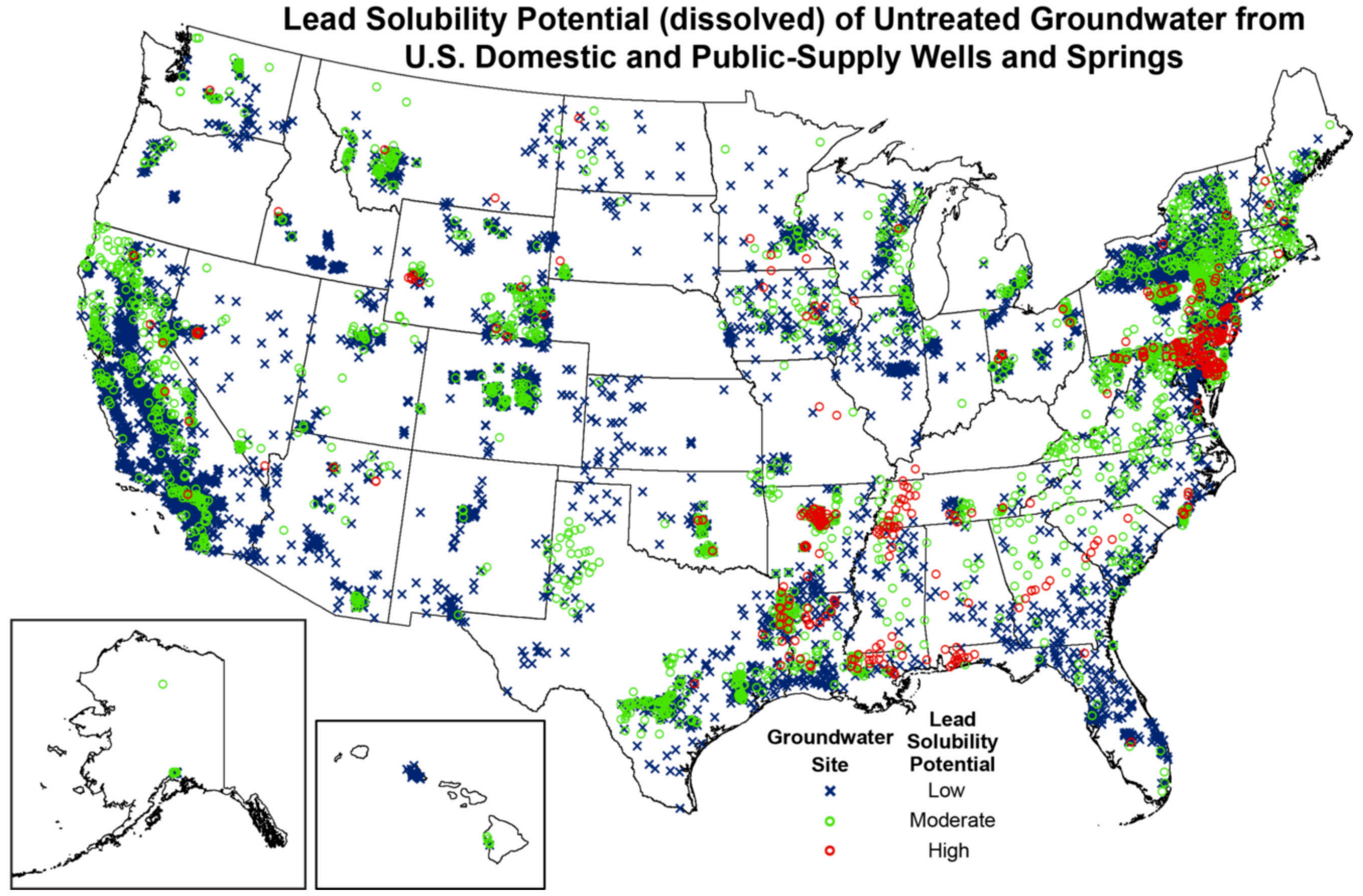


Weather stations

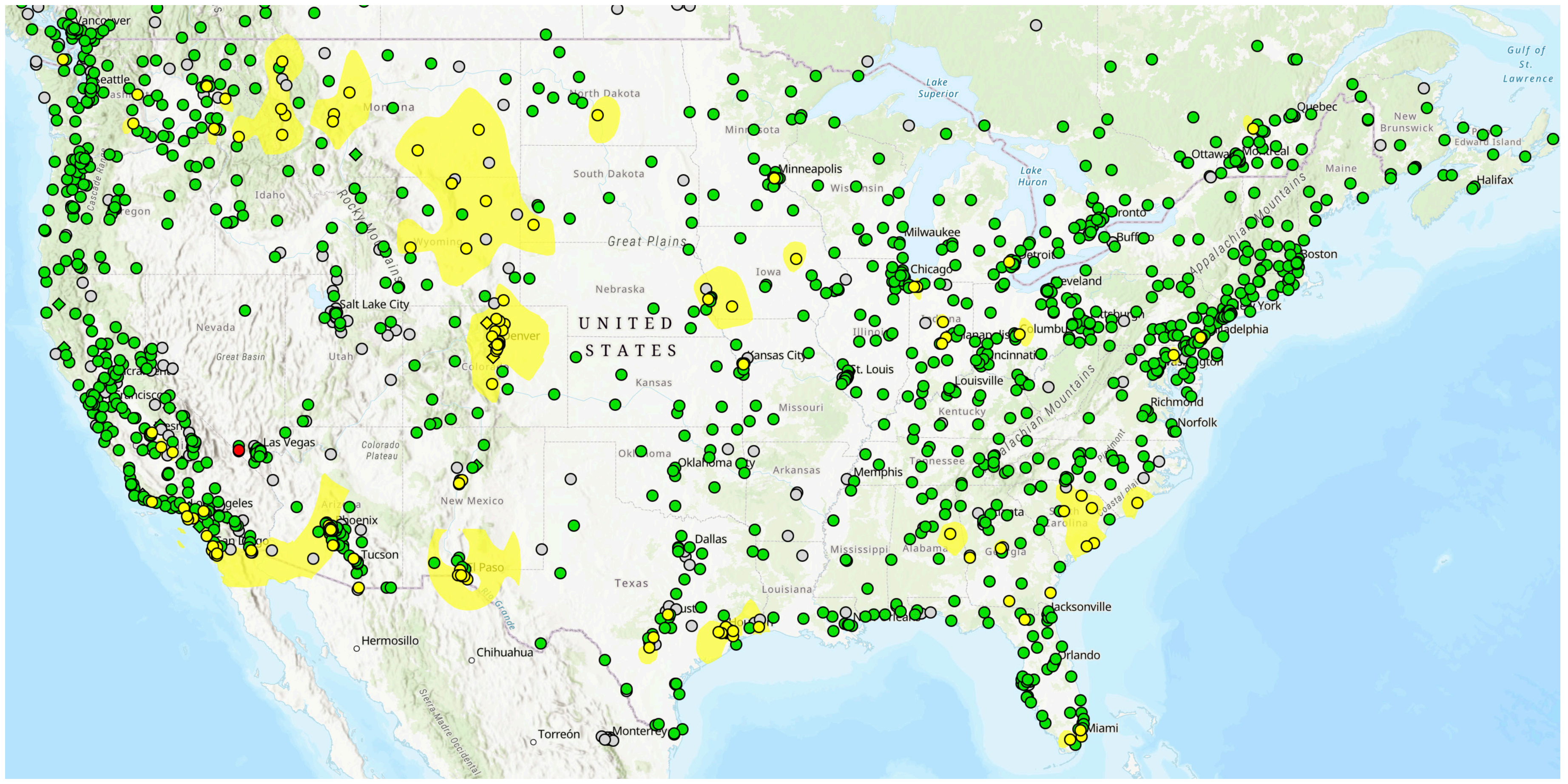


Examples

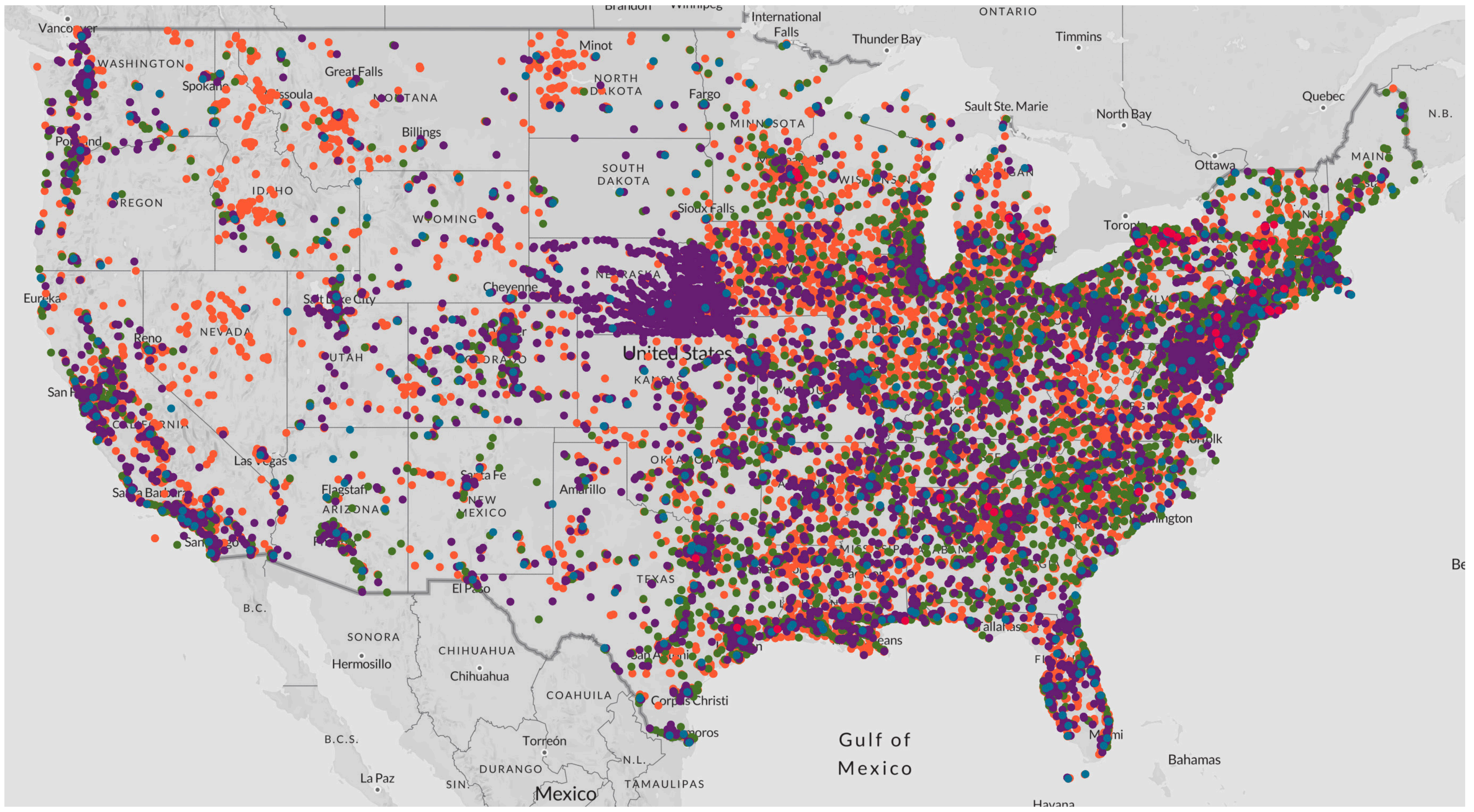
Lead



PM 2.5



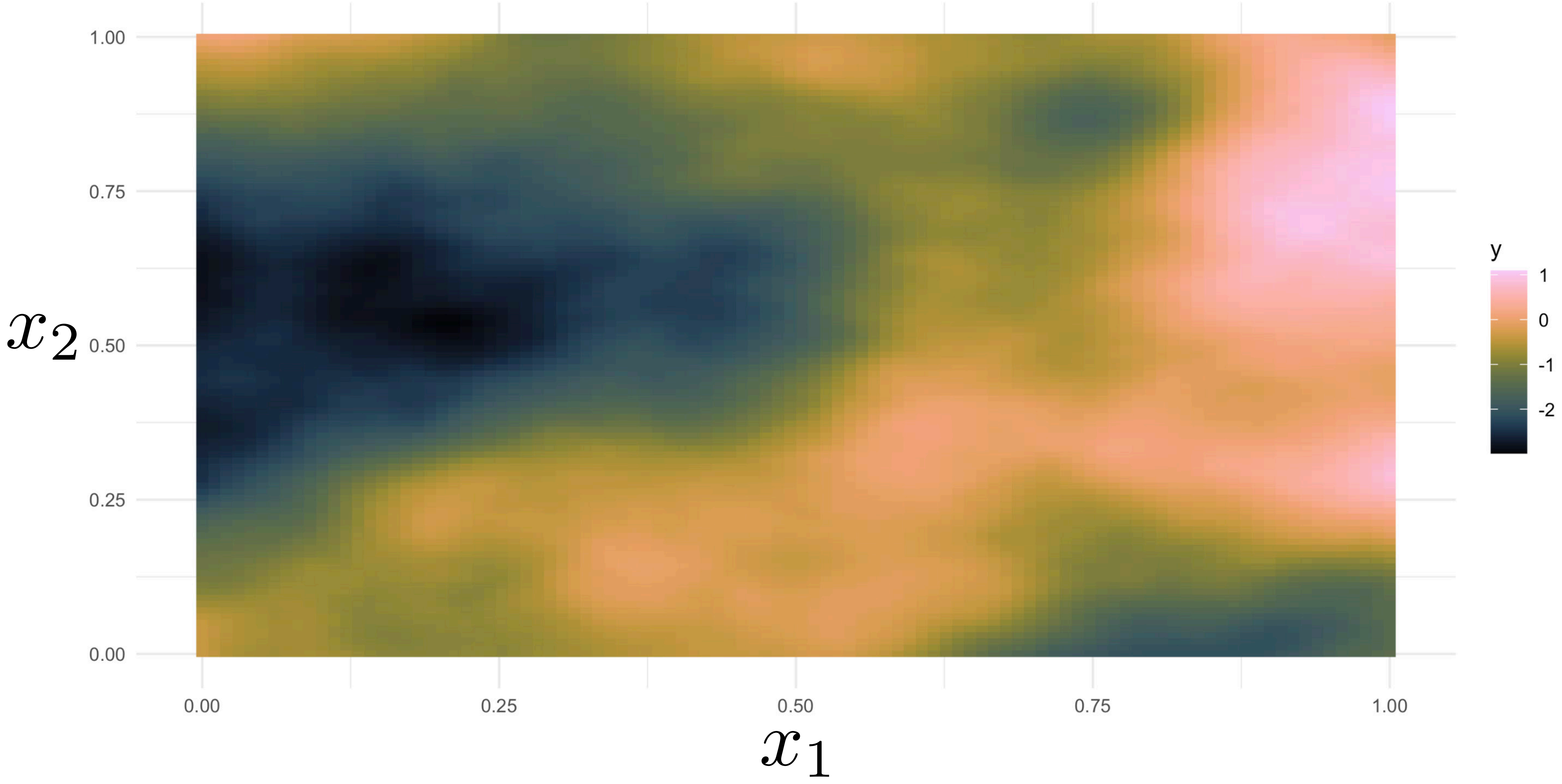
PFAS



Spatial multivariate data

Features

- Several variables observed over 2-D domain (earth)
- **Spatial** dependence
  - » "Near things are more related than distant things"
