Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

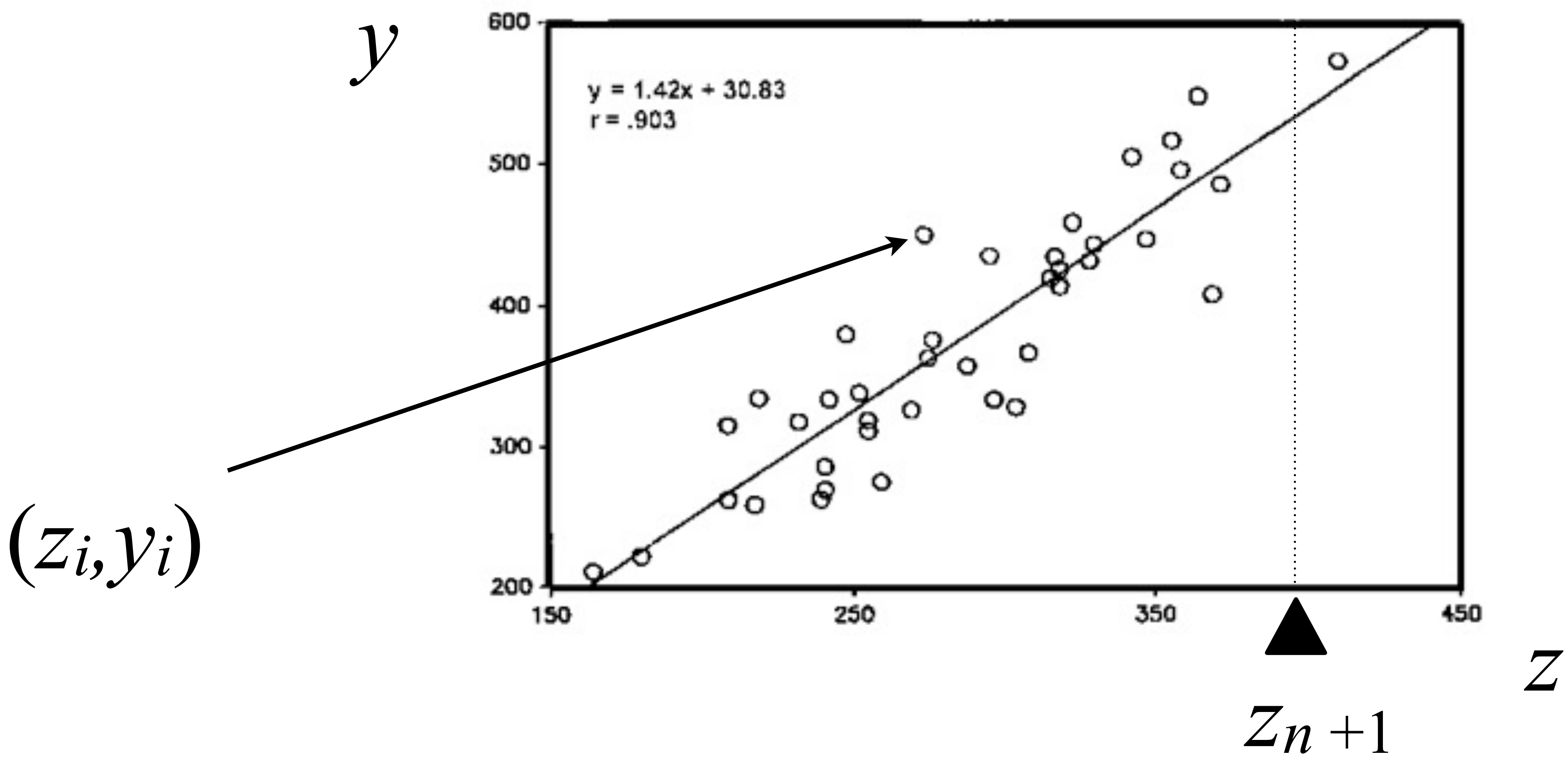
Lecture 9, Monday March 28

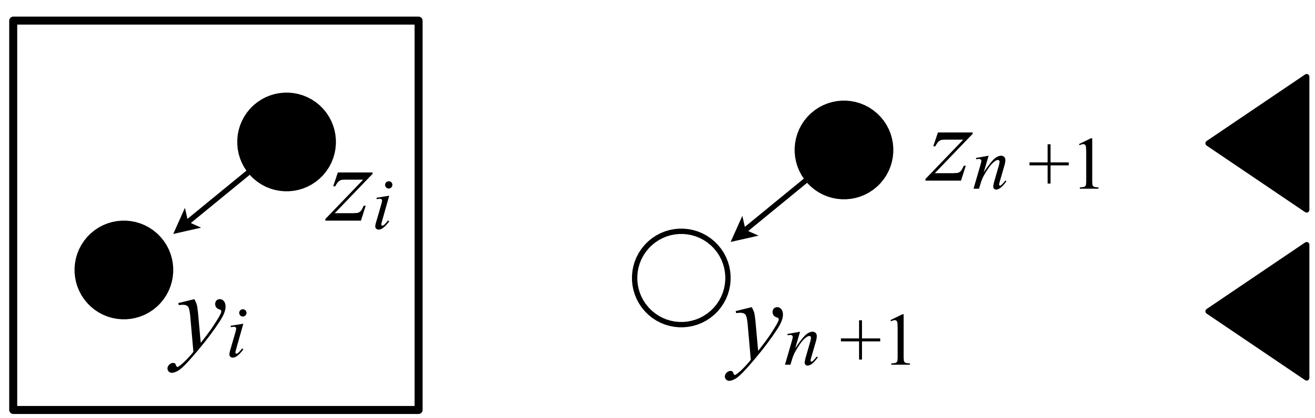
Program for today

- Applications
  - NLP: language modelling, segmentation, alignment
