Identifying cluster structures and relevant variables in high-dimensional datasets

Mahlet G. Tadesse

Department of Mathematics & Statistics
Georgetown University

SADA 2013, Cotonou, Bénin

Outline

- Review of Bayesian stochastic search variable selection
  - Bayesian variable selection in linear model
  - Bayesian variable selection in classification
- Bayesian variable selection in clustering
  - Finite mixtures with unknown number of components
  - Infinite mixtures with DPM
- Discovering cluster structures & relationships between high-dimensional data sets
  - Stochastic Partitioning Method
  - MCMC implementation

Outline

Outline

SSVS for linear model

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- Introduce latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$, where

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ is included in the model} \\ \gamma_j = 0 & \text{otherwise} \end{cases}$$

- γ_j used to induce mixture prior on the regression coefficients

$$\beta_j \sim (1 - \gamma_j)\mathcal{I}_0 + \gamma_j \mathcal{N}(0, \tau_j^2)$$

- Prior of $\boldsymbol{\gamma}$ updated via Gibbs sampling or Metropolis algorithm.

- MCMC procedure results in a list of visited models $\boldsymbol{\gamma}^{(t)}$ and their relative posterior probabilities $p(\boldsymbol{\gamma}^{(t)}|\boldsymbol{X}, Y)$, $t = 0, \ldots, T$.
- Inference for variable selection can be based on:
  - vector $\boldsymbol{\gamma}$ with highest joint posterior probability, $p(\boldsymbol{\gamma}|\boldsymbol{X}, Y)$
  - γ_j's with largest marginal posterior probabilities, $p(\gamma_j = 1|\boldsymbol{X}, Y)$.
- Outcome for future observations can be predicted via Bayesian model averaging

$$\widehat{Y}_f = \sum_{\gamma} \left(\boldsymbol{X}_{f(\gamma)} \widehat{\boldsymbol{\beta}}_{(\gamma)} \right) \cdot p(\boldsymbol{\gamma}|\boldsymbol{X}, Y).$$

George and McCulloch, *JASA*, 1993; *Statistica Sinica*, 1997.

Outline

Variable selection in multinomial probit model

- Observed data consist of $(\boldsymbol{Z}_{n\times 1}, \boldsymbol{X}_{n\times p})$ with categorical response Z_i taking values $0, \ldots, K-1$.
- Use a multinomial probit model to link $P(Z_i = k)$ to $\boldsymbol{X}_i\beta$.
- Introduce latent matrix $\boldsymbol{Y}_{n\times(K-1)}$, where $\boldsymbol{Y}_i = (y_{i,1}, \ldots, y_{i,K-1})$

  (Albert and Chib, *JASA*, 1993)

$$Z_i = \begin{cases} 0 & \text{if } y_{i,k} < 0 \ \ \forall k \\ k & \text{if } y_{i,k} = \max_{1\le j\le K-1} y_{i,j} \end{cases}$$

$$\boldsymbol{Y}_i = \boldsymbol{\alpha}' + \boldsymbol{X}_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \quad i = 1 \ldots, n$$

- Introduce binary latent vector γ to induce mixture prior on β

$$\beta_{(\gamma)} \sim \mathcal{N}\left(\beta_{0(\gamma)}, \boldsymbol{H}_{(\gamma)} \otimes \Sigma\right)$$

- MCMC procedure iterates between the following steps:
  (1) Update $\boldsymbol{Y}$ from $p(\boldsymbol{Y}|\gamma, \boldsymbol{X}, \boldsymbol{Z})$, a truncated multivariate-t distribution.
  (2) Update γ using a Metropolis search.

- Discriminating variables are selected based on $p(\gamma|\boldsymbol{X}, \widehat{\boldsymbol{Y}}, \boldsymbol{Z})$.

Sha, Vannucci, Tadesse, *et al.*, *Biometrics*, 2004.

Outline

Motivation: Discover disease subtypes & detect biomarkers

Interest in identifying homogeneous subgroups of samples and selecting discriminating variables.

- For various malignancies, existing disease classes are too broad.
- Biomarker profiles may better capture disease heterogeneities.

Challenges:

- unknown number of classes
- class membership of samples not observed
- discriminating markers need to be identified.