- **Cross-variable** dependence
  - » temperature, humidity
  - » industrial pollutants of water
  - » air quality



Two inputs with a non-linear interaction effect on 1 outcome variable

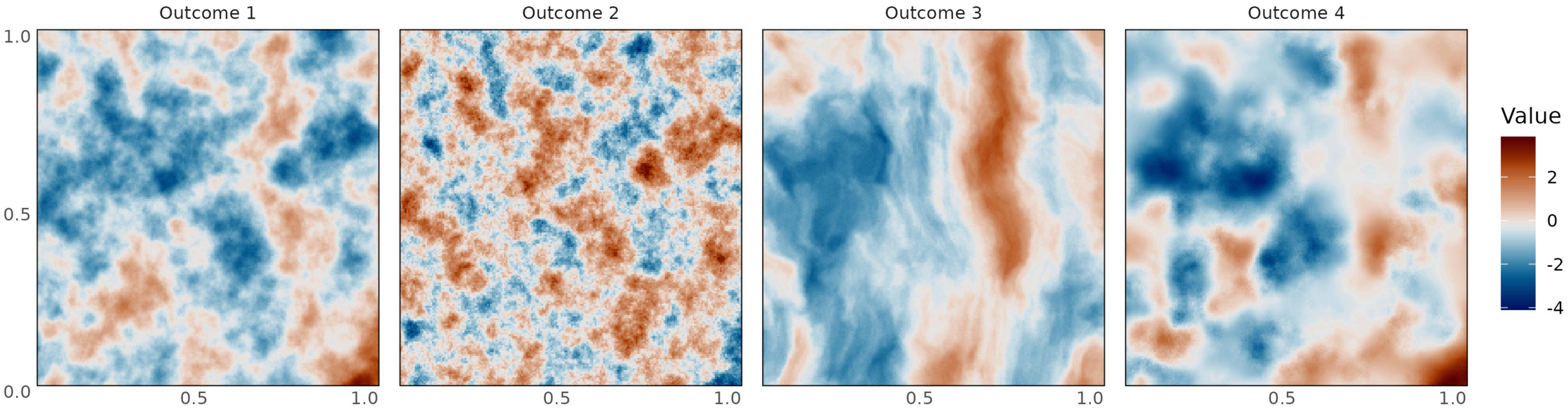
Joint model of many variables

$$y_i(\boldsymbol{x}) = f_i(\boldsymbol{x}) + \boldsymbol{\varepsilon}_i(\boldsymbol{x})$$

