Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

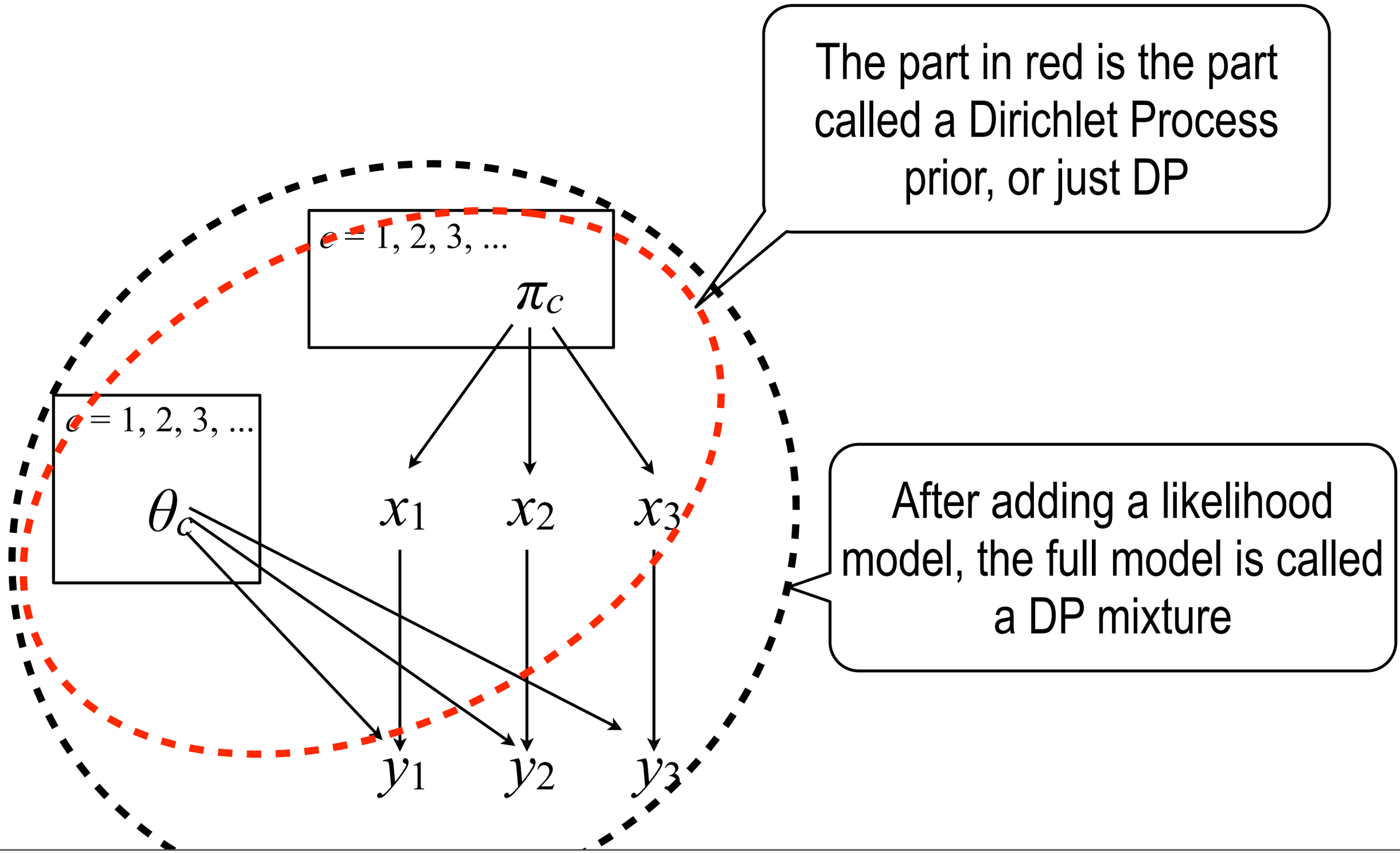
Lecture 8, Wednesday March 23

Program for today

- Applications
  - GLMs: Regression and classification
  - NLP: language modelling, segmentation, alignment
- Extensions
  - Hierarchies and sequences
  - Pitman-Yor process

