Exploring the Spatial Dependency of Gene Expression Using Markov Random Fields

Genome Informatics 2019

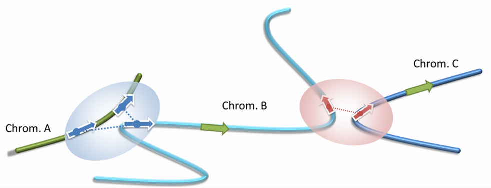
Naihui Zhou, Iddo Friedberg and Mark S. Kaiser

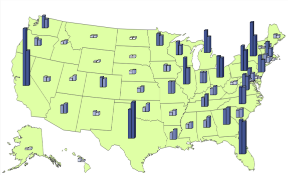
Iowa State University

Problem Statement

Hypothesis

- The quantitative expression of is are **spatially dependent**.



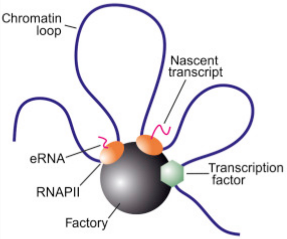


Significance

- Deeper underlying mechanisms of gene regulation could manifest as global spatial dependency

Homouz and Kudlicki (2013)

Hypotheses for the spatial clustering of genes



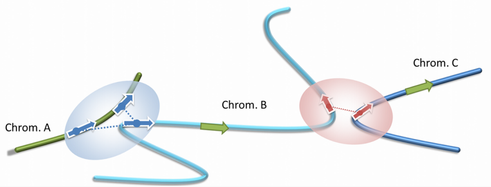
- **Transcription factories**: spatial clustering of genes for active transcription
- **Hub-enhancers**: Spatial clustering of genes to share common regulatory elements such as enhancers

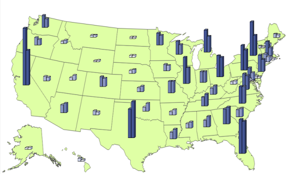
Sofiadis and Papantonis (2018)

Problem Statement

Hypothesis

- The quantitative expression of genes are **spatially correlated**.



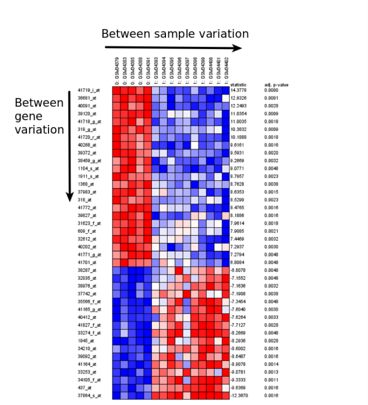


Significance

- A more comprehensive stochastic model for RNA-seq accounting for spatial dependency.

Homouz and Kudlicki (2013)

Stochastic Model for the Between-gene Variation of Gene Expression Quantification



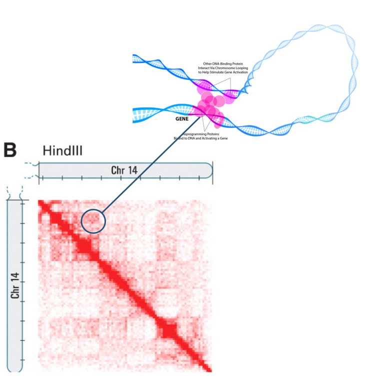
- **Differential Expression** analyses models the between-sample variation of RNA-seq data
- Borrowing information from between-gene variation (limma)
- Taking into account the spatial location of these genes enables better modelling of the between-gene variation

Al-Shahrour et al. (2005); Smyth (2005)

Poisson Hierarchical Markov Random Field Model (PHiMRF)

- Direct modeling of RNA-seq count data
- Infer spatial gene neighbors from HiC data

HiC Network Inference


Other DNA-Binding Protein
Interact Via Chromosome Looping
to Help Stimulate Gene Activation
GENE
Reprogramming Proteins
Bound to DNA and Activating a Gene
B
HindIII
Chr 14
Chr 14