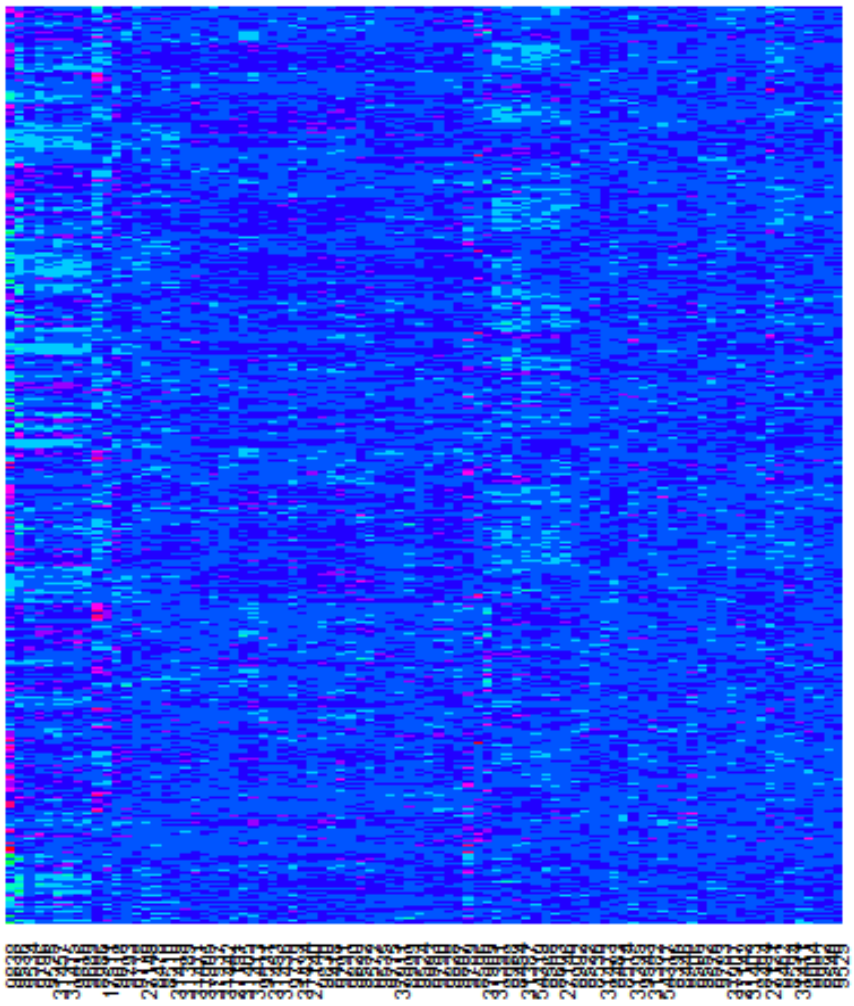
Example: Idiopathic dilated cardiomyopathy (IDC)

- IDC results in dilated and weakened heart that does not pump blood efficiently.
- The causes of IDC are unknown (viral infection, inherited or spontaneous mutations, toxic exposures) and response to treatment varies across patients.

Can gene expression profiles capture disease heterogeneities?

- Myocardial cells obtained from 86 patients with IDC at time of heart transplant.
- RNA samples isolated and hybridized to Affymetrix HU133 arrays.

Heatmap using all probe sets considered for analysis



Bayesian variable selection in model-based clustering

Proposed methods provide a unified approach for discovering cluster structure among samples and selecting relevant variables.

- Use model-based clustering with an unknown number of components to uncover cluster structure:
  - finite mixture models with reversible jump MCMC techniques.
  - infinite mixture models with Dirichlet process mixture priors.
- Use stochastic search MCMC techniques to explore space of variable subsets and identify discriminating genes.

Outline

Model-based clustering

- Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be independent p-dimensional observations arising from K populations

$$f(\boldsymbol{x}_i|\boldsymbol{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k).$$

- We consider $f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ multivariate normal with $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \Sigma_k)$.
- Introduce latent vector $\boldsymbol{y} = (y_1, \ldots, y_n)'$, where $y_i = k$ if the i^{th} observation comes from cluster k

$$p(y_i = k) = w_k.$$

Binder, 1978; McLachlan and Basford, 1988; Fraley and Raftery, 1998.

Stochastic search variable selection

- We need to select the variables that provide information about the cluster structure.
- Introduce latent p-vector $\boldsymbol{\gamma}$ with binary entries

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ defines a mixture distribution} \\ \gamma_j = 0 & \text{otherwise.} \end{cases}$$

- γ is used to explore the space of variable subsets and search for models with highest posterior probabilities.
- The use of $\boldsymbol{\gamma}$ is different from the regression setting, where it is used to induce mixture priors on regression coefficients.

- The likelihood function is given by

$$L(K, \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}, \boldsymbol{\Omega} | \mathbf{X}, \boldsymbol{y}) = \prod_{k=1}^{K} (2\pi)^{\frac{-pn_k}{2}} |\Sigma_{(\gamma)k}|^{\frac{-n_k}{2}} w_k^{n_k}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{x_i \in C_k} (\boldsymbol{x}_{(\gamma)i} - \boldsymbol{\mu}_{(\gamma)k})^T \Sigma_{(\gamma)k}^{-1} (\boldsymbol{x}_{(\gamma)i} - \boldsymbol{\mu}_{(\gamma)k}) \right\}$$
$$\times \phi(\boldsymbol{X}_{(\gamma^c)} | \boldsymbol{\eta}_{(\gamma^c)}, \boldsymbol{\Omega}_{(\gamma^c)}),$$

where $C_k = \{x_i | y_i = k\}$ with cardinality n_k, $\phi(.)$ is multivariate normal density.

- Specify conjugate priors and integrate out mean and covariance parameters.

