Bayesian Multivariate Density Regression with Coordinate-Wise Predictor Selection

Giovanni Toto[1], Peter Müller[2,3], Abhra Sarkar[2]

1: Department of Statistical Sciences, University of Padova

2: Department of Statistics and Data Sciences, University of Texas at Austin

3: Department of Mathematics, University of Texas at Austin

1 Introduction

We discuss the estimation of the joint density of a multivariate random variable in the presence of **categorical covariates**.

Focusing on density regression rather than modeling summaries, such as expected value or variance, provides a more complete description of the phenomenon of interest.

Goals:
- density estimation of bounded continuous multivariate responses,
- separate variable selection for each dimension of the response.

2 Contributions

We propose a Bayesian approach for conditional multivariate density estimation that

- employs a **Gaussian copula** to model dependence across dimensions,
- employs **mixtures of truncated normals with common atoms** to model marginal distributions,
- models covariate-dependent mixture weights through a **tensor factorization**,
- replaces mode matrices in the tensor factorization with **partitions over the covariate levels**.

3 Notation

Statistical unit i

multivariate response, $\mathbf{x}_i = (x_{1,i}, \dots, x_{d,i})^{\mathrm{T}} \in [A, B]^d$

covariate combination, $\mathbf{c}_i = (c_{1,i}, \dots, c_{p,i})^{\mathrm{T}} \in \{1, \dots, d_1\} \times \cdots \times \{1, \dots, d_p\}$

Tensor factorization for the ℓ^{th} dimension of the response

partition for covariate h, $\mathbf{s}_{\ell,h} = \{s_{\ell,h}^{(1)}, \dots, s_{\ell,h}^{(d_h)}\}^{\mathrm{T}} \in \{1. \dots, d_h\}^{d_h}$

core tensor element in position $\mathbf{m}$, $\boldsymbol{\lambda}_{\ell,\mathbf{m}} = \{\lambda_{\ell,\mathbf{m}}(1), \dots, \lambda_{\ell,\mathbf{m}}(K)\}^{\mathrm{T}} \in \Delta^{K-1}$

mapping of covariate combination $\mathbf{c}_i$, $\mathbf{s}_\ell^{(\mathbf{c}_i)} = \{s_{\ell,1}^{(c_{1,i})}, \dots, s_{\ell,p}^{(c_{p,i})}\}^{\mathrm{T}} \in \{1, \dots, d_1\} \times \cdots \times \{1, \dots, d_p\}$

4 Model specification

Gaussian copula with mixtures of truncated normals with common atoms as marginals

$$f_{\mathbf{x}|\mathbf{c}}(\mathbf{x}_i \mid \mathbf{c}_i) = |\mathbf{R}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}_i^{\mathrm{T}}(\mathbf{R}^{-1} - \mathbf{I}_d)\mathbf{y}_i\right\} \prod_{\ell=1}^{d} \left\{\sum_{k=1}^{K} \lambda_{\ell,\mathbf{s}_\ell^{(\mathbf{c}_i)}}(k) \mathsf{TN}\left(x_{\ell,i}; \mu_k, \sigma_k^2, [A, B]\right)\right\}$$