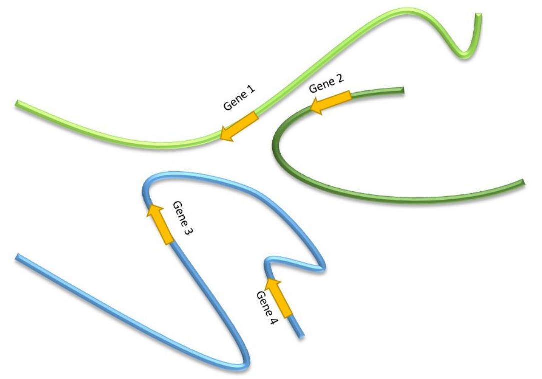
- 10kb resolution

- All possible pairwise interactions between genomic fragments, shown as heatmaps.
- Two genes are called neighbors if their HiC interaction is higher than threshold
- We infer a **spatial gene network**

Rao et al. (2014) Lieberman-Aiden et al. (2009)

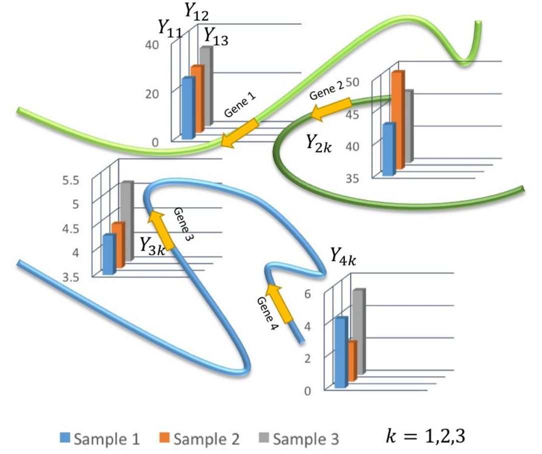
Model Specification

- Let Y_{ik} be the random variable connected with the RNA-seq count for gene i (located at s_i) from sample k, $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, M$.



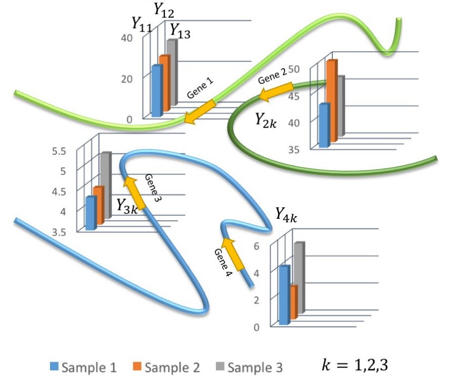
Model Specification

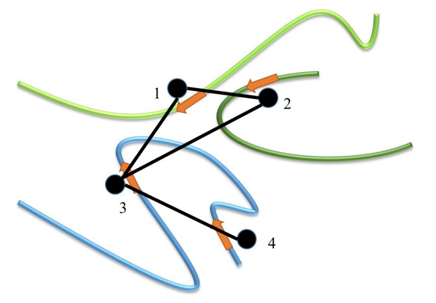
- Let Y_{ik} be the random variable connected with the RNA-seq count for gene i (located at s_i) from sample k, $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, M$.



Model Specification

- Let Y_{ik} be the random variable connected with the RNA-seq count for gene i (located at s_i) from sample k, $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, M$.