- $\boldsymbol{x}$ coordinates in the spatial domain
- $f_i(\cdot)$ unknown function that explains outcome $y_i(\cdot)$
- Gaussian error without spatial or cross-variable dependence $\varepsilon_i(\cdot) \overset{iid}{\sim} N(0, \sigma^2)$
- Multivariate dependence: the functions $f_i(\cdot), i = 1, \ldots, q$ are **related to each other**
- Deal with **missing data**
- Resolve **confounding** issues
- Learn graphical/**network** structure

Spatial multivariate data: example

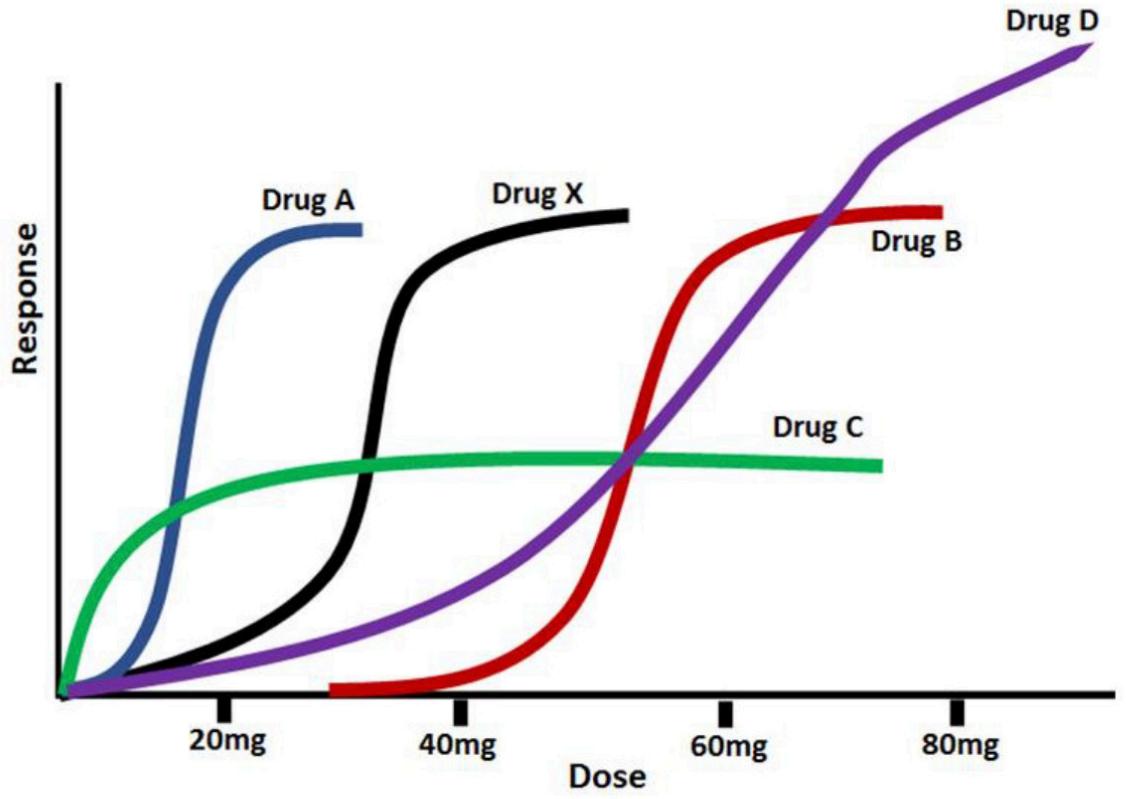
What spatial cross-correlation may look like

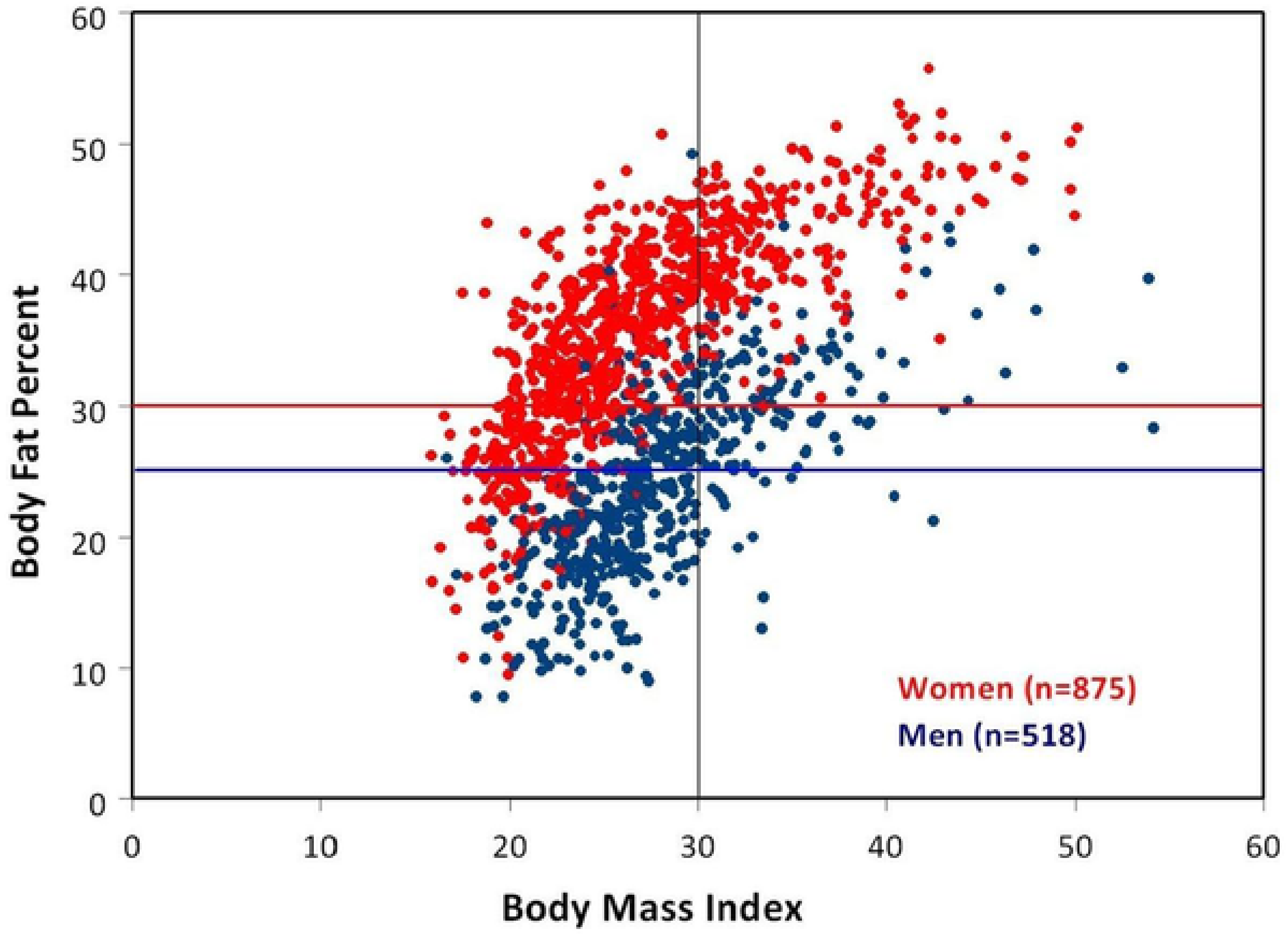


Spatial multivariate data: broader interpretation

Features

- Variables observed over p-dimensional (feature space)
- All may depend on inputs
  - » dosage of drugs
  - » mixture of exposures
  - » interactions
- Multiple outcomes are related to each other
  - » BMI, cardiovascular health, ...