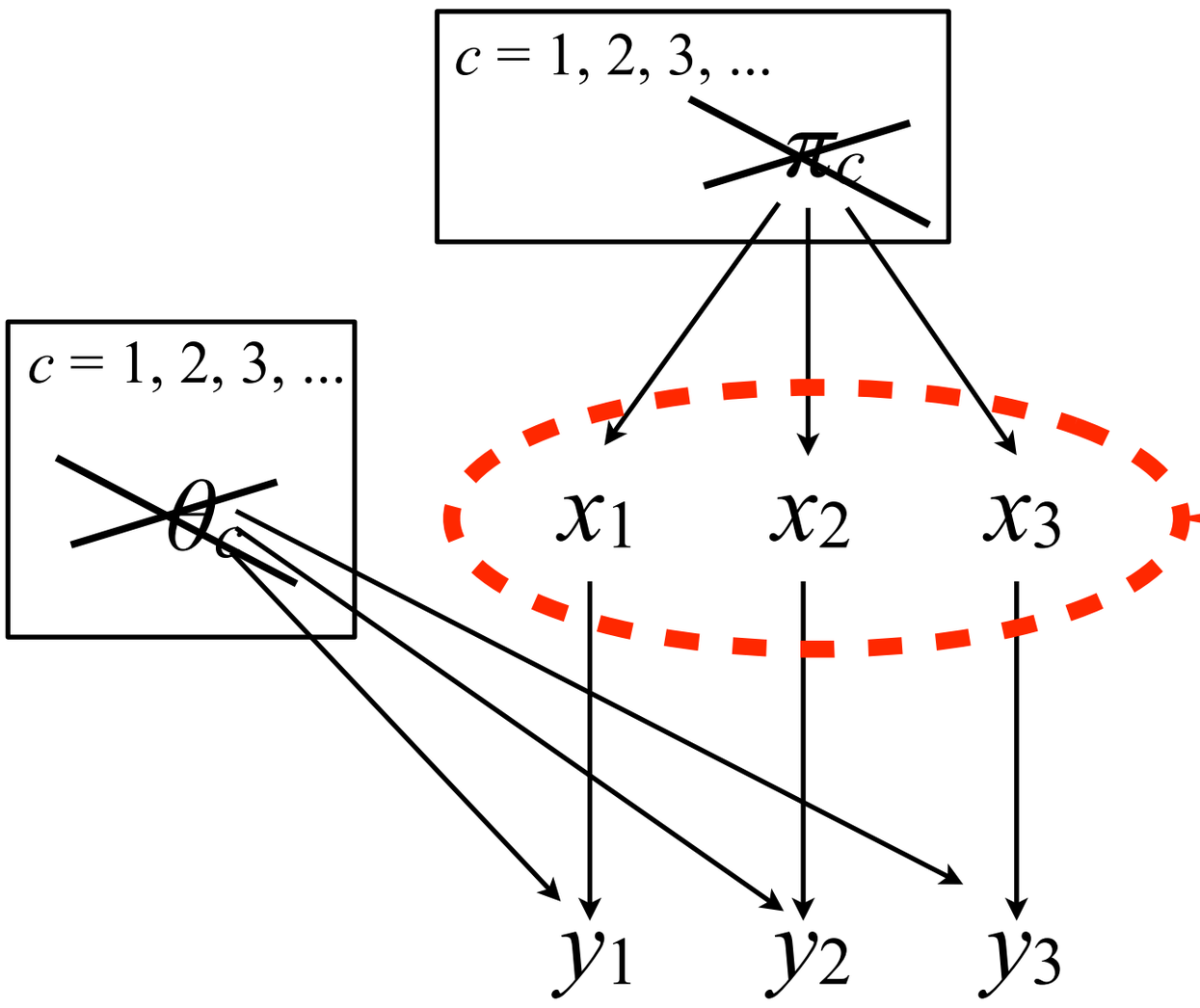
Review

Terminology



Terminology

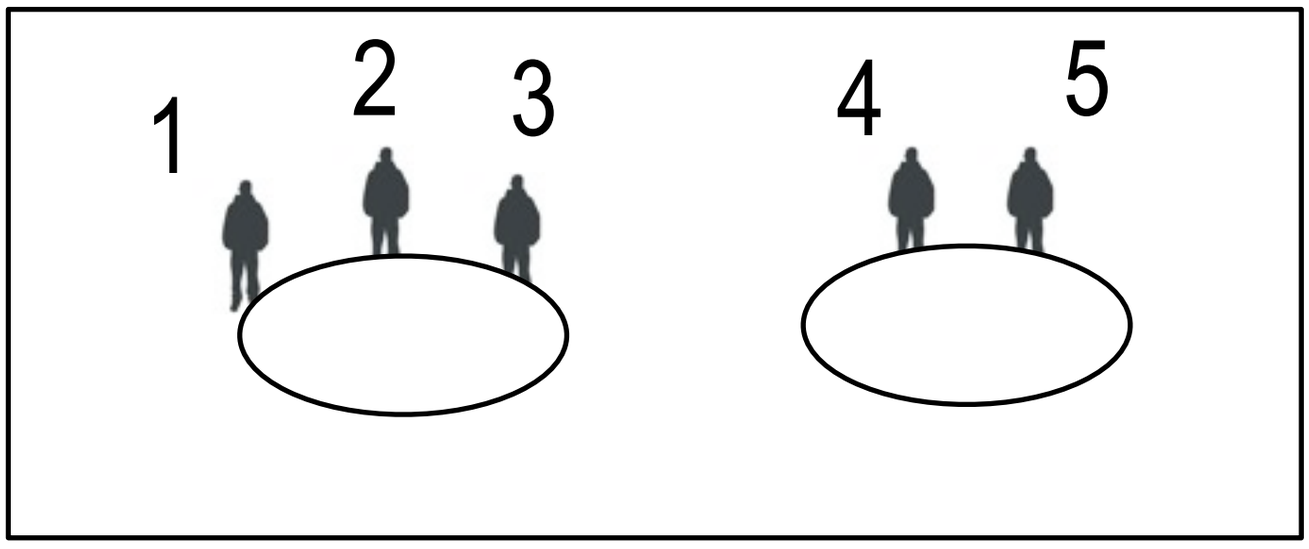
CRP: Chinese Restaurant Process



We denote the marginal distribution over the cluster indicators x's by $\mathrm{CRP}(\alpha_0)$

CRP: Quiz

What is the marginal CRP(1) of this table assignment, $\alpha_0 = 1$



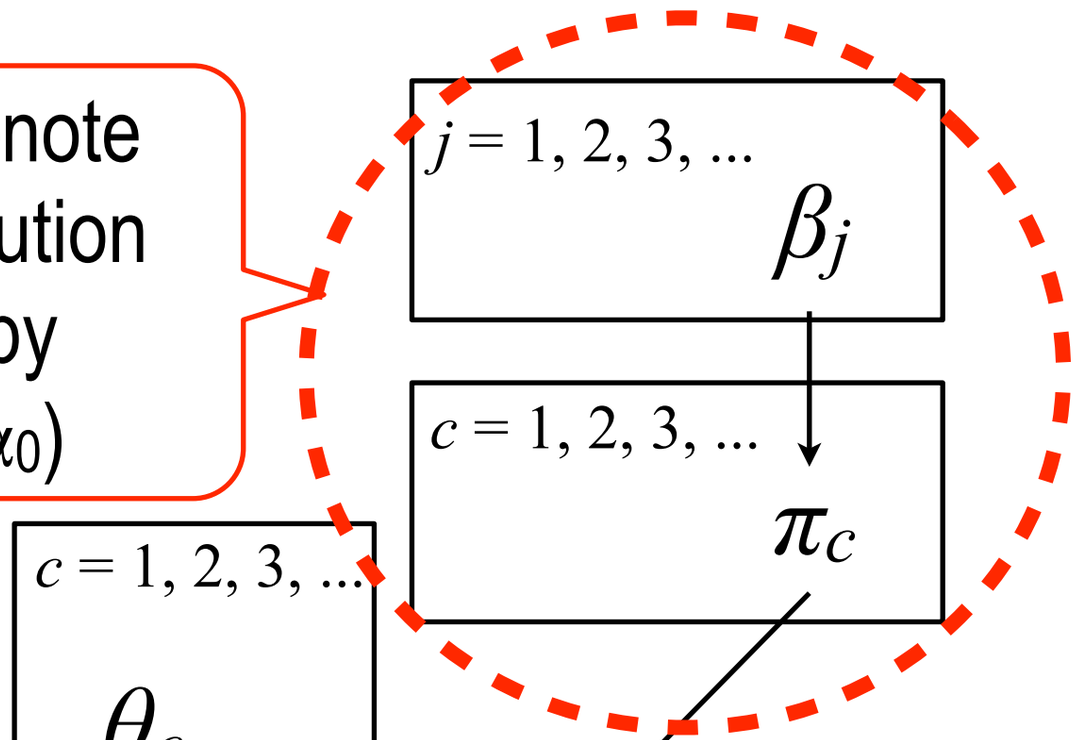
(Solution shown in the board)