- Extensions
  - Hierarchies and sequences
  - Pitman-Yor & Beta processes

Review

Regression: notation





Training data

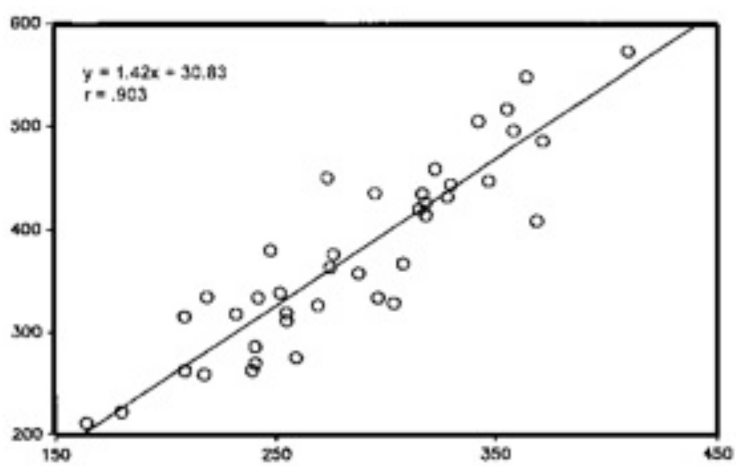
Test data/prediction

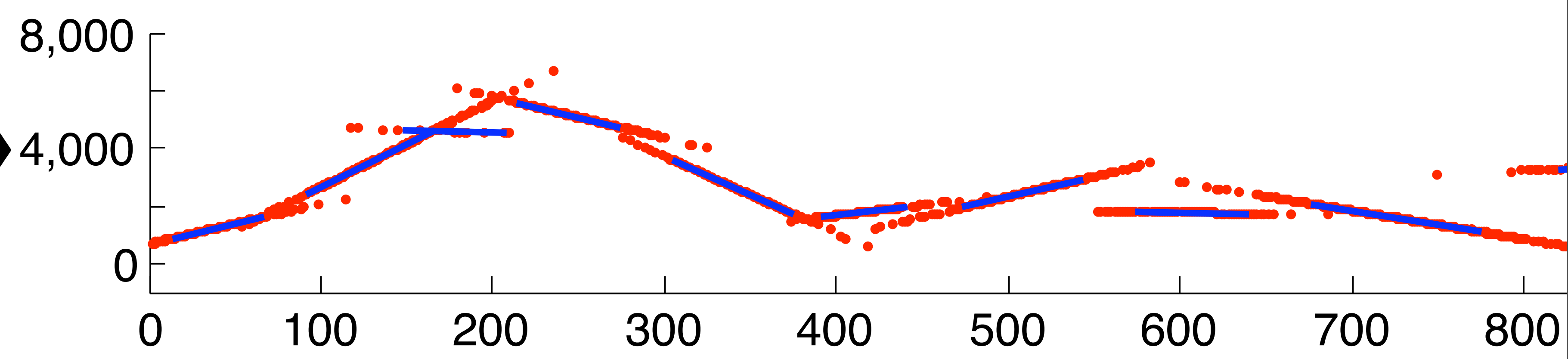
Input/covariate

Output/dep. var.