Finite mixtures with reversible jump MCMC

Specify prior for number of components, K

$$K \sim \text{trunc-Poisson}(\lambda) \text{ or Uniform on } 2, \ldots, K_{\max}$$

MCMC procedure iterates:

(1) Update $\boldsymbol{\gamma}$ from its full conditional.

(2) Update $\boldsymbol{w}$ from its full conditional.

(3) Update $\boldsymbol{y}$ from its full conditional.

(4) Split one cluster into two, or merge two into one.

(5) Birth or death of an empty component.

Steps (4) and (5) use reversible jump MCMC (Green 1995; Richardson and Green 1997).

Tadesse, Sha and Vannucci, *JASA*, 2005.

Outline

Dirichlet process mixtures

- Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be independent p-dimensional observations arising from a mixture of distributions $F(\boldsymbol{\theta}_i)$.
- $\boldsymbol{\theta}_i$ are assumed to be independent draws from a distribution G, which follows a Dirichlet process prior (Antoniak 1974).

$$\begin{aligned} \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim DP(G_0, \alpha), \end{aligned}$$

Dirichlet process mixtures

Equivalent models can be obtained by taking the limit as $K \to \infty$ of finite mixture models with K components.

$$\begin{aligned} \boldsymbol{x}_i | y_i, \boldsymbol{\psi} &\sim F(\boldsymbol{\psi}_{y_i}) \\ y_i | \boldsymbol{w} &\sim \text{Discrete}(w_1, \ldots, w_K) \\ \boldsymbol{\psi}_y &\sim G_0 \\ \boldsymbol{w} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \end{aligned}$$

where the latent variable y_i indicates the cluster allocation of sample i and $\boldsymbol{\psi}_{y_i}$ corresponds to the identical $\boldsymbol{\theta}_i$'s.

Dirichlet process mixtures

- Integrating over $\boldsymbol{w}$ and taking $K \to \infty$ (Neal 2000)

$$p(y_i = y_l \text{ for some } l \neq i | \boldsymbol{y}_{-i}) = \frac{n_{-i,y_l}}{n - 1 + \alpha}$$

$$p(y_i \neq y_l \text{ for all } l \neq i | \boldsymbol{y}_{-i}) = \frac{\alpha}{n - 1 + \alpha}$$

- Specify conjugate priors and integrate out mean and covariance parameters.

Infinite mixtures with DPM

- Avoids dimensions jumping scheme

$$p(y_i = y_l \text{ for some } l \neq i | y_{-i}, \boldsymbol{x}_i) = b \frac{n_{-i,y_l}}{n-1+\alpha} \int F(\boldsymbol{x}_i; \boldsymbol{\psi}) dG_{-i,y_l}(\boldsymbol{\psi})$$

$$p(y_i \neq y_l \text{ for all } l \neq i | y_{-i}, \boldsymbol{x}_i) = b \frac{\alpha}{n-1+\alpha} \int F(\boldsymbol{x}_i; \boldsymbol{\psi}) dG_0(\boldsymbol{\psi}),$$

- MCMC procedure iterates the following steps:

  (1) Update $\boldsymbol{\gamma}$ using a Metropolis algorithm

  (2) Update $\boldsymbol{y}$ using split-merge MCMC of Jain & Neal (*JCGS* 2004).

Kim, Tadesse and Vannucci, *Biometrika*, 2006.

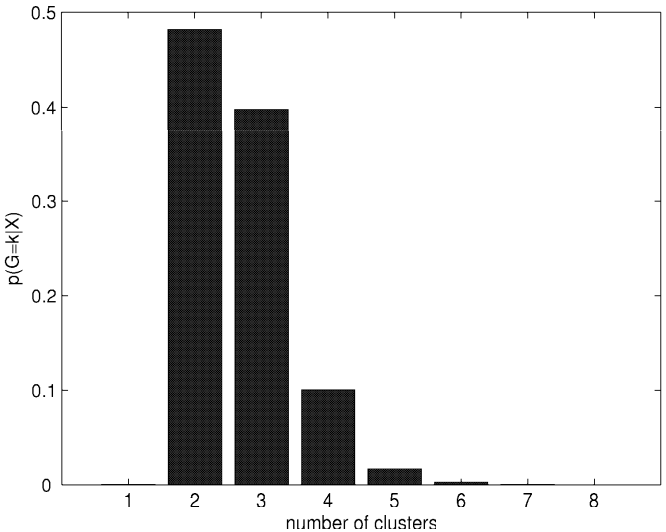
Posterior inference

For cluster structure:

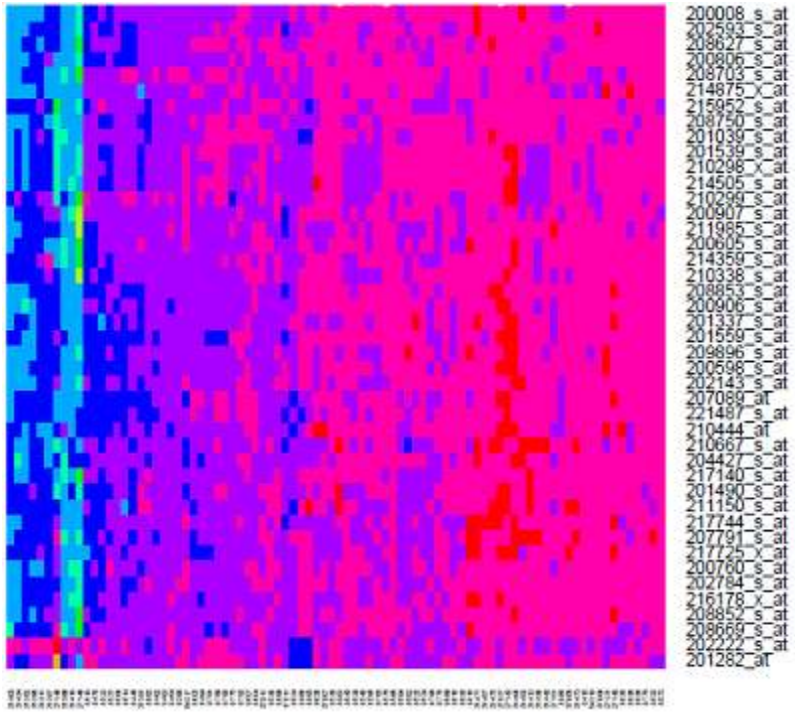
- Inference conditional on $\widehat{K}$
  - Estimate G by number of clusters most visited by MCMC sampler
  - Relabel by minimizing posterior expectation of $\mathcal{L}_0(\boldsymbol{y};\boldsymbol{\phi}) = -\sum_{i=1}^{n} \log\{p_{iy_i}(\boldsymbol{\phi})\}$ (Stephens, *JRSS-B*, 2000).
  - Estimate sample allocation by the mode of the marginal posterior distribution, $\widehat{y}_i = \text{argmax}_{1\leq k\leq K}\left\{p(y_i = k|\boldsymbol{X},\widehat{K})\right\}$.
- Use posterior pairwise probabilities $p(y_i = y_j|\boldsymbol{X})$.

For relevant variables:

- Select variables with largest marginal posterior probabilities, $p(\gamma_j = 1|\boldsymbol{X})$.



(a) Visited number of clusters



(b) Heatmap of selected probe sets

Outline

Motivation: Integrate genomic data sets

- Relating data sets from various genome-wide technologies may give insights into the complex DNA-RNA-protein relationships.
- There are several ongoing efforts in this area:
  - eQTL studies – association between gene expression and SNP array data (Morley *et al.* 2004; Cheung *et al.*, *Nature* 2005).
  - Association between gene expression profiles and aCGH data (Bussey *et al.* 2006; Stranger *et al.*, *Science* 2007)

Relating genomic datasets

- The goal in these studies is to identify DNA sequence variations that explain changes in mRNA transcript levels.
- The standard methods of analysis consist of fitting univariate linear regression models for each outcome on each regressor.
  - Morley *et al.* assessed each of 3 554 expression levels on each of 2 455 SNP markers.
  - Stranger *et al.* examined each of 14 925 expression levels with each of 24 963 autosomal CGH clones one at a time.

Existing variable selection methods

Several methods have been proposed to relate high-dimensional covariate data to univariate outcomes

e.g., identify gene expression levels associated with disease status or time to event in DNA microarray studies.

$$Y_{N\times 1} \qquad X_{n\times p},\ p \gg N$$

$$\begin{array}{|c|} \hline y_1 \\ \vdots \\ y_N \\ \hline \end{array} \qquad \begin{array}{|cccccc|} \hline x_{11} & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ x_{N1} & \cdots & \cdots & \cdots & x_{Np} \\ \hline \end{array}$$

Goal:

Relate response and covariate data when both are high-dimensional

$Y_{N\times q},\ \ q \gg N$

$$\begin{bmatrix} y_{11} & \cdots & \cdots & \cdots & y_{1q} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ y_{N1} & \cdots & \cdots & \cdots & y_{Nq} \end{bmatrix}$$

$X_{N\times p},\ \ p \gg N$

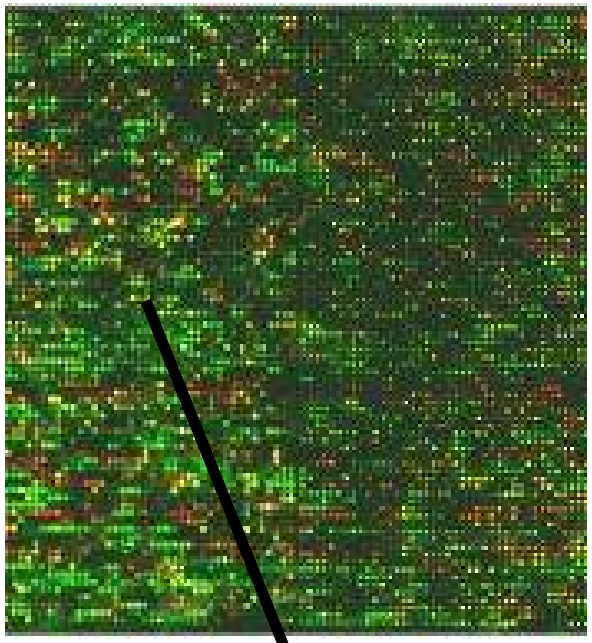
$$\begin{bmatrix} x_{11} & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ x_{N1} & \cdots & \cdots & \cdots & x_{Np} \end{bmatrix}$$