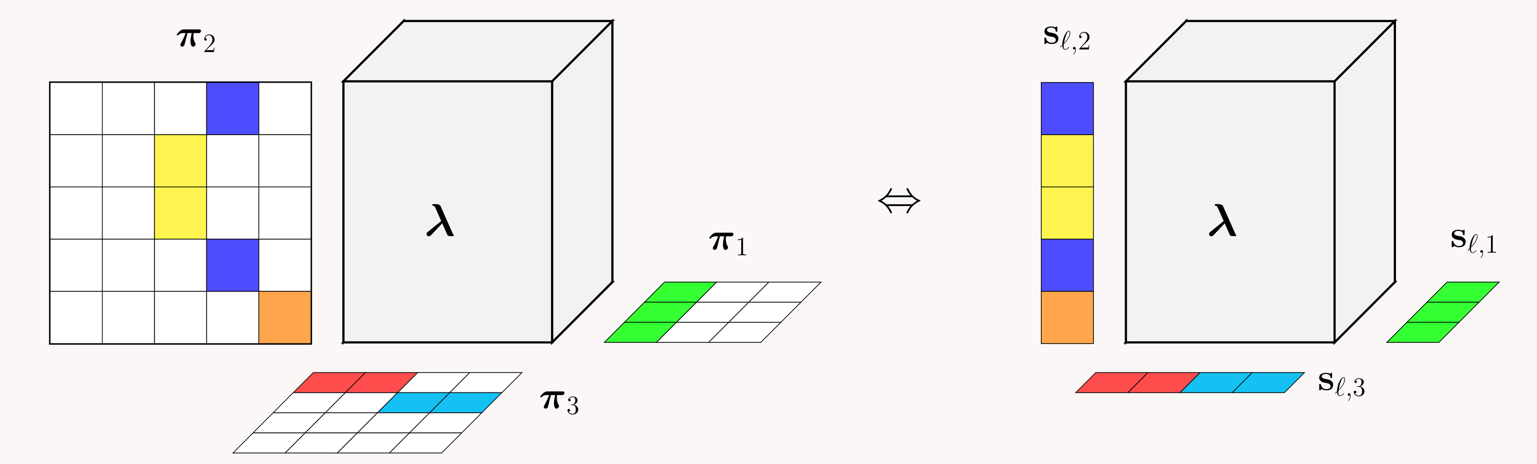
Covariate-dependent weights modeled through a tensor factorization

$$\lambda_{\ell,s_1,\dots,s_p}(k) = \sum_{m_1=1}^{d_1} \cdots \sum_{m_p=1}^{d_p} \lambda_{\ell,m_1,\dots,m_p}(k) \underbrace{\prod_{h=1}^{p} 1(s_h = m_h)}_{\text{selects a single probability}} \qquad \text{with} \quad s_h = s_{\ell,h}^{(c_{h,i})}$$

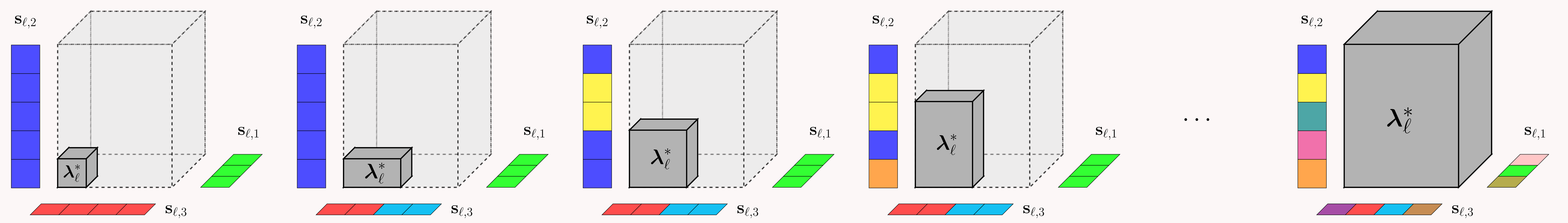
Partitions over covariate levels

$$\{s_{\ell,h}^{(1)}, \dots, s_{\ell,h}^{(d_h)}\} \mid \boldsymbol{\pi}_{\ell,h} \sim \mathsf{Cat}_{d_h}(\boldsymbol{\pi}_{\ell,h}) \times \cdots \times \mathsf{Cat}_{d_h}(\boldsymbol{\pi}_{\ell,h})$$