Terminology

GEM: Griffiths-Engen-McCloskey

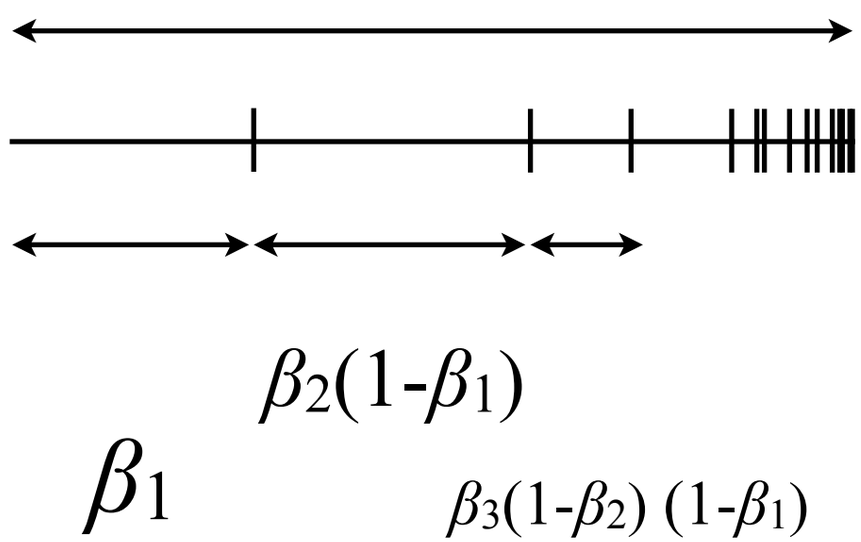
$$\beta_j \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_0)$$

We will denote this distribution over π by GEM(α_0)



$$\theta_c \overset{\text{iid}}{\sim} G_0$$

$$G' = \sum_{c=1}^{\infty} \pi_c \delta_{\{\theta(c)\}}$$



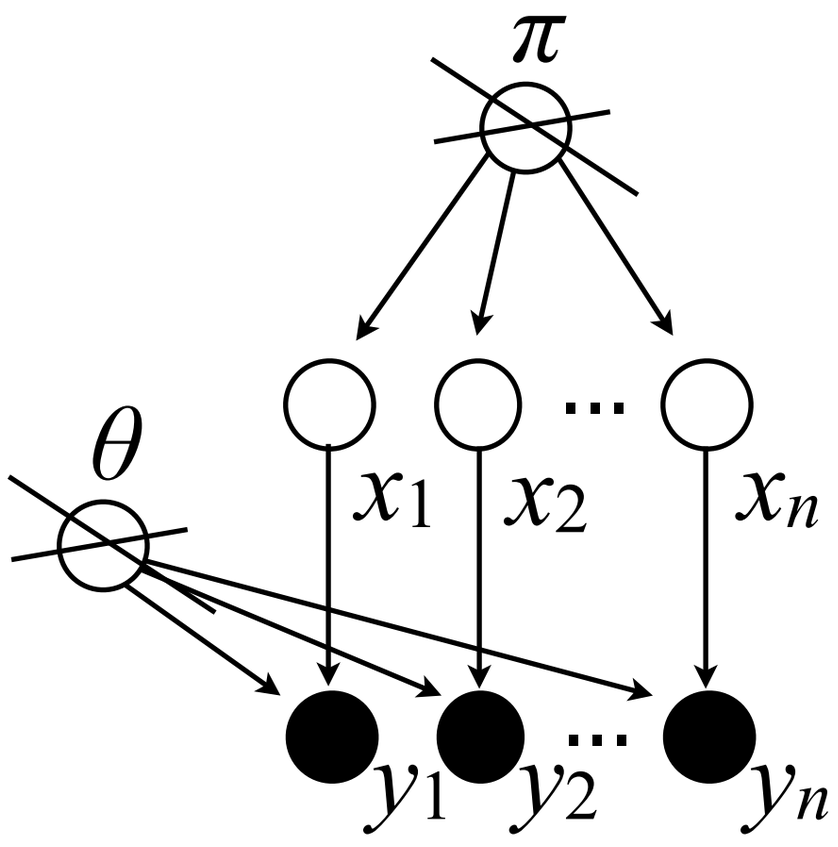
Probabilistic inference with DPs

Goal: computing a conditional expectation (e.g. for a Bayes estimator)

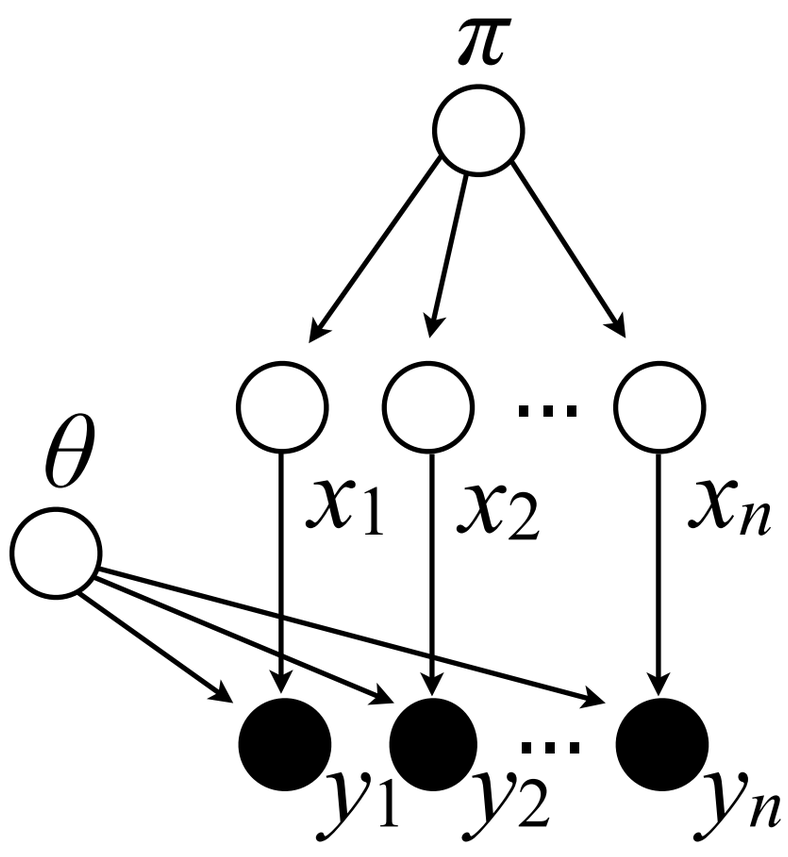
$$\mathbb{E}[f(\pi, \theta, x, y)|y]$$

We covered two samplers:

Collapsed sampler

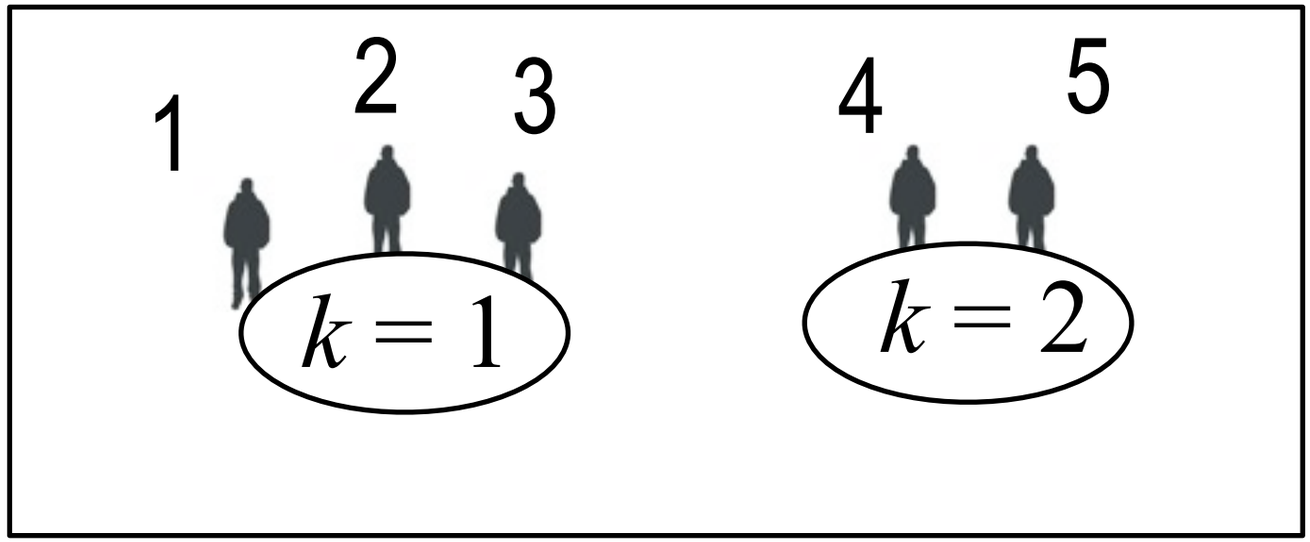


Slice sampler



Collapsed Gibbs sampler

Current state:



Notation: $L(\mathrm{d}y|\theta)$ = likelihood, $B \subseteq \{1, 2, ..., n\}$, and $L(\mathrm{d}y_B)$ = cluster marginal likelihood:

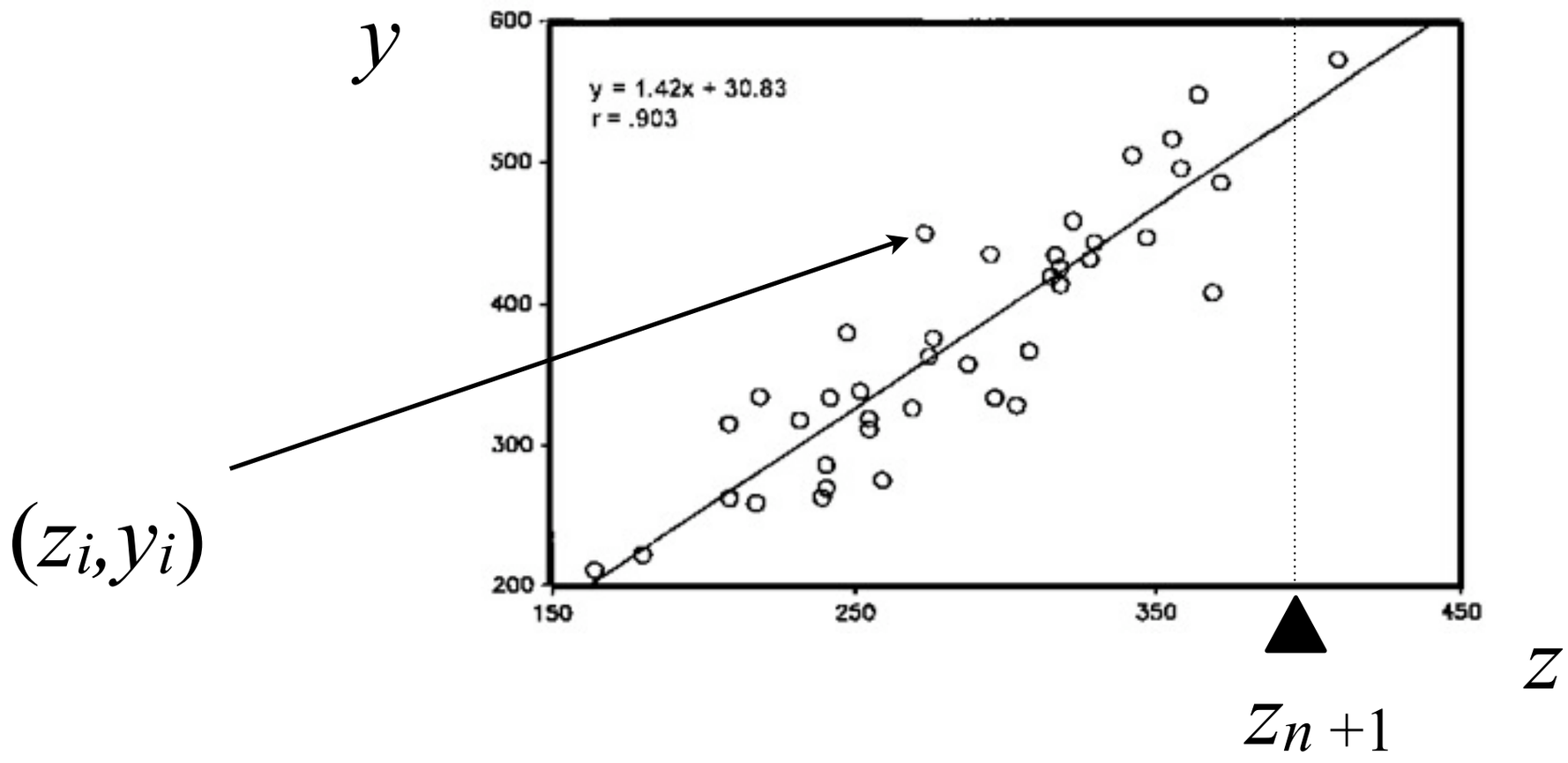
$$L(\mathrm{d}y_B) = \int \prod_{i \in B} L(\mathrm{d}y_i|\theta) G_0(\mathrm{d}\theta)$$

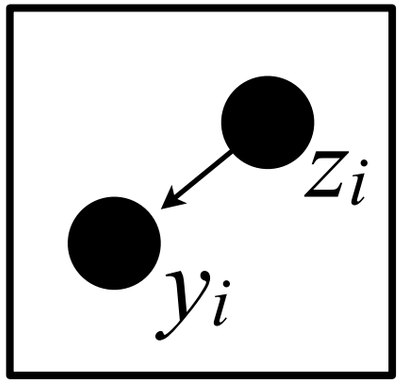
E.g.: $P(z_2 = k \mid \text{rest})$

(Derivation on the board)