Model Specification

- Let Y_{ik} be the random variable connected with the RNA-seq count for gene i (located at s_i) from sample k, $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, M$.
- Y_{ik} follows a Poisson - lognormal mixture:
    - $Y_{ik} \sim \text{Poisson}(\lambda_i)$.
    - Let $w_i = log(\lambda_i)$.
    - $$w_i | \mathbf{w}(N_i) \sim N(\mu_i, \tau^2). \tag{1}$$

      where $N_i = \{s_j : s_j$ **is a neighbor of** $s_i\}$, and $\mathbf{w}(N_i) = \{w_j : s_j \in N_i\}$

Interpretation

MRF:

$$w_i|\boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \sum_{j \in N_i} \frac{\eta}{|N_i| + |N_j|}(w_j - \alpha).$$

Simple Linear Regression:

$$Y_i \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \beta X_i.$$

- MRFs are a type of **auto-regressive** models.

Interpretation

MRF:

$$w_i | \boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \sum_{j \in N_i} \frac{\eta}{|N_i| + |N_j|}(w_j - \alpha).$$

Simple Linear Regression:

$$Y_i \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \beta X_i.$$

- MRFs are a type of **auto-regressive** models.
- α: basal expression

Interpretation

MRF:

$$w_i | \boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \eta \sum_{j \in N_i} \frac{1}{|N_i| + |N_j|}(w_j - \alpha).$$

Simple Linear Regression:

$$Y_i \sim N(\mu_i, \tau^2),$$
$$\mu_i = \alpha + \beta X_i.$$

- MRFs are a type of **auto-regressive** models.
- α
- η: dependency parameter
  - Null hypothesis: $\eta = 0$
  - $\hat{\eta}$: **S**patial **I**nteraction **E**stimate (**SIE**)

Interpretation

MRF:

$$w_i | \boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2),$$

$$\mu_i = \alpha + \eta \sum_{j \in N_i} \frac{1}{|N_i| + |N_j|}(w_j - \alpha).$$

Simple Linear Regression:

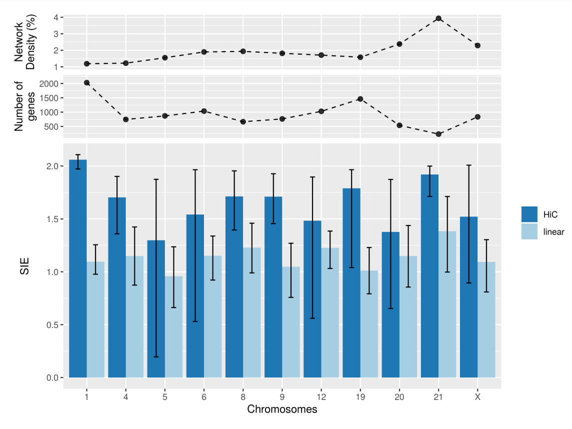
$$Y_i \sim N(\mu_i, \sigma^2),$$

$$\mu_i = \alpha + \beta X_i.$$

- MRFs are a type of **auto-regressive** models.
- α
- η: dependency parameter
  - Null hypothesis: $\eta = 0$
  - $\hat{\eta}$: **S**patial **I**nteraction **E**stimate (**SIE**)
- τ^2: residual conditional variance
- Bayesian framework with double Metropolis-Hastings MCMC

Intra-chromosomal dependency

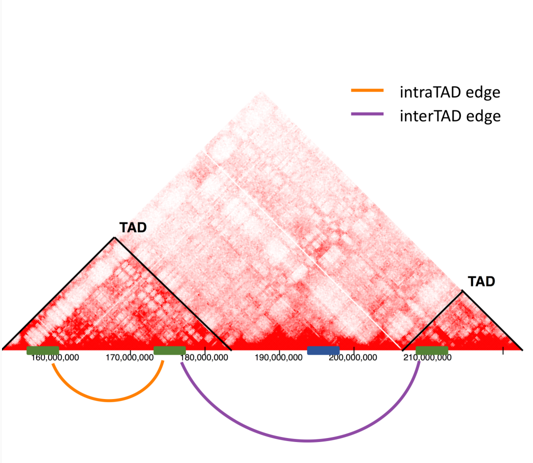
Meaningful positive spatial dependency found for Chromosomes 1, 4, 5, 6, 8, 9, 12, 19, 20, 21 and X.