5 Mode matrices vs. partitions over covariate levels



6 Tensor evolution



7 Density estimation on aggregate level combinations

The model

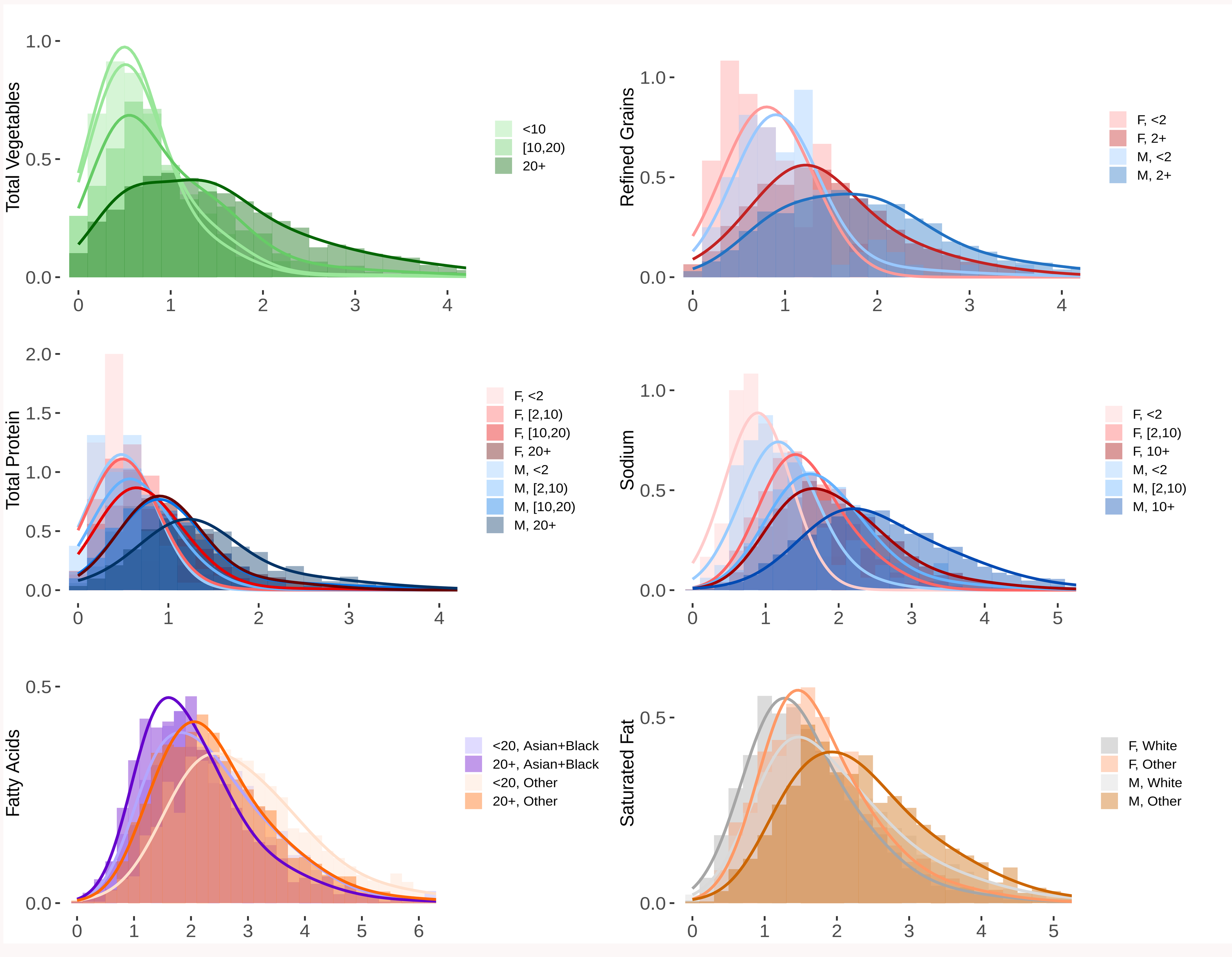
- **aggregates covariate levels** according to the dimension-specific partitions,
- defines **new covariate combinations** based on the aggregate levels,
- assign a **different mixture distribution to each of the new covariate combinations**.

remark: variable selection is a by-product of level aggregation.