Joint model of many variables

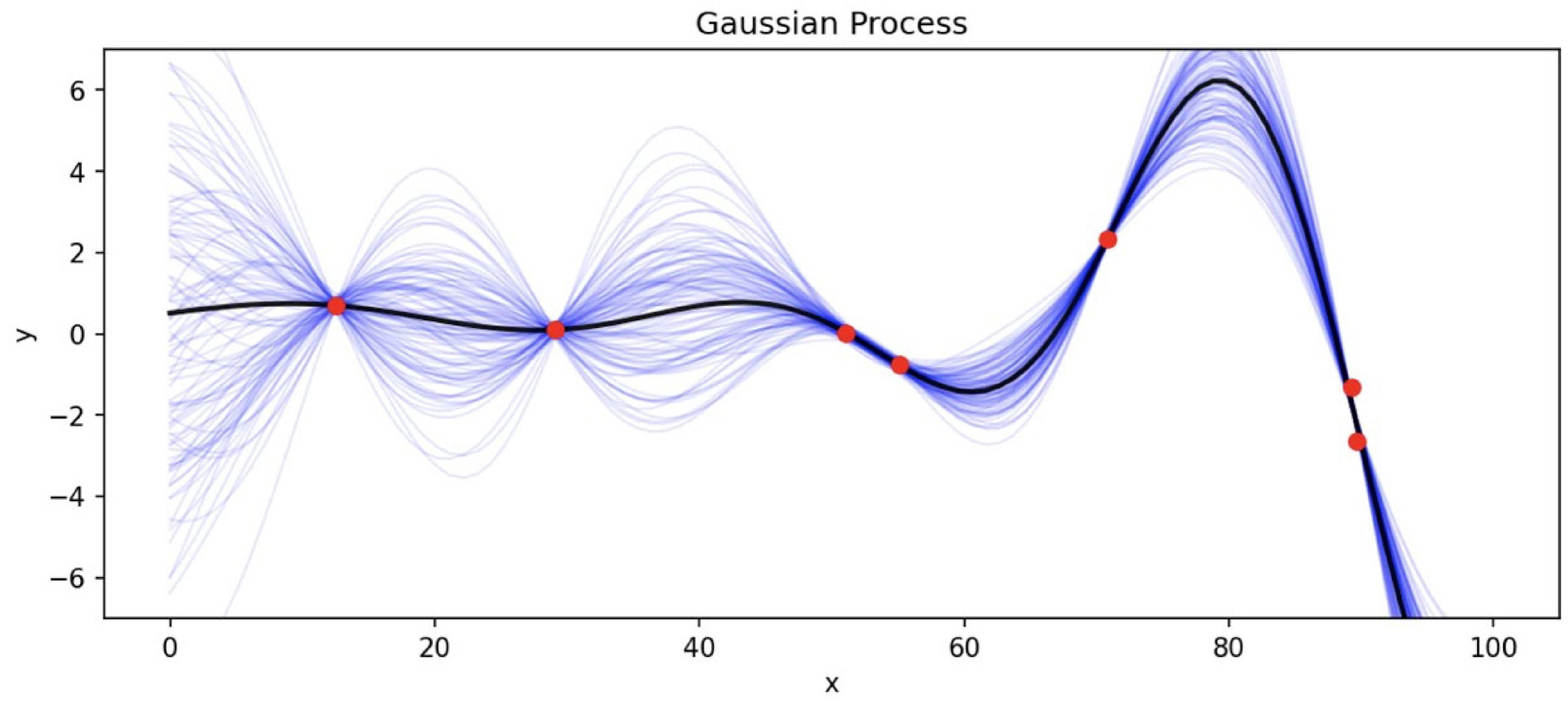
$$y_i(\boldsymbol{x}) = f_i(\boldsymbol{x}) + \boldsymbol{\varepsilon}_i(\boldsymbol{x})$$

- $\boldsymbol{x}$ coordinates in feature domain (p dimensions)
- $f_i(\cdot)$ unknown function that explains outcome $y_i(\cdot)$
- Gaussian error without spatial or cross-variable dependence $\varepsilon_i(\cdot) \overset{iid}{\sim} N(0, \sigma^2)$
- Multivariate dependence: the functions $f_i(\cdot), i = 1, \ldots, q$ are **related to each other**
- Deal with **missing data**
- Resolve **confounding** issues
- Learn graphical/**network** structure

From **univariate** to **multi-output** Gaussian Processes

- (Univariate) GP is a prior process over functions
- Completely determined by the covariance function or kernel $K_\theta(\cdot,\cdot)$
- Parametric model for $K_\theta(\cdot,\cdot)$ leads to interpretable outputs (e.g., ARD kernel and length scales)

$$f(\cdot) \sim GP(0, K_\theta(\cdot,\cdot))$$



- Multivariate or **multi-output GP** is prior over **vector-valued** functions
- Completely determined by the cross-covariance matrix function $\boldsymbol{C}_\theta(\cdot,\cdot)$
- Parametric model for $\boldsymbol{C}_\theta(\cdot,\cdot)$ leads to interpretation on each margin $f_r(\cdot)$, as well as cross-dependence, i.e. how $f_r(\cdot)$ is related to $f_s(\cdot), \quad r \neq s$

$$\begin{bmatrix} f_1(\cdot) \\ \vdots \\ f_q(\cdot) \end{bmatrix} = \boldsymbol{f}(\cdot) \sim GP(\boldsymbol{0}, \boldsymbol{C}_\theta(\cdot,\cdot))$$

Multi-output Gaussian Processes and **cross-covariance matrix functions**

$$\begin{bmatrix} f_1(\cdot) \\ \vdots \\ f_q(\cdot) \end{bmatrix} = \boldsymbol{f}(\cdot) \sim GP(\mathbf{0}, \boldsymbol{C}_\theta(\cdot,\cdot))$$

- Multivariate or **multi-output GP** is prior over **vector-valued** functions
- Completely determined by the cross-covariance matrix function $\boldsymbol{C}_\theta(\cdot,\cdot) : \Re^d \times \Re^d \to \mathcal{M}$
- $\boldsymbol{C}_\theta$ is a parametric model of cross-covariance, i.e. by construction we have

$$\boldsymbol{C}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{cov}\left\{\boldsymbol{y}(\boldsymbol{x}_i), \boldsymbol{y}(\boldsymbol{x}_j)\right\},$$

  which is a symmetric positive definite matrix of size $q \times q$
- We choose the function $\boldsymbol{C}_\theta$ and then estimate its parameters θ using the data
- Extends covariance function or kernel function to multivariate setting
- Equivalent to joint modeling of $q(q+1)/2$ covariance functions
- Must be a **valid** cross-covariance matrix function – some **conditions** need to hold
- Determines all spatial and cross-variable dependence under a GP
- For non-Gaussian or multi-type data, use latent GP in GLMM

Summary so far

Multivariate GPs are useful!

- **Nonlinear** effect of exposures (latitude, longitude, covariates) on outcomes
- **Interaction** effects of exposures on outcomes
- Joint model of exposures' effects on **multiple related outcomes**

as long as we have a cross-covariance matrix function $C_\theta(\cdot, \cdot)$ that is

- valid (!!)
- interpretable
- flexible
- useful downstream in many different settings

Unfortunately

- **Difficult** to create valid cross-covariance matrix functions
- Some valid specifications lead to
  - » **difficult computations**
  - » lack identifiability of parameters
  - » lack easy interpretations
- **Very flexible models** work only for small q (e.g., multivariate Matérn model)
- **Scalable models** are inflexible and not very interpretable

Example: **linear coregionalization** *aka* **spatial factor model**

Matheron 1982, Wackernagel 2003, Schmidt & Gelfand 2003

$$\boldsymbol{C}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\Lambda} \begin{bmatrix} \rho_1(\boldsymbol{x}_i, \boldsymbol{x}_j) & & \\ & \ddots & \\ & & \rho_k(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{bmatrix} \boldsymbol{\Lambda}^\top$$

- $\boldsymbol{\Lambda}$ is a “tall and skinny” factor loadings matrix of size $q \times k, \quad k < q$
- Each $\rho_h(\cdot, \cdot), h = 1, \ldots, k$ is a univariate correlation function
- **Easy** to build!
- Dimension reduction by choosing small k
- By far the most used model of cross-covariance
  - » model nonstationarity *Gelfand et al. 2004*
  - » spatially varying coefficients models *Gelfand et al. 2003 and Reich et al. 2010*
  - » space-time data *Berrocal et al. 2010, De Iaco et al. 2019*
  - » for non-Gaussian data *Peruzzi & Dunson 2024*
  - » scalable spatial factor models *Taylor-Rodriguez et al. 2019, Zhang & Banerjee 2022*
  - » applications in many fields *Teh et al. 2005, Finley et al. 2008, Álvarez & Lawrence 2011, Fricker et al. 2013, Moreno-Muñoz et al. 2018, Liu et al. 2022, Townes & Engelhardt 2023*
  - » software *Pebesma 2004, Finley et al. 2015, Tikhonov et al. 2020, Finazzi & Fassò 2014, Krainski et al. 2019, Peruzzi 2022*

Example: **linear coregionalization** *aka* **spatial factor model**

Matheron 1982, Wackernagel 2003, Schmidt & Gelfand 2003

$$\boldsymbol{C}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\Lambda} \begin{bmatrix} \rho_1(\boldsymbol{x}_i, \boldsymbol{x}_j) & & \\ & \ddots & \\ & & \rho_k(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{bmatrix} \boldsymbol{\Lambda}^\top$$

- $\boldsymbol{\Lambda}$ is a “tall and skinny” factor loadings matrix of size $q \times k, \quad k < q$
- Each $\rho_h(\cdot, \cdot), h = 1, \ldots, k$ is a univariate correlation function
- **Easy** to build!