Identify sets of correlated outcomes modulated by sets of covariates

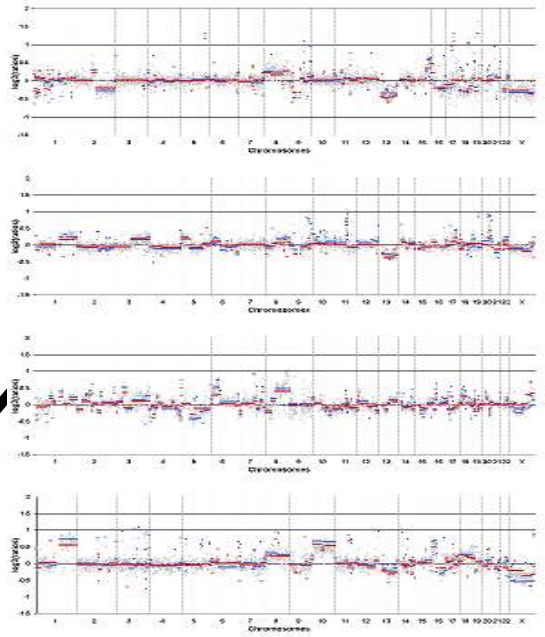
e.g., $([X_{34}, X_{590}, X_{1015}], [Y_1, Y_{26}, Y_{927}])$

$([X_5, X_{590}, X_{369}, X_{872}], [Y_{47}, Y_{168}, Y_{2156}])$

Gene expression



CGH profile



$$Y_{ji}|S_k \sim \mathcal{N}(\alpha_j + \sum_{r=1}^{m_k} \beta_{s_r} \cdot X_{s_r i}\ , \sigma_k^2)$$

$[Y_5, Y_{11}, Y_{91}]$ $[X_{22}, X_{52}, X_{82}, X_{83}, X_{106}]$

$[Y_2, Y_{23}, Y_{37}, Y_{43}, Y_{57}, Y_{79}, Y_{80}, Y_{81}, Y_{88}]$ $[X_1, X_{23}, X_{77}, X_{102}, X_{135}, X_{145}, X_{175}, X_{198}]$

$[Y_{27}, Y_{93}, Y_{96}]$ $[X_{12}, X_{23}, X_{87}, X_{104}, X_{135}, X_{145}, X_{149}, X_{151}, X_{176}, X_{177}]$

$[Y_{21}, Y_{40}, Y_{50}, Y_{53}, Y_{55}, Y_{63}, Y_{76}, Y_{78}, Y_{95}, Y_{99}, Y_{100}]$ $[\emptyset]$

Outline

Proposed stochastic partitioning method

- The data consist of N independent samples with $\mathcal{X} = (X_1, \ldots, X_p)$ covariates and $\mathcal{Y} = (Y_1, \ldots, Y_q)$ outcomes.
- We propose pairwise partitioning the data into subsets of $\mathcal{X}$ and $\mathcal{Y}$ to identify sets of associated markers.
- An element of the pairwise partition is a pair

$$\mathcal{S} = (X_I, Y_J), \quad I \subset \{1, \ldots p\}, \; J \subset \{1, \ldots q\},$$

such that the X_I will jointly explain changes in and have the same effect on Y_J.

- Each Y_j is allowed to belong to one component only, whereas a variable X_r may belong to many components or to none.

- A decomposition of the variables into K components (i.e., a configuration) is given by

$$\mathcal{S}_1 \oplus \ldots \oplus \mathcal{S}_K = (X_{I_1}, Y_{J_1}) \oplus \ldots \oplus (X_{I_K}, Y_{J_K}).$$

- For simpler notation, a component $\mathcal{S}_k$ is labeled by its cardinalities

$$(|I_1|, |J_1|) \oplus \ldots \oplus (|I_K|, |J_K|),$$

$0 \leq |I_k| \leq p, \quad 1 \leq |J_k| \leq q, \quad \sum_{k=1}^{K} |J_k| = q.$

- Components of type $(m, 0)$ are equivalent to components of type $(1, 0)$ since X is viewed as a fixed covariate matrix.