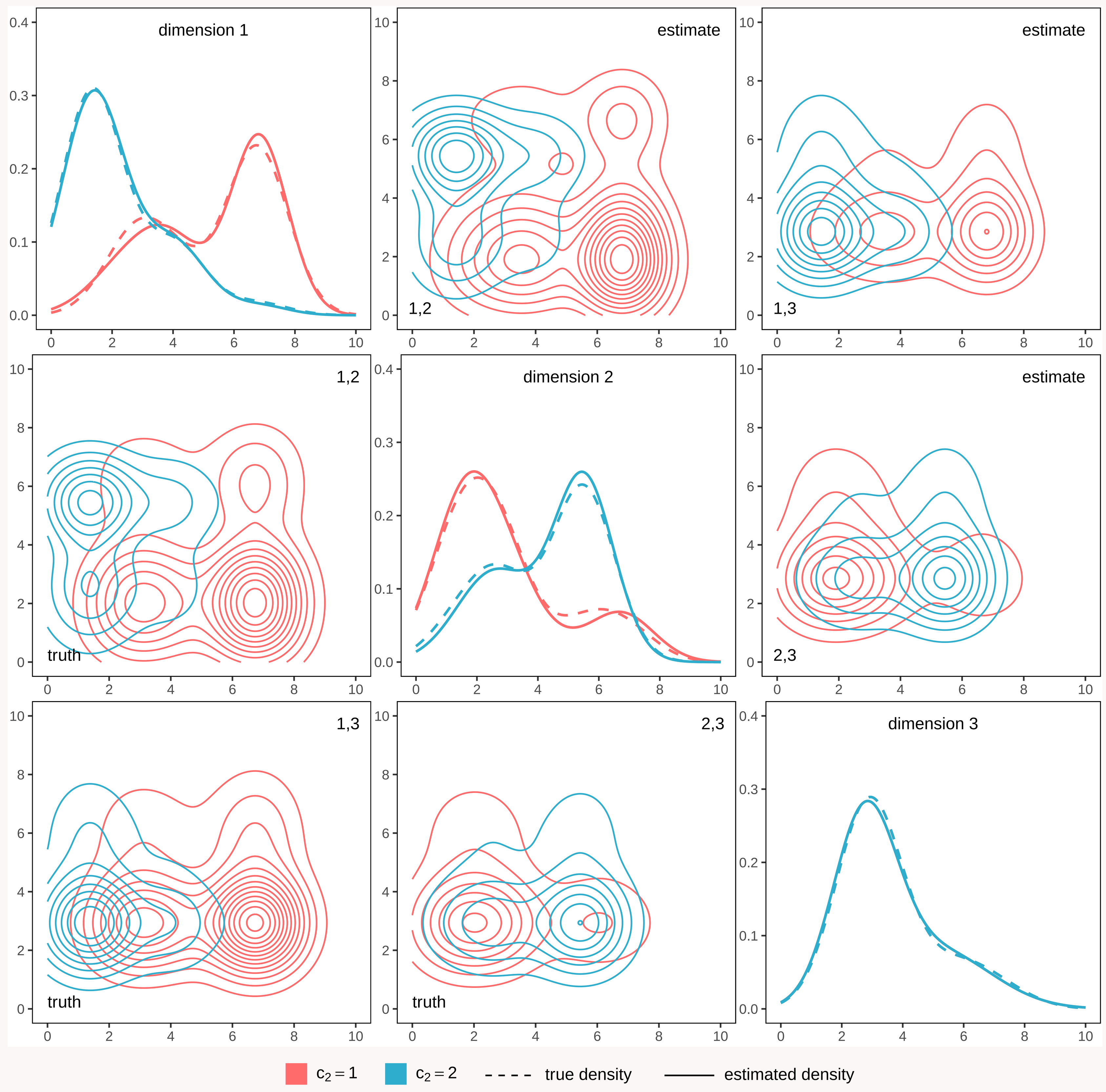
8 NHANES dietary data

Multivariate response: $d = 6$ regularly consumed dietary components

Covariates (n. of levels): sex (2), age (10), race (6), income (17)



9 Simulations



References

Yang, Y. and Dunson, D. B. (2016). Bayesian Conditional Tensor Factorizations for High Dimensional Classification. Journal of the American Statistical Association.

Sarkar, A. (2022). Bayesian Semiparametric Covariate Informed Multivariate Density Deconvolution. Journal of Computational and Graphical Statistics.