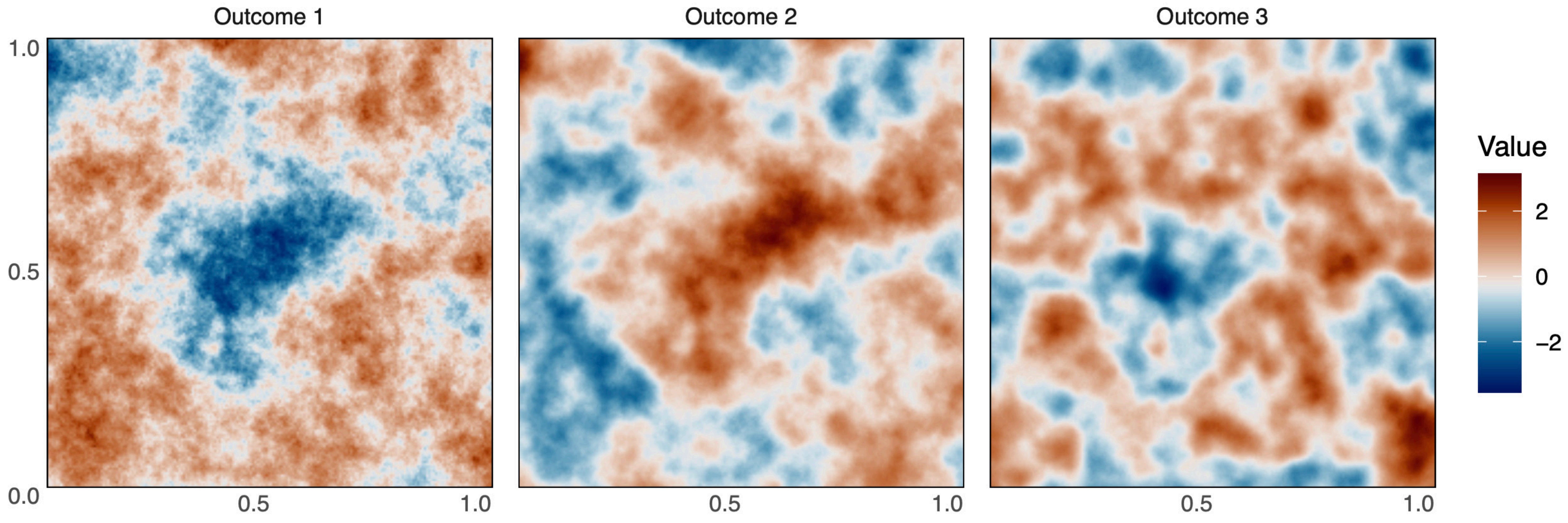
But suffers from **major problems**!

- Parameters of $\rho_h(\cdot, \cdot)$ have non-linear relationships with $C_{rs}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \text{cov}\{y_r(\boldsymbol{x}_i), y_s(\boldsymbol{x}_j)\}$
- Therefore, parameters of $\rho_h(\cdot, \cdot)$ are not directly or easily interpretable
- Cannot be used to model outcomes with different **smoothness**
  - » Smoothness plays important role in spatial confounding settings *Gilbert et al 2023*
- Cannot be used to estimate **networks** of spatial variables
- Cannot incorporate measurement error into model – must model measurement error separately
- Cannot model **outcome-specific** spatial characteristics

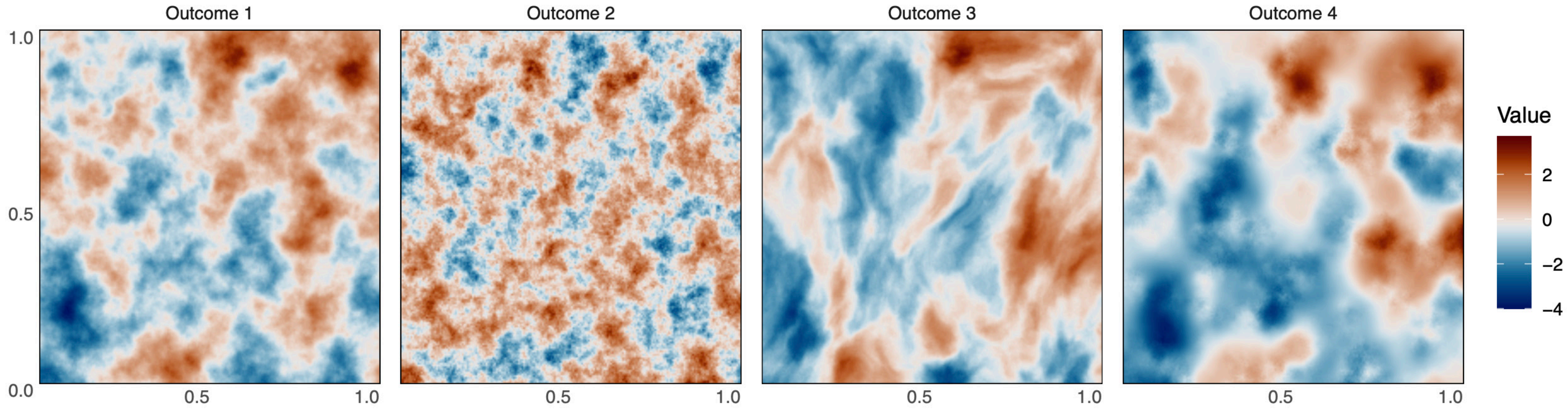
Example: **linear coregionalization** *aka* **spatial factor model**

Matheron 1982, Wackernagel 2003, Schmidt & Gelfand 2003

- Cannot be used to model outcomes with different **smoothness**



- Cannot model **outcome-specific** spatial characteristics, e.g., stationary vs nonstationary outcomes



If **coregionalization does not work**, then what do we do?

- **Multivariate Matérn model**

Gneiting et al. 2010, Apanasovich et al. 2012, Emery et al. 2022, Yarger et al. 2024

  - » Difficult conditions to check for validity
  - » Effectively only works for small q
  - » Difficult to extend to non-stationarity or other more complex spatial behavior
  - » Cannot use for dimension reduction

- **Convolution methods**

Gaspari & Cohn 1999, Majumdar & Gelfand 2007

  - » Computationally intricate
  - » May require numerical integration for each element of covariance matrix
  - » Cannot use for dimension reduction

IOX–Inside-out cross-covariance: definition

Ingredients

- Specify q univariate correlation functions (some may be the same): $\rho_r(\cdot,\cdot)$
- Specify a set of “special” locations $\mathcal{S}$. Typically, choose this as the set of observed locations
- Compute $\boldsymbol{L}_r$ such that $\boldsymbol{L}_r \boldsymbol{L}_r^\top = \rho_r(\mathcal{S})$
- Define $\boldsymbol{h}_r(\boldsymbol{x}) = \rho_r(\boldsymbol{x}, \mathcal{S})\rho_r(\mathcal{S})^{-1}$
- Define $e_r(\boldsymbol{x}) = \rho_r(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{h}_r(\boldsymbol{x})\rho_r(\mathcal{S}, \boldsymbol{x})$
- Define $\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{1}_{\{\boldsymbol{x}_i = \boldsymbol{x}_j\}} \sqrt{e_r(\boldsymbol{x}_i) e_s(\boldsymbol{x}_i)}$

Definition

- Define the r, s element of $\boldsymbol{C}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as

$$\boldsymbol{C}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_{rs} \left[\boldsymbol{h}_r(\boldsymbol{x}_i) \boldsymbol{L}_r \boldsymbol{L}_s^\top \boldsymbol{h}_s(\boldsymbol{x}_j)^\top + \varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j) \right]$$

- Entirely **new model** of cross-covariance
- **Valid** cross-covariance matrix function

from “the other” Inside Out

IOX–Inside-out cross-covariance: it's simpler than it looks...