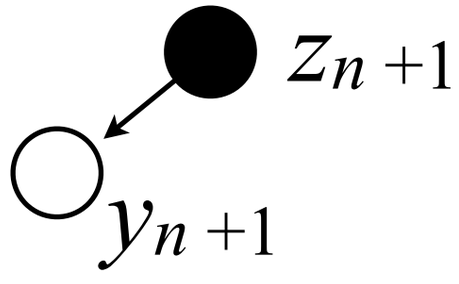
Applications of Dirichlet Processes: Regression

Regression: notation





Training data



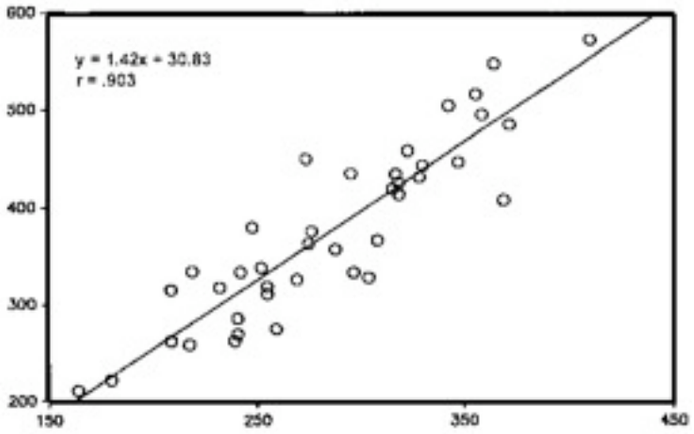
Test data/prediction

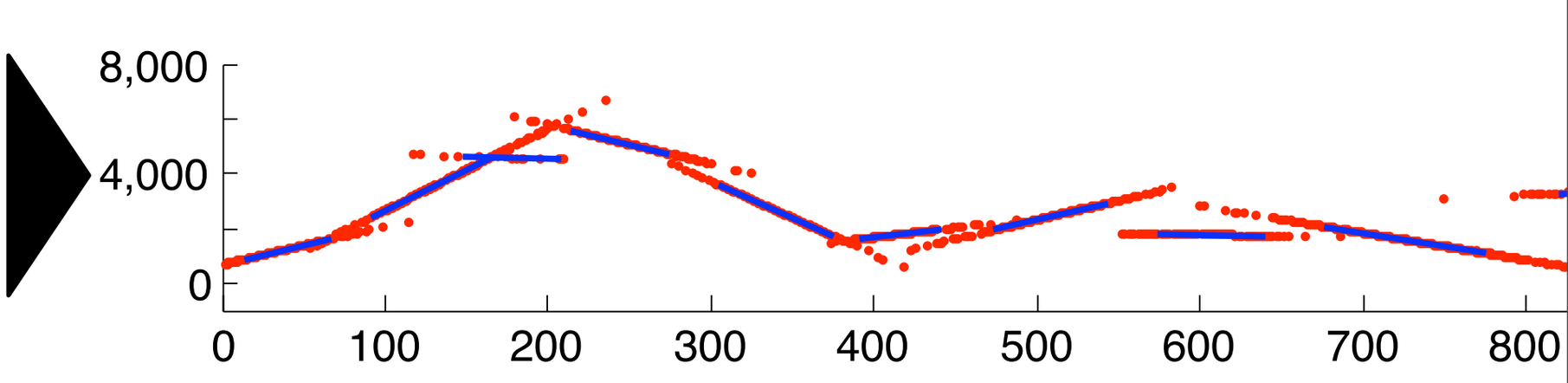
Input/covariate

Output/dep. var.

Goals

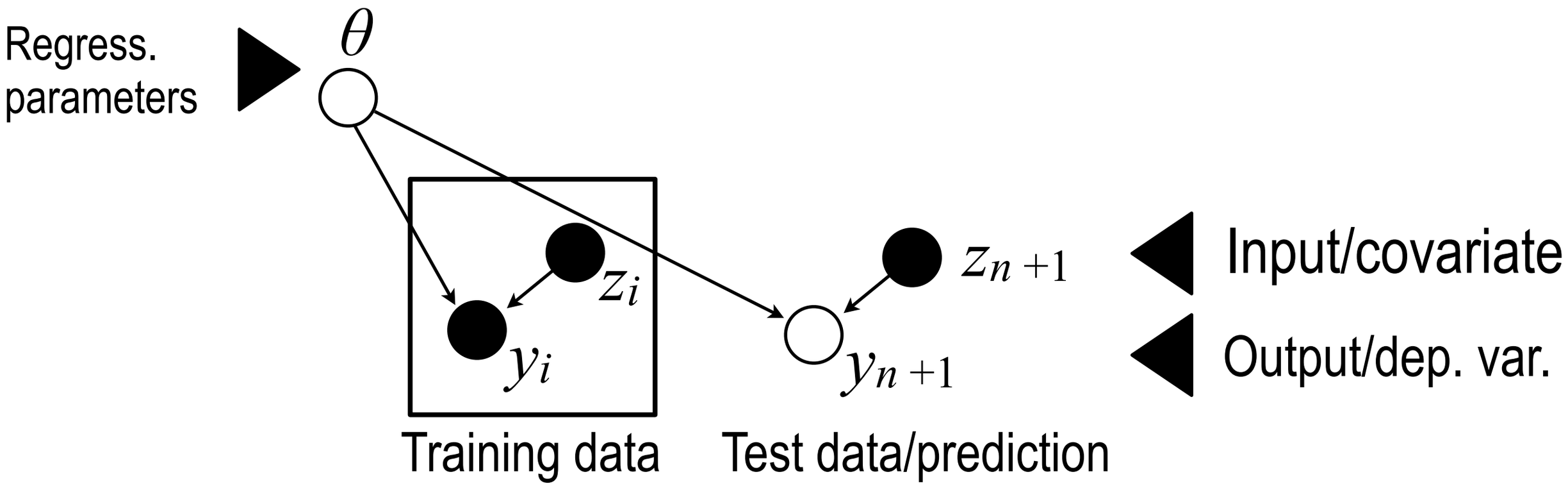
- Globally linear > locally linear
- More generally, globally GLM > locally GLM