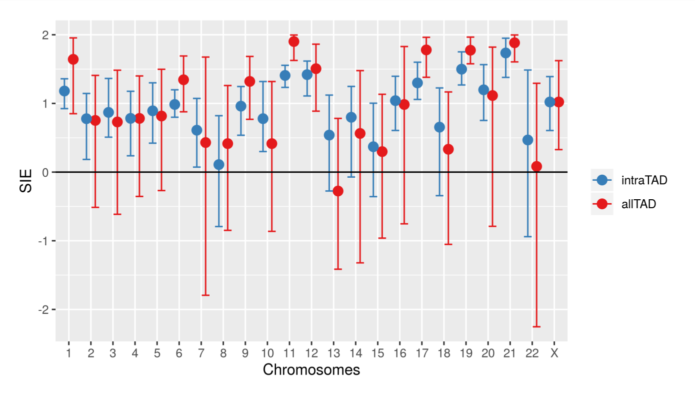
The *linear* baseline: Gene network inferred only from upstream and downstream neighboring genes, no HiC data used.

Topologically Associating Domains (TADs)

- TADs are spatial chromosomal structures with frequent interaction within.
- Frequent enhancer-promoter interactions
- Active transcription

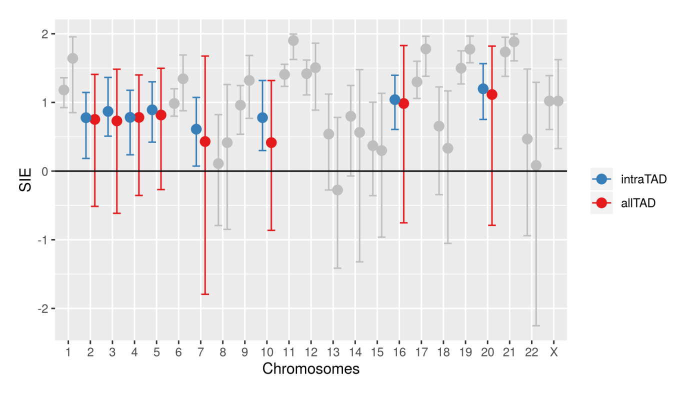


Topologically Associating Domains (TADs)



- Isolate gene neighbors within each TAD (intraTAD)
-

Topologically Associating Domains (TADs)

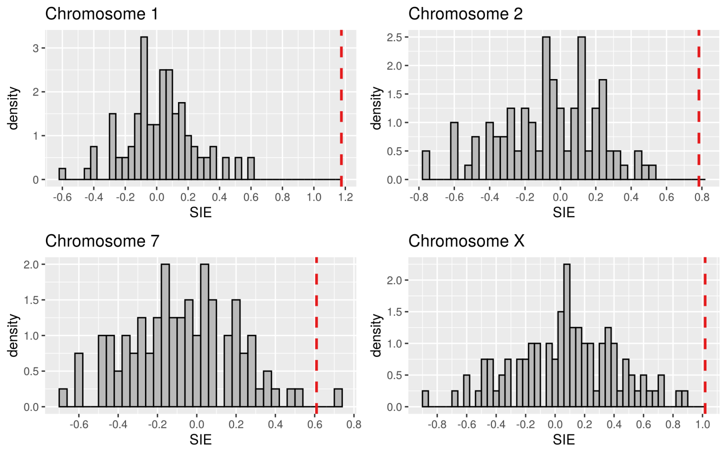


- Isolate gene neighbors within each TAD (intraTAD)
- Seventeen chromosomes show meaningful spatial dependency when considering only intraTAD neighbors, while only nine chromosomes show the same when considering all neighbors.