- Specify a set of "special" locations $\mathcal{S}$. Typically equal to the set of **observed locations**
- When evaluated at $\mathcal{S}$...

$$\boldsymbol{C}_{\mathrm{IOX}} = \begin{bmatrix} \boldsymbol{L}_1 & & \\ & \ddots & \\ & & \boldsymbol{L}_q \end{bmatrix} (\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n) \begin{bmatrix} \boldsymbol{L}_1 & & \\ & \ddots & \\ & & \boldsymbol{L}_q \end{bmatrix}^\top$$

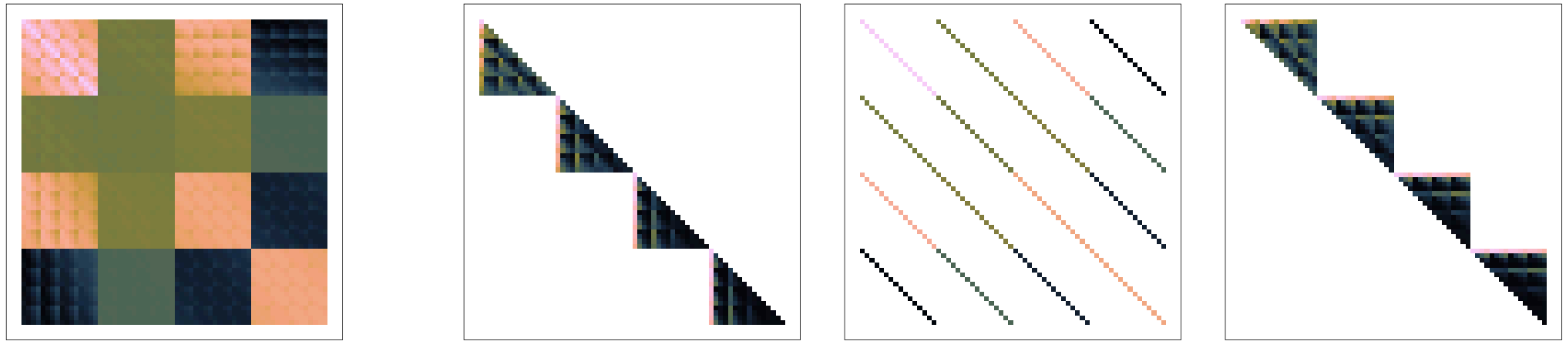
- Compare with coregionalization (LMC)

$$\boldsymbol{C}_{\mathrm{LMC}} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_n) \begin{bmatrix} \rho_1(\mathcal{S}) & & \\ & \ddots & \\ & & \rho_k(\mathcal{S}) \end{bmatrix} (\boldsymbol{\Lambda}^\top \otimes \boldsymbol{I}_n)$$

IOX–Inside-out cross-covariance: it's simpler than it looks...

- Specify a set of "special" locations $\mathcal{S}$. Typically equal to the set of **observed locations**
- When evaluated at $\mathcal{S}$...

$$C_{\text{IOX}} = \begin{bmatrix} \boldsymbol{L}_1 & & \\ & \ddots & \\ & & \boldsymbol{L}_q \end{bmatrix} (\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n) \begin{bmatrix} \boldsymbol{L}_1 & & \\ & \ddots & \\ & & \boldsymbol{L}_q \end{bmatrix}^\top$$

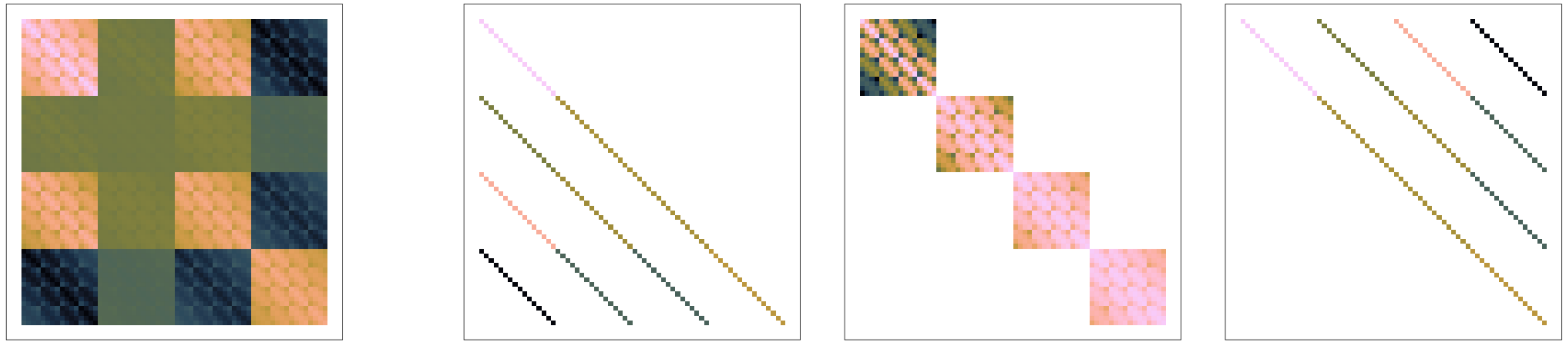


key to interpret:

$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{\Sigma}$$
$$\boldsymbol{L}_r \boldsymbol{L}_r^\top = \rho_r(\mathcal{S})$$

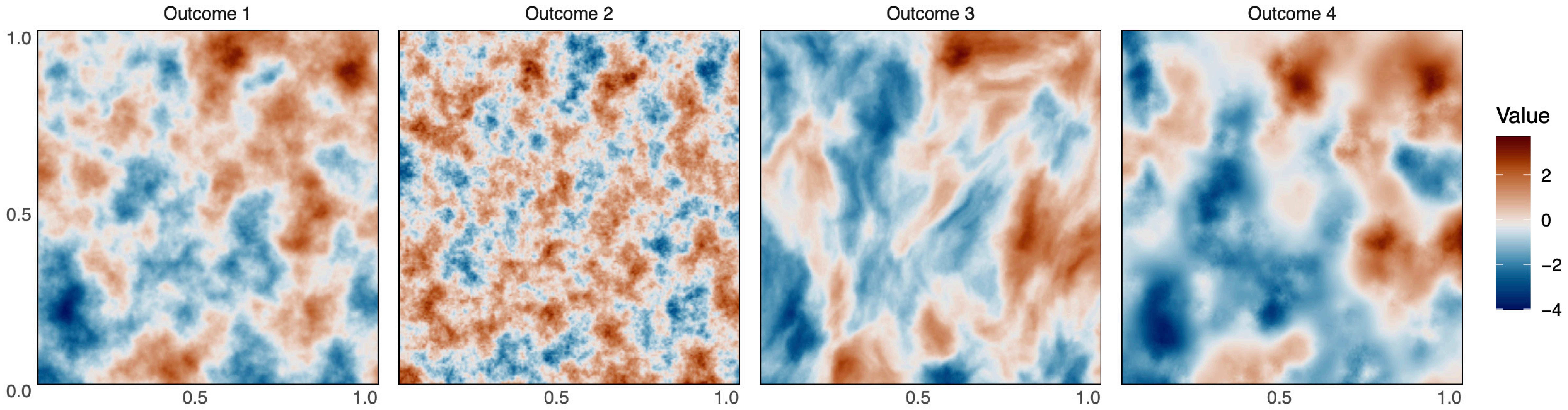
- It is "inside-out" compared to coregionalization!
- Essentially the **same ingredients**

$$C_{\text{LMC}} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_n) \begin{bmatrix} \rho_1(\mathcal{S}) & & \\ & \ddots & \\ & & \rho_k(\mathcal{S}) \end{bmatrix} (\boldsymbol{\Lambda}^\top \otimes \boldsymbol{I}_n)$$



IOX–Inside-out cross-covariance: features

- Inside-out cross-covariance is valid if each of the univariate functions are valid: **simple**!
- $\rho_r(\cdot, \cdot)$ is (essentially) the **marginal covariance for the r-th outcome**
- **Direct interpretation** of its parameters
- Can be used to introduce **outcome specific features**
- *Example:* only some outcomes are affected by some exposures
- *Example:* some outcomes exhibit non-stationarity
- *Example:* some outcomes have different smoothness