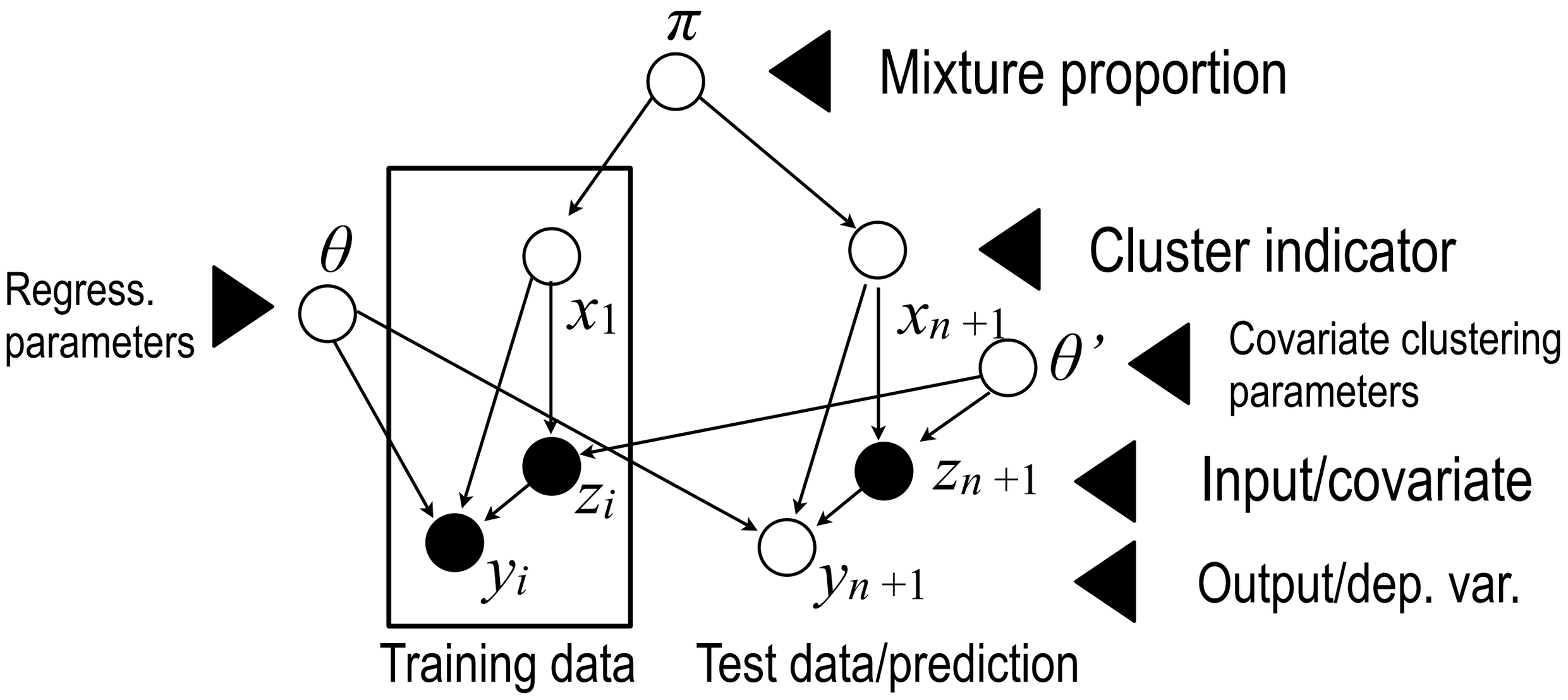
- Posterior distribution over predictions
- Optionally, over parameters as well

Basic Bayesian regression

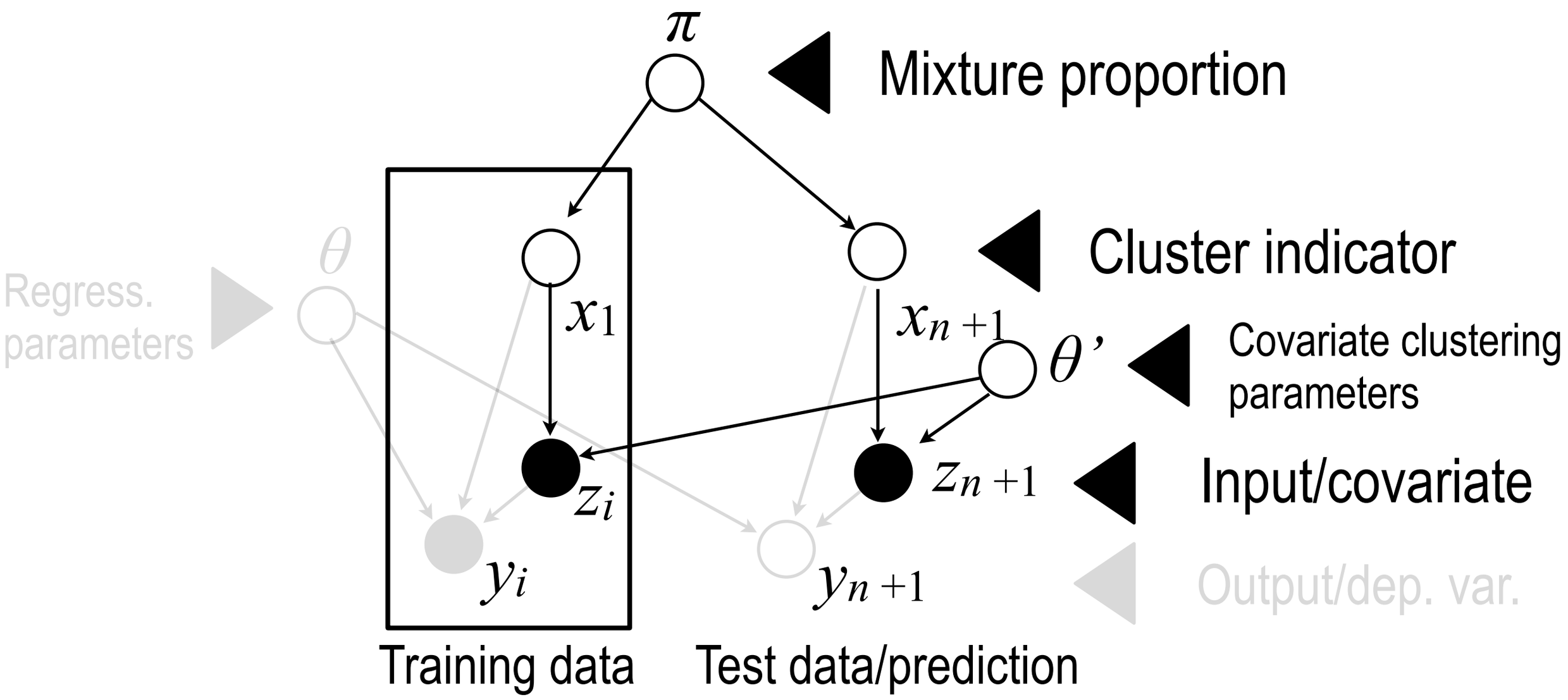


Note: in this basic setup, distribution on z_i does not affect prediction (but we will need dist on z later, so G-prior excluded)