Example: $p = 12,\ q = 10$

$$([X_1], [Y_1, Y_2]) \oplus ([X_1, X_{12}], [Y_5, Y_8, Y_9, Y_{10}]) \oplus ([\,], [Y_3, Y_6, Y_7]) \oplus ([X_9, X_{10}], Y_4)$$

Compact notation

$$(1, 2) \oplus (2, 4) \oplus (0, 3) \oplus (2, 1)$$

Conditional on a partition $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$, outcomes in the same component are assumed to have similar profiles

$$Y_{ji}|\mathcal{S}_k \overset{iid}{\sim} \mathcal{N}(\alpha_j + \mu_k, \sigma_k^2), \qquad j = t_1, \ldots, t_{n_k}, \;\; i = 1, \ldots, N,$$

where $\mu_k = g_k(X_{s_1}, \ldots, X_{s_{m_k}})$.

Priors and marginal likelihood

- We assign a prior to each configuration

$$p((m_1, n_1) \oplus \ldots \oplus (m_K, n_K)) \propto \prod_{k=1}^{K} \rho^{m_k \cdot n_k}, \qquad 0 < \rho \leq 1.$$

- We take conjugate priors for the regression parameters and integrate them out

$$f(m_k, n_k) = \int \phi(m_k, n_k)\, dp(\theta_k|\sigma_k)\, dp(\sigma_k)$$

- The marginalized likelihood of a configuration reduces to

$$f\left((m_1, n_1) \oplus \ldots \oplus (m_K, n_K)\right) = \prod_{k=1}^{K} f(m_k, n_k)$$

Outline

MCMC implementation

- The number of possible configurations of type (m, n), with $n > 0$, is

$$\sum_{k=1}^{q} S_2(q,k)\, 2^{p.k},$$

  where S_2 are the Stirling numbers of the second kind.

- We construct a Markov chain, where transitions between configurations are implemented by splitting or merging components.

MCMC implementation

- To ensure better mixing among both regressors and response variables, we implement the Markov chain as a two-step process:
  - **Step 1**: propose moves that allow creation or deletion of $(1, 0)$ components.
  - **Step 2**: propose moves that allow split or merge of (m, n) components $(n > 1)$.

- In addition, we use parallel tempering (Geyer, 1991) to prevent the sampler from being trapped in local modes.

Monni and Tadesse, *Bayesian Analysis*, 2009.

Parallel tempering implementation

- Define R distributions $\xi_i(x) = \xi(x)^{1/T_i}$, $1 = \frac{1}{T_0} > \ldots > \frac{1}{T_{R-1}} > 0$.
- $\xi_0(C) = \xi(C) = f(C) \cdot p(C)$ is the posterior distribution from which we want to sample.

The tempering algorithm iterates between the following steps:

(i) **parallel scan:** for each $\xi_i(.)$, perform a fixed number of updates.

(ii) **state exchange:** swap neighboring chains and accept the exchange between configurations at T_i and T_{i+1} with probability

$$P(C(T_{i+1}) \leftrightarrow C(T_i)) = \min\left\{1, \left(\frac{f(C(T_{i+1}))}{f(C(T_i))} \cdot \frac{p(C(T_{i+1}))}{p(C(T_i))}\right)^{1/T_i - 1/T_{i+1}}\right\}.$$

Posterior inference

- Pairwise posterior probabilities take into account the contributions of different configurations
  - $p \times q$ matrix of posterior probabilities for association between X_i and Y_j
  - $q \times q$ matrix of posterior probabilities for allocation of (Y_i, Y_j) to same components.
- Consider most likely models by locating different modes of the posterior probability.