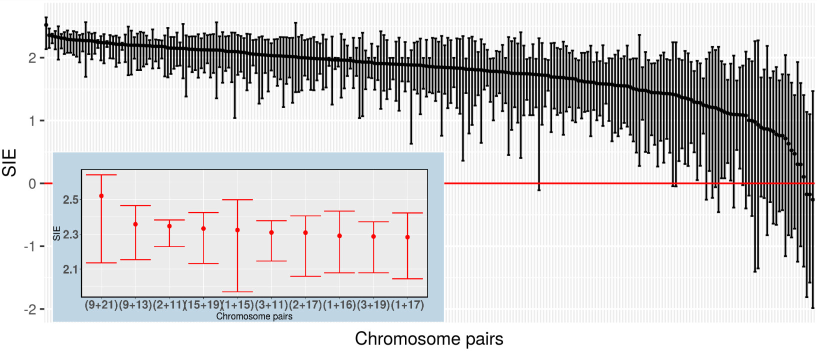
Disruption of TAD Boundaries with Permutation

- Simulate 100 random networks with the same TAD genes (nodes) but random edges (neighbors), the same number as intraTAD neighbors
- Obtain SIE for each of these networks
- Compare with SIE obtained from the observed network with edges inferred from HiC

Spatial Interaction Estimates

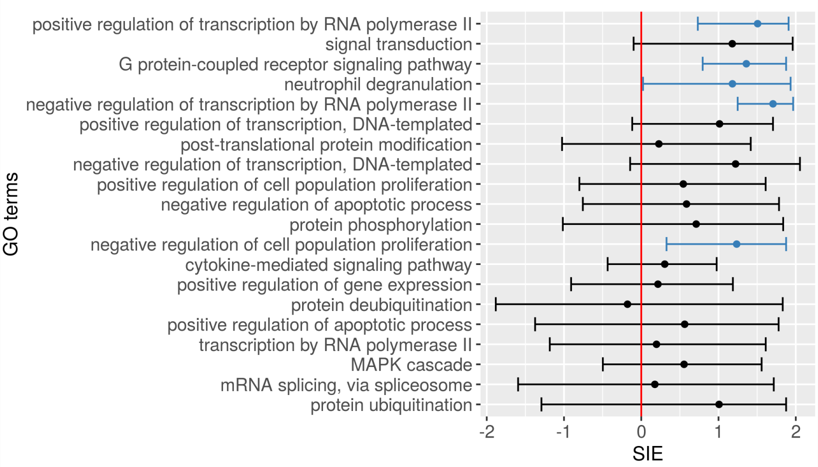


Inter-chromosomal Dependency

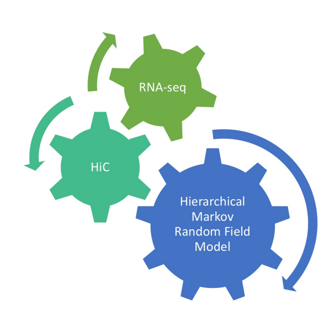


All genes and all edges: SIE = 2.561 (2.551, 2.570).

Functional gene groups



Summary



- Probabilistic model for RNA-seq data accounting for gene locations in the 3D genome
- **SIE** to estimate the strength of spatial dependency
- Global spatial dependency of gene expression detected **within chromosomes**, **between chromosomes** and in **functional gene groups**.
- Applicable to any gene groups.

- General purpose R package **PhiMRF**: https://github.com/ashleyzhou972/PhiMRF
- HiC data processing: https://github.com/ashleyzhou972/bioMRF

Acknowledgment

Major Professors

- Dr. Iddo Friedberg
- Dr. Mark Kaiser

Friedberg Lab Members

- Md Nafiz Hamid
- Huy Nguyen
- Parnal Joshi
- Xiao Hu

References i

Fatima Al-Shahrour, Pablo Minguez, Juan M Vaquerizas, Lucía Conde, and Joaquín Dopazo. Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic acids research*, 33(suppl_2): W460–W464, 2005.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Dirar Homouz and Andrzej S Kudlicki. The 3d organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PLoS One*, 8(1):e54699, 2013.