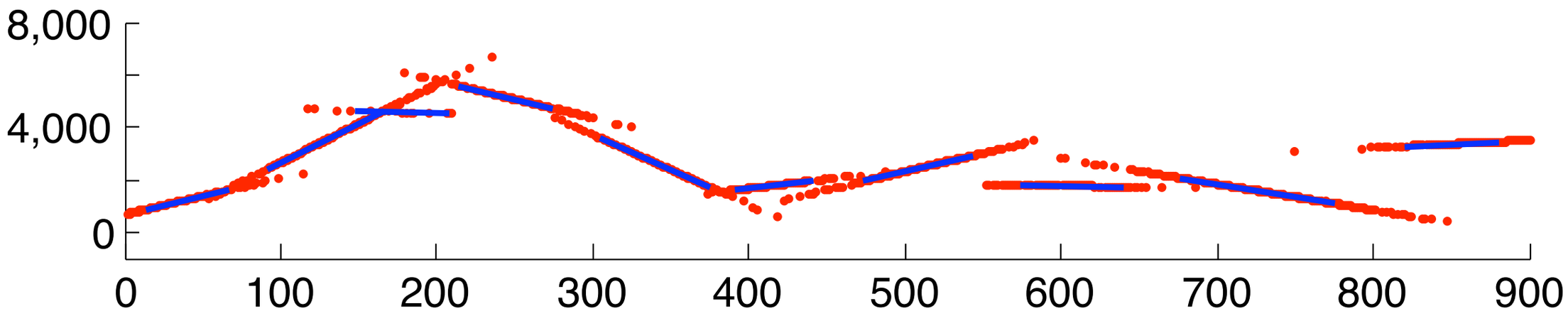
Nonparametric Bayesian regression



Nonparametric Bayesian regression



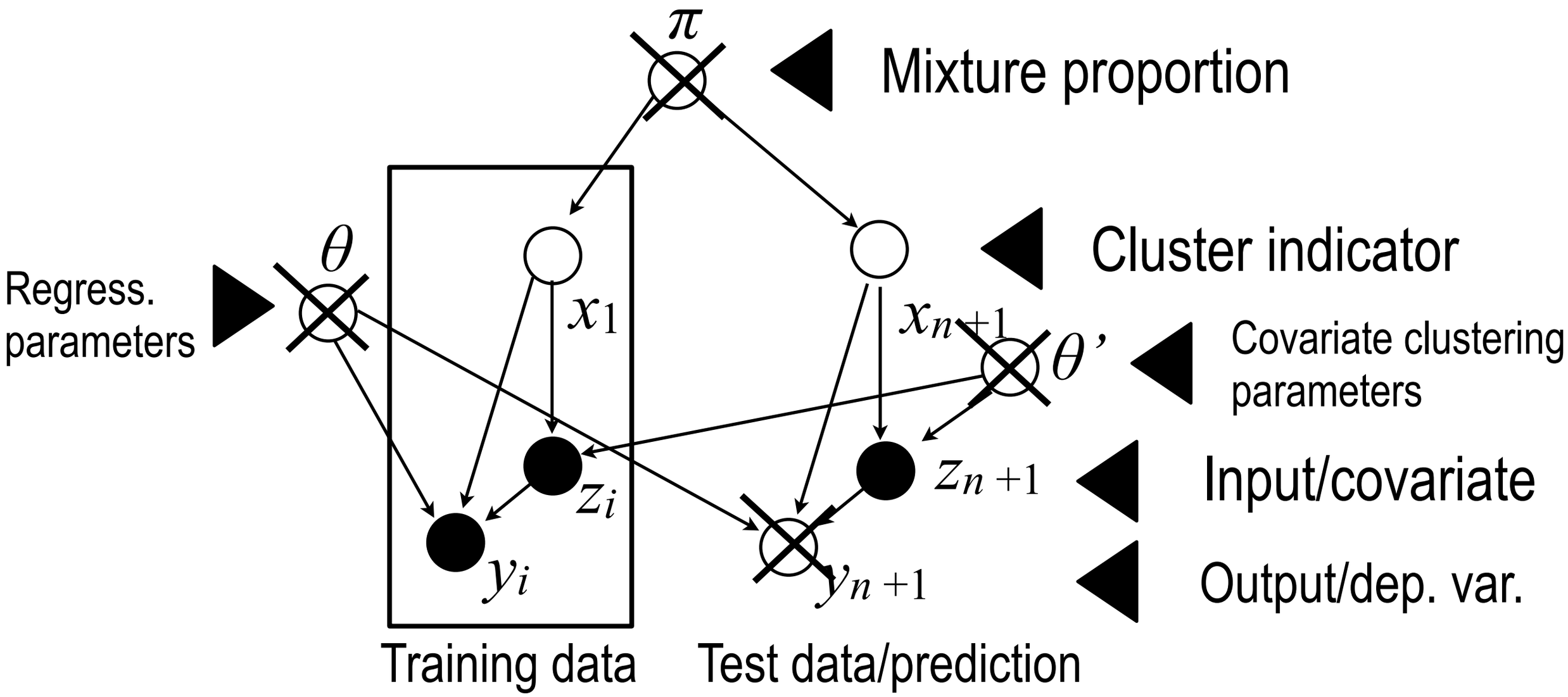
Intuition



Given a new datapoint, the prior on the z's enable us to get a posterior over which cluster it belongs to. For each cluster, we have a standard Bayesian linear regression model

Computing the posterior

Collapse sampling is possible:



Back to previous remark

Goal: computing a conditional expectation (e.g. for a Bayes estimator)

$$\mathbb{E}[f(\pi, \theta, x, y)|y]$$

Special case: sometimes, f depends only on the cluster indicators,

$$f(\pi, \theta, x, y) = f(x)$$

Example: clustering, where we only care about the posterior fraction of the time each pair of points is in the same mixture component

Note: can be made a bit less restrictive
(will come back to this point later)