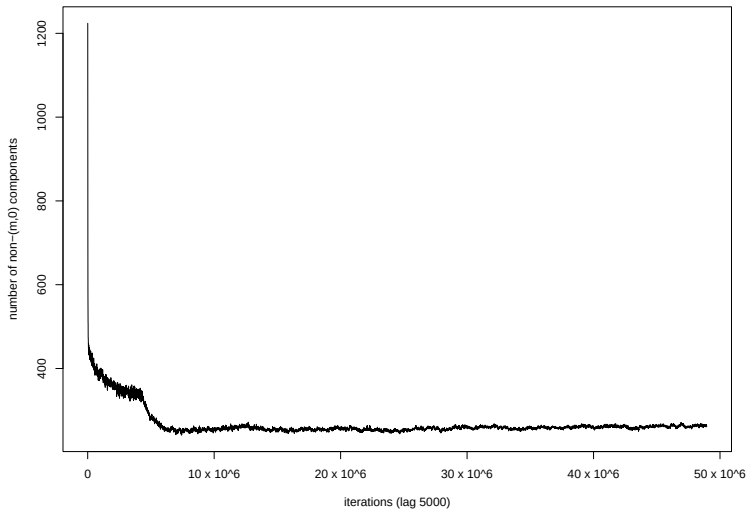
Application to eQTL analysis

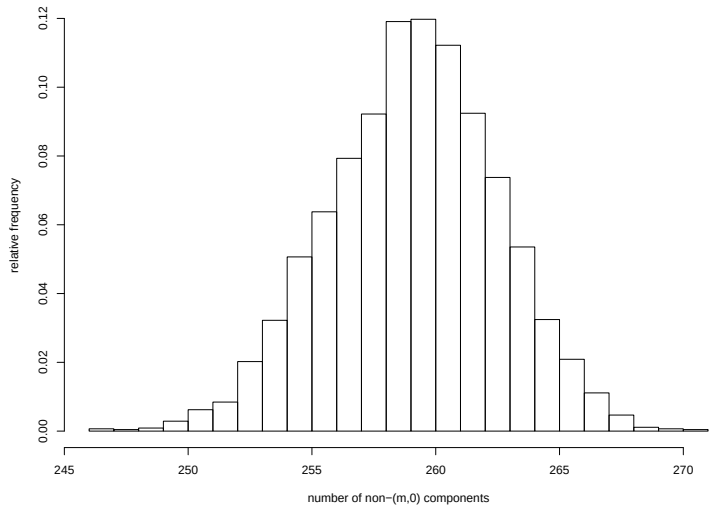
- Expression quantitative trait loci (eQTL) studies assess the genetic basis of variations in mRNA transcript abundance.
- We use the data from Morley *et al.* (*Nature*, 2004)
  - 56 unrelated individuals from 14 CEPH families examined.
  - RNA samples from each individual hybridized to Affymetrix arrays; 3554 probe sets considered for analysis.
  - SNP genotypes for each individual obtained on 2455 markers from the SNP Consortium database.

Posterior inference

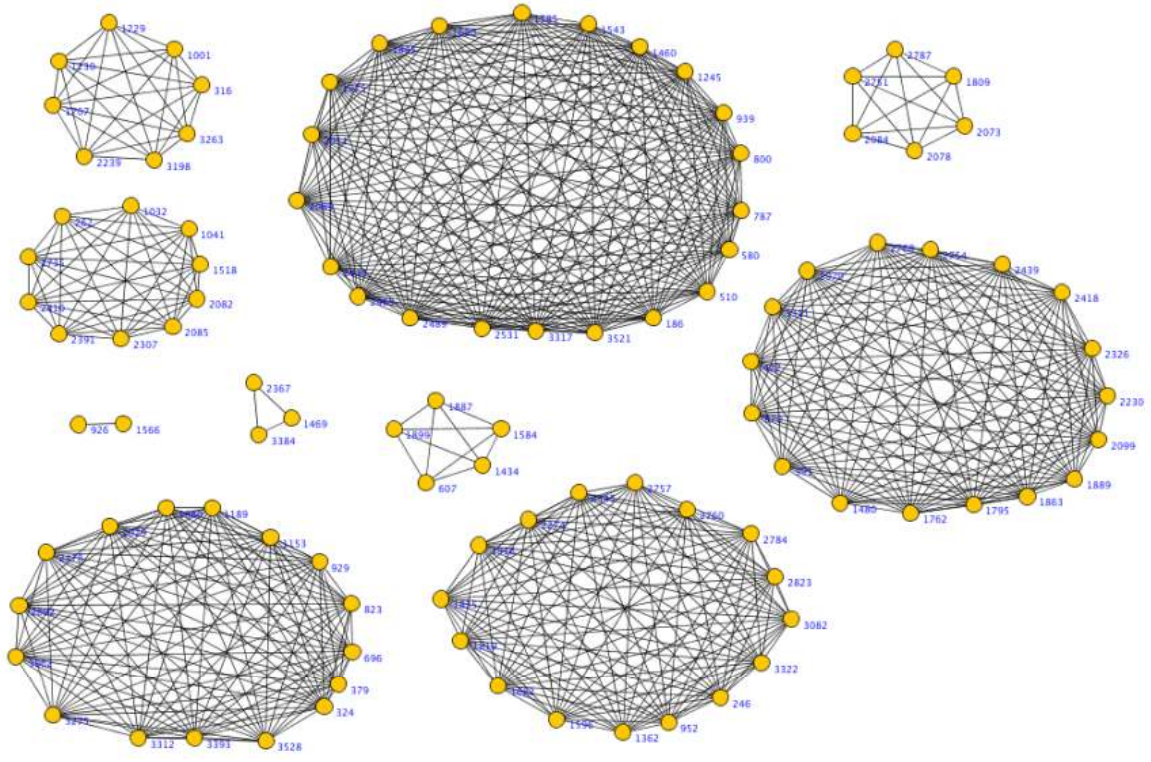
Ran MCMC chains for 50 million iterations and focused on last 15 million iterations, sub-sampling configurations every 20 000 scans.

- Use 3554×3554 matrix of posterior pairwise probabilities that two probe sets be allocated to the same component to identify correlated outcomes.
- Use the 2455×3554 matrix of marginal posterior probabilities that each SNP be associated with each probe set
  - examine rows to focus on specific markers and identify expression phenotypes to which they are strongly associated;
  - examine columns to focus on specific gene expression phenotype and locate its related markers.