Mark S Kaiser and Noel Cressie. The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220, 2000.

References ii

Faming Liang. A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

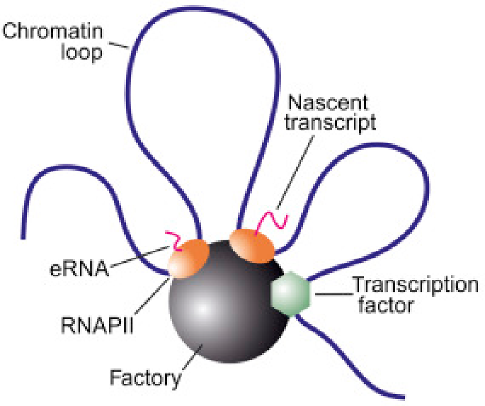
Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

References iii

Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

Konstantinos Sofiadis and Argyris Papantonis. Transcription factories as spatial and functional organization nodes. In *Nuclear Architecture and Dynamics*, pages 283–296. Elsevier, 2018.

Transcription factories

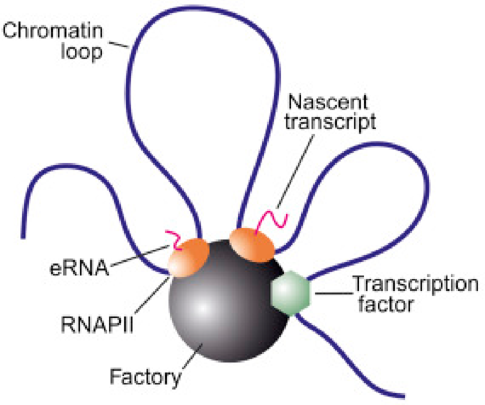


... the RNA polymerase transiently immobilized on the surface of a supramolecular protein complex as it reels in its template to copy it and produce a transcript ...

We define a transcription factory as multiprotein, supramolecular, nuclear body containing at least two RNA polymerases engaged on two different transcription units at any given time ...

Sofiadis and Papantonis (2018)

Transcription factories



- Co-regulated genes are **co-transcribed** in 3D nuclear space
- Changes in gene expression profile (e.g. upon differentiation) are mediated by changes in transcription factories. (For example, from enhancer to silencer)
- Chromosomal rearrangements (hallmark of cancer etiology) are governed by spatial proximity

Sofiadis and Papantonis (2018)

The joint distribution

Existence

Positivity condition(Besag (1974)) and Markov random field condition (Kaiser and Cressie (2000))

Conditional Autoregressive Models (CAR)

$$y_i|\mathbf{y}_{-i} \sim N(\mu_i, \tau_i^2) \tag{2}$$

or

$$f_i(y(s_i)|\{y(s_j) : j \neq i\}) = \frac{1}{\sqrt{2\pi\tau_i^2}} exp[-\frac{1}{2\tau_i^2}\{y(s_i) - \mu\}^2],$$

We can further model the conditional mean with

$$\mu(\{y(s_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} c_{i,j}\{y(s_j) - \alpha_j\}. \tag{3}$$

such that

$$c_{i,j}\tau_j^2 = c_{j,i}\tau_i^2, c_{i,i} = 0; \text{for } i, j = 1, \ldots, n$$

Conditional Autoregressive Model

CAR

$$y_i|\boldsymbol{y}_{-i} \sim N(\mu_i, \tau_i^2)$$

where

$$\mu(\{y(s_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} c_{i,j}\{y(s_j) - \alpha_j\}.$$

such that

$$c_{i,j}\tau_j^2 = c_{j,i}\tau_i^2, c_{i,i} = 0; \text{for } i, j = 1, \ldots, n$$