Extensions

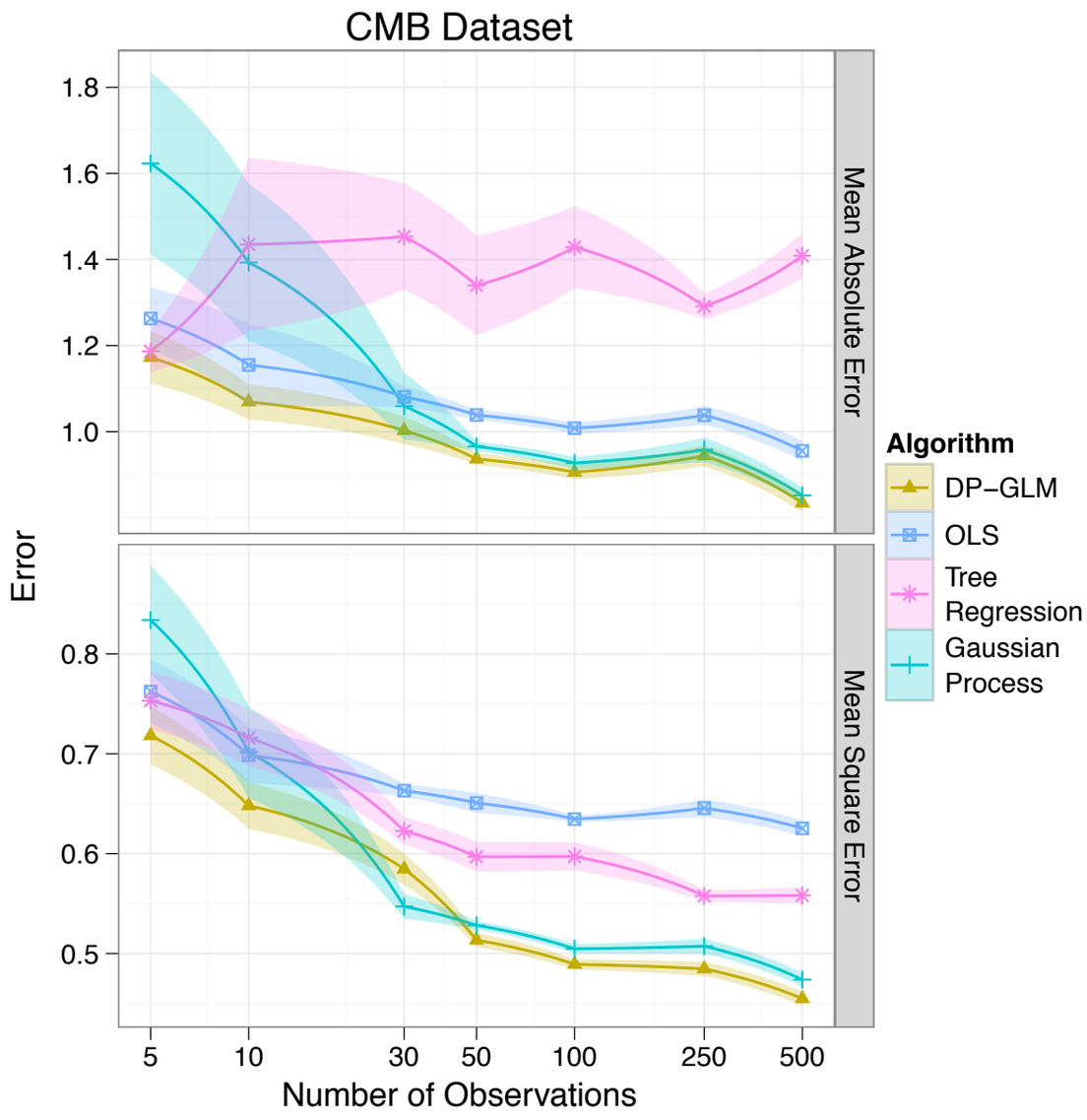
Other types of input/output:
Categorical/simplex, count, positive reals

Simple,unified model: replace Normal likelihoods by GLMs
Multinomial, Poisson, Gamma

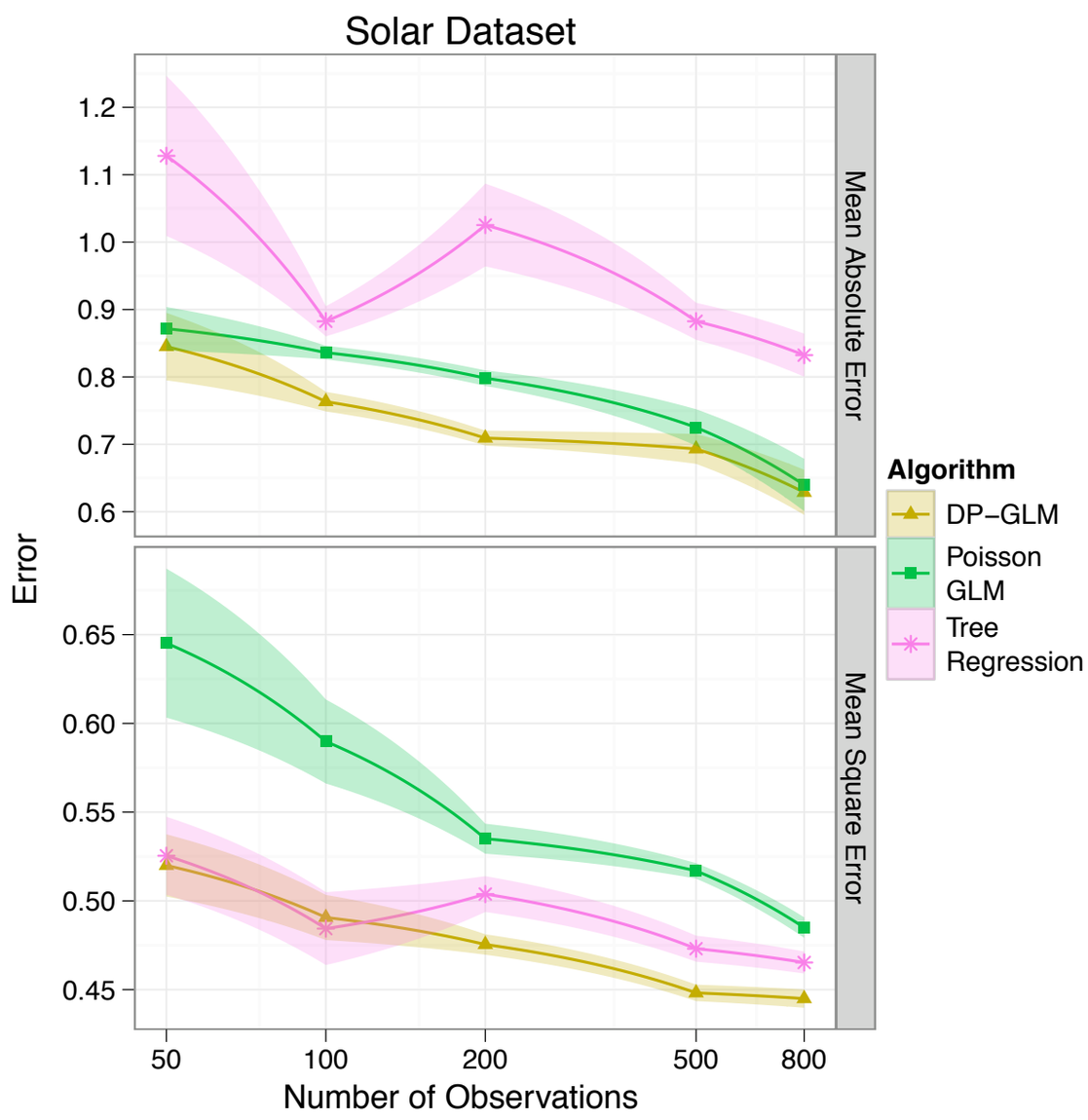
Difficulty: loss of analytic conjugate priors

Solution: use slice sampler or other auxiliary variables

Empirical evaluation



Continuous inputs & outputs



Continuous inputs & count outputs