4-variable GP using IOX. Outcomes 3 and 4 are non-stationary

IOX–Inside-out cross-covariance: features

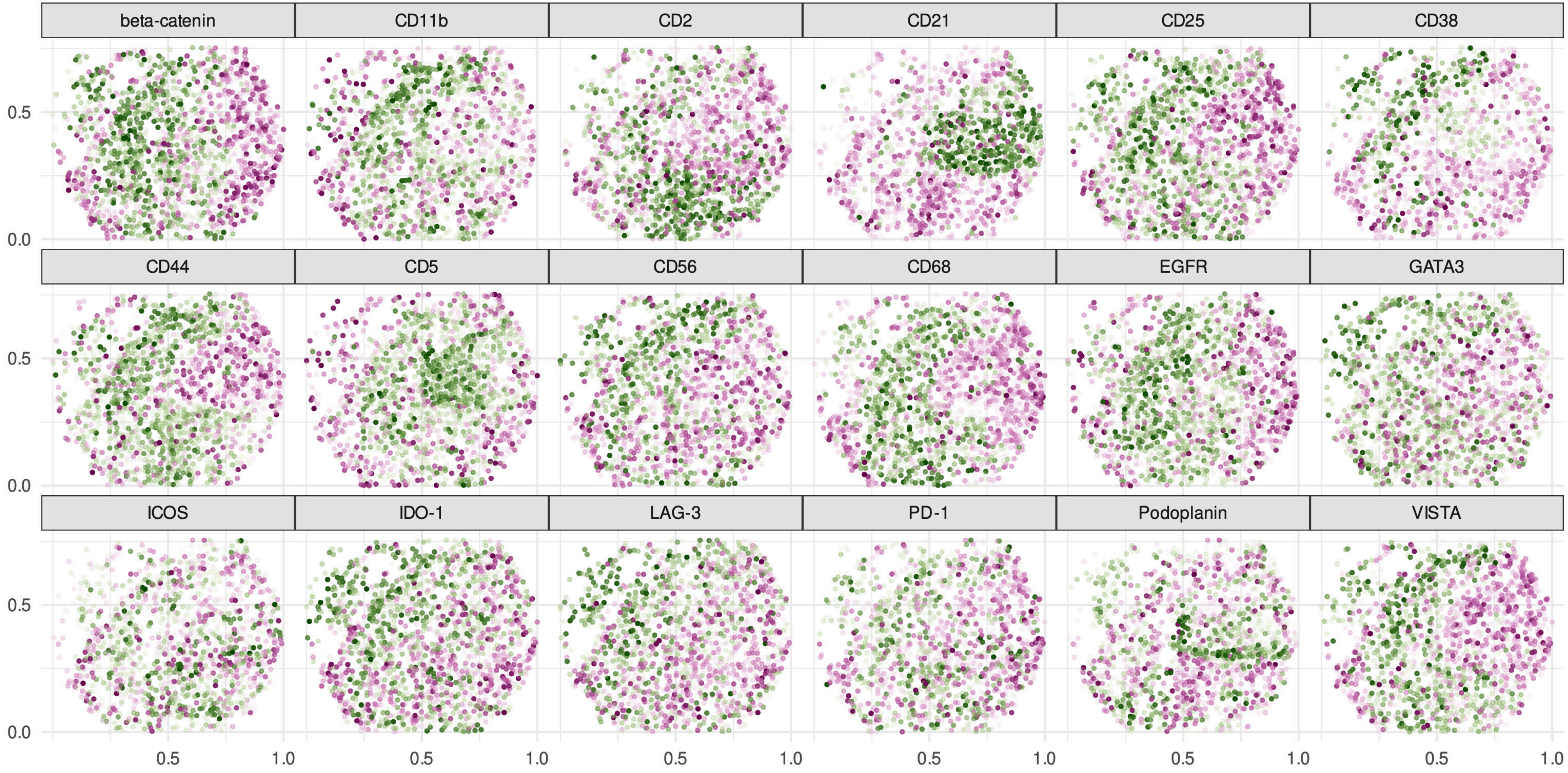
- **Simple to build**: inside-out cross-covariance is valid if each of the univariate functions are valid – easy.
- $\rho_r(\cdot, \cdot)$ is (essentially) the **marginal covariance for the r-th outcome**
- **Direct interpretation** of all parameters, e.g., outcome-specific length-scales of exposures
- Easy to specify **priors** for the parameters
- Can be used to introduce **outcome specific features**
- *Example:* only some outcomes are affected by some exposures
- *Example:* some outcomes exhibit **non-stationarity**
- *Example:* some outcomes have different **smoothness**
- Can be used for **dimension reduction**
- Can incorporate outcome-specific measurement error (**nugget** effect)
- Can be paired with **scalable** methods for GPs (low-rank, NNGP, RadGP, MGP, MRA...)
- Multiple avenues for computations
- Can model **networks** of spatial outcomes (future work)
- As easy to implement as a coregionalization model, but resolves most shortcomings

Shortcomings of IOX:

- Must choose $\mathcal{S}$
- All cross-covariances $C_{ij}(\cdot, \cdot)$ are derived indirectly and are less interpretable
- Still, $C_{ij}(\cdot, \cdot)$ in IOX is as interpretable as in a coregionalization model: must use plots.
- Intuition: IOX **prioritizes** marginal inference while **accounting for** cross-variable dependence

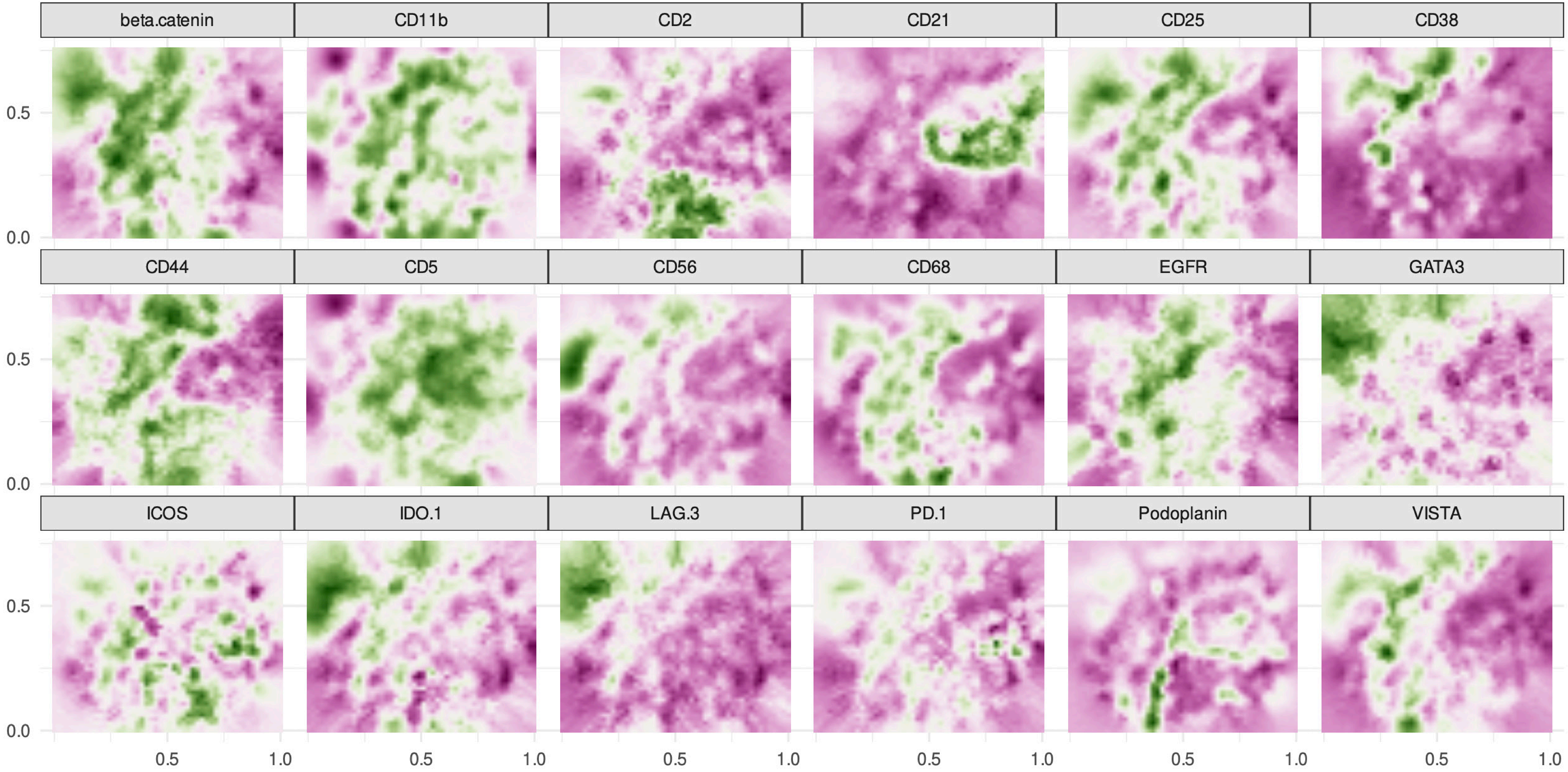
IOX–Inside-out cross-covariance: CODEX colorectal cancer data

- Take 1 patient, look at biomarker expression in tissue biopsy
- Total of 18 biomarkers with spatial dependence
- 2,873 spatial locations. Effective dimension of the problem: 51,714