Our model

$$w_i|\boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2)$$

where

$$\mu_i = \alpha + \sum_{j \in N_i} \frac{\eta}{|N_i| + |N_j|}(w_j - \alpha).$$

obviously,

$$\eta\tau^2 = \eta\tau^2$$

Conditional Autoregressive Model

If we let C denote the $n \times n$ matrix with elements $c_{i,j}$, and M the $n \times n$ matrix with diagonal elements τ_i^2, then the joint distribution of $Y(s_1), \ldots, Y(s_n)$ is

$$\mathbf{Y} \sim N(\boldsymbol{\alpha}, (I_n - C)^{-1} M),$$

if $(I_n - C)$ is invertible and $(I_n - C)^{-1}M$ is positive definite.
Denote the covariance matrix as $\mathbf{\Sigma}$, then the precision matrix is

$$\mathbf{Q} = \mathbf{\Sigma}^{-1} = M^{-1}(I_n - C)$$

For $i \neq j$, $Y(s_i)$ and $Y(s_j)$ are conditionally independent given the rest if and only if $q_{ij} = 0$. The neighborhood enters the joint through the precision matrix Q.

Model Extension

- Extension:

  $$N_i = N_{i1} \cup N_{i2} \cup \cdots \cup N_{iL}.$$

  .

  Each N_{il} is a different neighborhood type, $l = 1, 2, \ldots, L$.

  $$\begin{aligned}\mu_i &= \alpha + \sum_{j \in N_{i1}} \frac{\eta_1}{|N_i| + |N_j|}(w_j - \alpha) + \cdots + \sum_{j \in N_{iL}} \frac{\eta_L}{|N_i| + |N_j|}(w_j - \alpha) \\ &= \alpha + \sum_{l=1}^{L} \sum_{j \in N_{il}} \frac{\eta_l}{|N_i| + |N_j|}(w_j - \alpha).\end{aligned}$$

Model inference

Let $g(w_i|\boldsymbol{w}(N_i), \alpha, \boldsymbol{\eta}, \tau^2)$ be the conditional distribution for w_i, and $g(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2)$ be the joint distribution of $\boldsymbol{w}$.

- The posterior distribution for w_i is

$$\begin{aligned} & p(w_i|\alpha, \boldsymbol{\eta}, \tau^2, \boldsymbol{w}, \boldsymbol{y}) \\ \propto & \prod_{k=1}^{M} f(y_{ik}|w_i) g(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2) \\ \propto & \prod_{k=1}^{M} f(y_{ik}|w_i) g(w_i|\boldsymbol{w}(N_i), \alpha, \boldsymbol{\eta}, \tau^2) \end{aligned}$$

- The posterior distribution for α is

$$p(\alpha|\boldsymbol{\eta}, \tau^2, \boldsymbol{y}, \boldsymbol{w}) \propto \pi(\alpha) g(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2) \tag{5}$$

Model inference

- The posterior distribution for α is

$$p(\alpha|\boldsymbol{\eta}, \tau^2, \boldsymbol{y}, \boldsymbol{w}) \propto \pi(\alpha) g(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2) \tag{5}$$

-

$$\begin{aligned} & p(\alpha|\eta, \tau^2, \boldsymbol{y}, \boldsymbol{w}) \\ \propto & \pi(\alpha) \frac{exp(Q(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2))}{\int exp(Q(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2)) d\boldsymbol{w}} \\ = & \pi(\alpha) C(\alpha) exp(Q(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2)) \end{aligned}$$

where

$$C(\alpha) = 1/\int exp(Q(\boldsymbol{w}|\alpha, \boldsymbol{\eta}, \tau^2)) d\boldsymbol{w}$$

MCMC

- Double Metropolis-Hasting algorithm (Liang (2010))
    - Posterior with intractable normalizing constant
    - Simulation of auxiliary variable
- Metropolis-within-Gibbs

Computation Acceleration

Each MC iteration, computational time complexity is $O(n^3)$.

- Written in C
- Parallelization (OpenMP)
- BLAS and Lapack routines