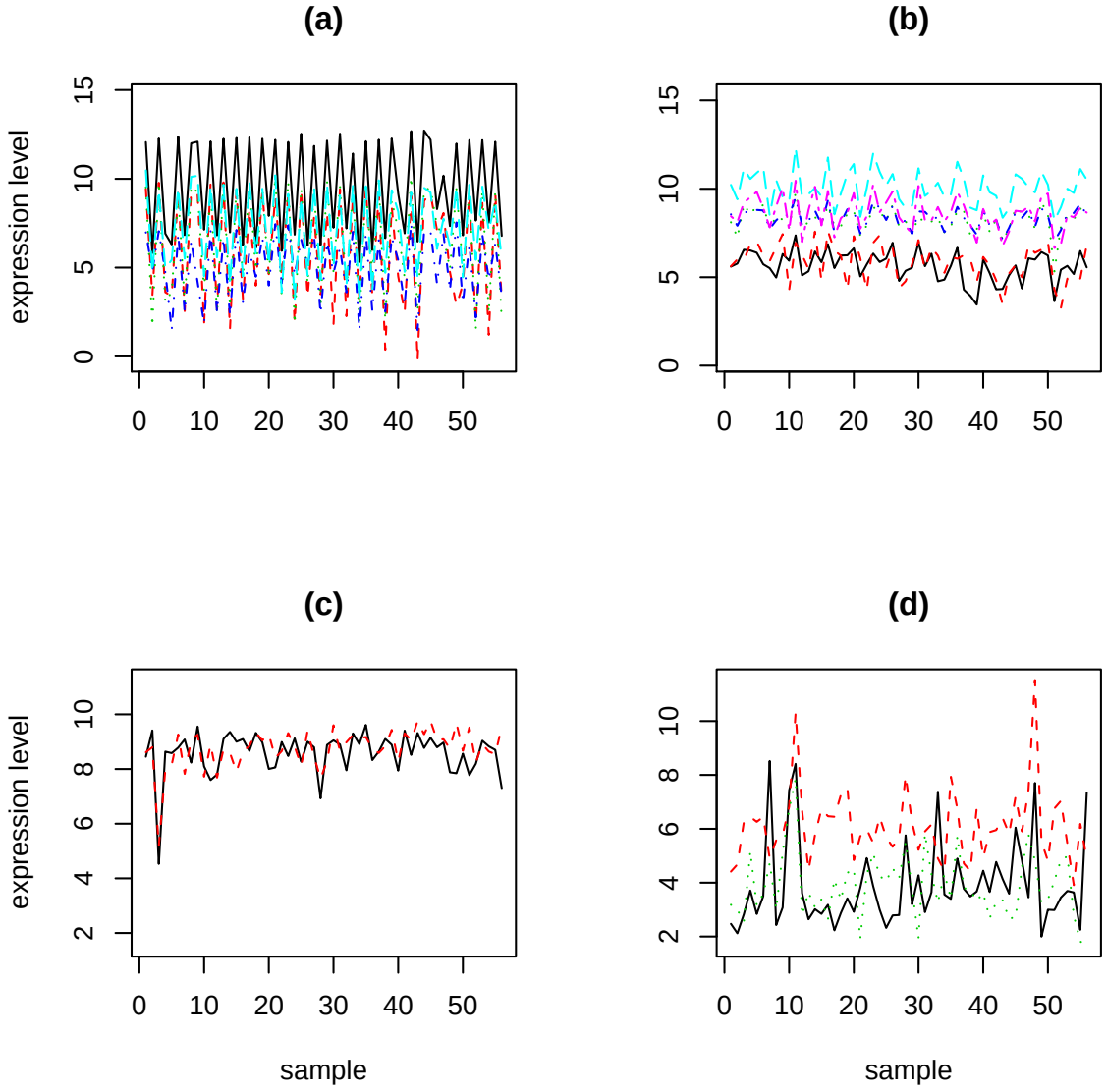




Number of components of type (m, n)



Network representation of some gene expressions with posterior pairwise probability for occurring in same components ≥ 0.7.


(a)
(b)
(c)
(d)
expression level
expression level
sample
sample
0
5
10
15
0
10
20
30
40
50
2
4
6
8
10

SNP marker		Gene expression	
RefSnp	Location (Mbp)	Name	Location
rs1859674	Chr X (116.29)	HDHD1A	Xp22.32
		UTX	Xp11.2
		U2AF1L2	Xp22.1
		XIST	Xq13.2
rs533569	Chr 11 (93.70)	HIST1H3H	6p21.3
		HIST1H2BF	6p21.3
		HIST1H2BE	6p21.3
		H2BFS	21q22.3
		HIST1H2BC	6p21.3
		HIST1H2AC	6p21.3
rs127503	Chr 6 (108.59)	SLC4A2	7q35
		CDK10	16q24
		LCAT	16q22.1
		CYP4F12	19p13.1

Example of markers and associated gene expressions

Summary

- Proposed methods provide exploratory tools to investigate key features and associations in high-dimensional data sets
  - mixture models with unknown number of components used to uncover cluster structures
  - stochastic search MCMC techniques used to explore space and identify variables related to components.

- Additional information may be incorporated to elicit priors and design better proposal distributions.