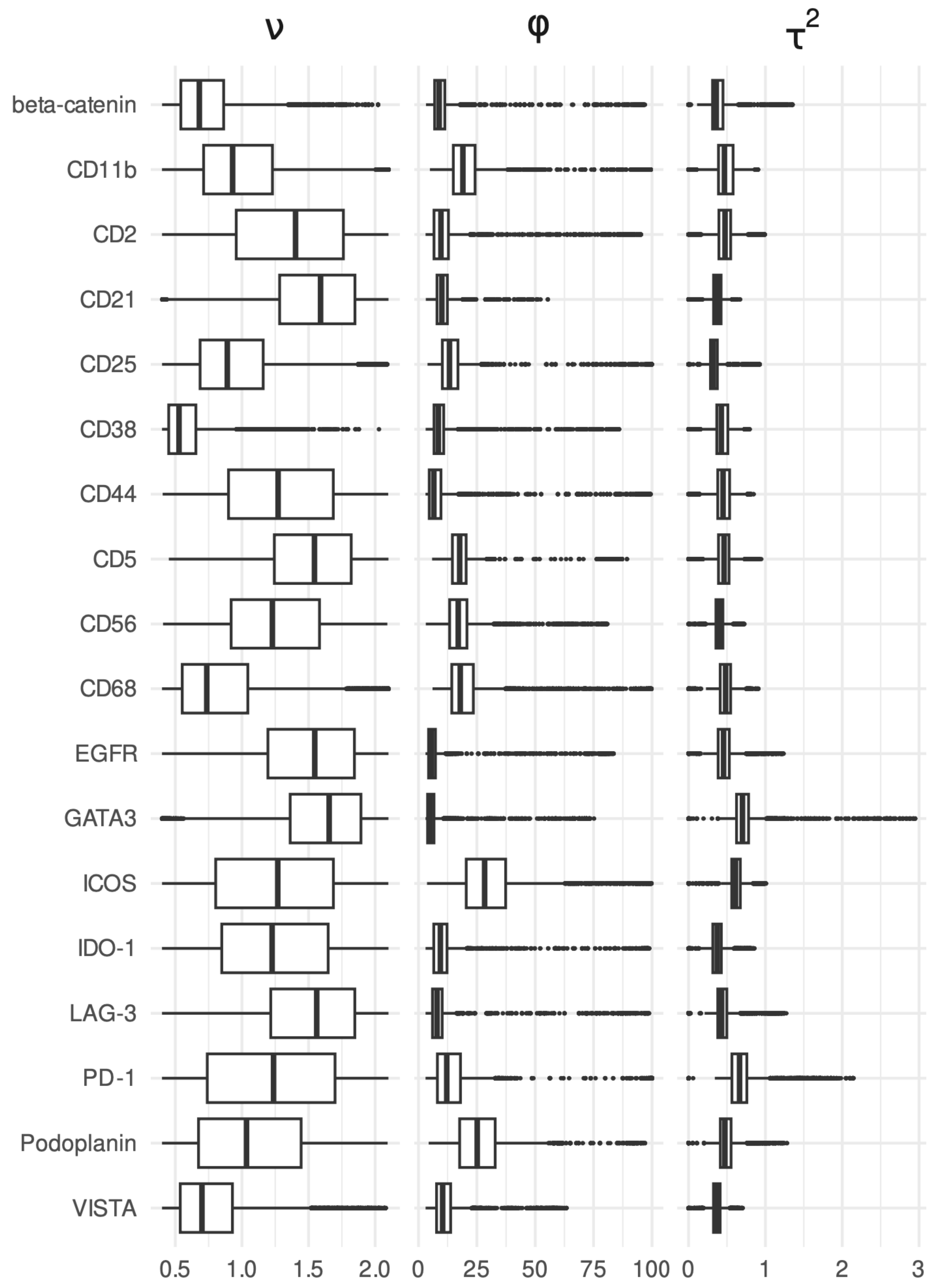
IOX–Inside-out cross-covariance: CODEX colorectal cancer data

- Estimated latent maps for biomarker expression
- Fitting time: 22 minutes



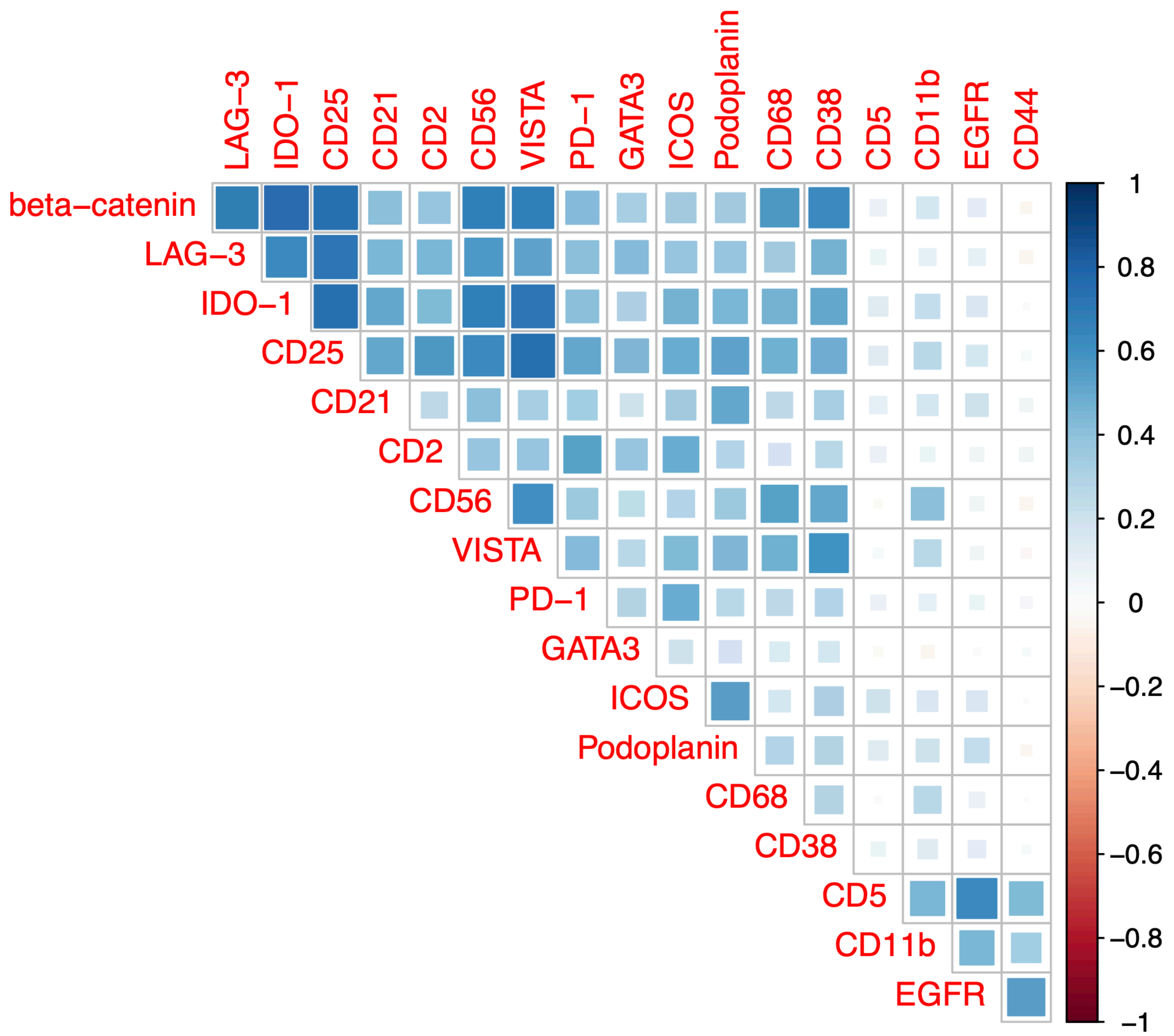
IOX–Inside-out cross-covariance: CODEX colorectal cancer data

- Estimated biomarker-specific spatial parameters (smoothness, spatial decay, error variance)



IOX–Inside-out cross-covariance: CODEX colorectal cancer data

- Estimated cross-correlation between biomarkers, at zero spatial distance



Summary, future work, comments

- IOX is a **powerful new cross-covariance model**
- Flexible, interpretable, scalable
- Easy to replace coregionalization/factor models
- Many possible avenues for **extensions and applications**
    - » joint modeling of variables to resolve confounding
    - » network learning
    - » non-stationary modeling
    - » outcome-specific variable selection in joint models

Preprint:
M Peruzzi (2024). Inside-out cross-covariance for spatial multivariate data.
https://arxiv.org/abs/2412.12407

Questions?

peruzzi@umich.edu