Goals

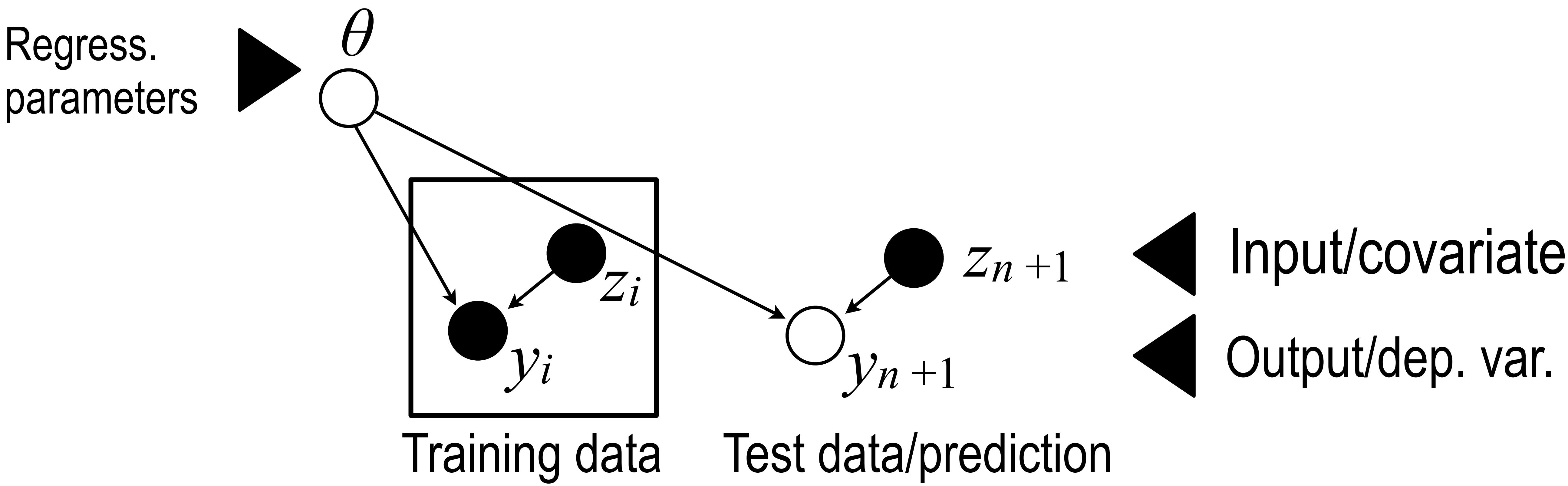
- Globally linear > locally linear
- More generally, globally GLM > locally GLM





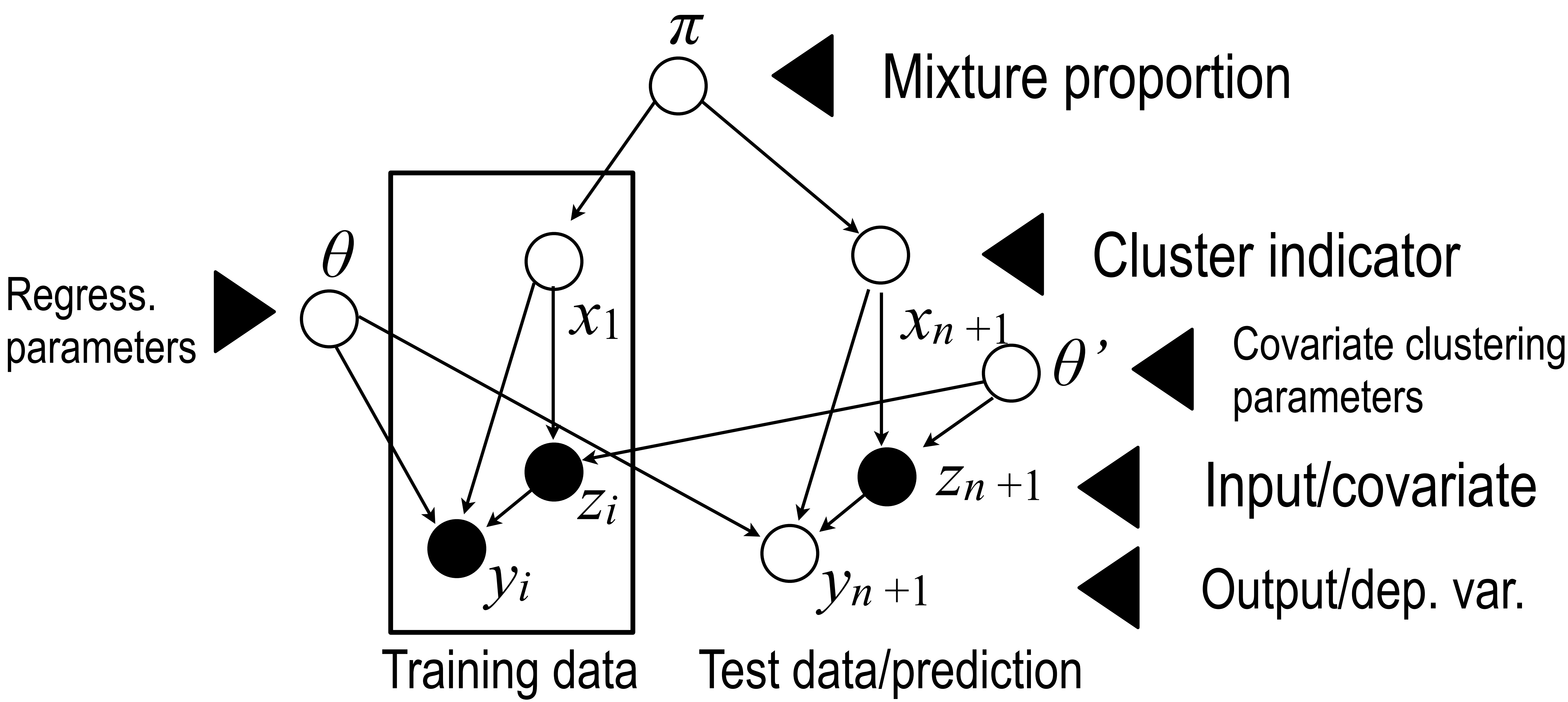
- Posterior distribution over predictions
- Optionally, over parameters as well

Basic Bayesian regression

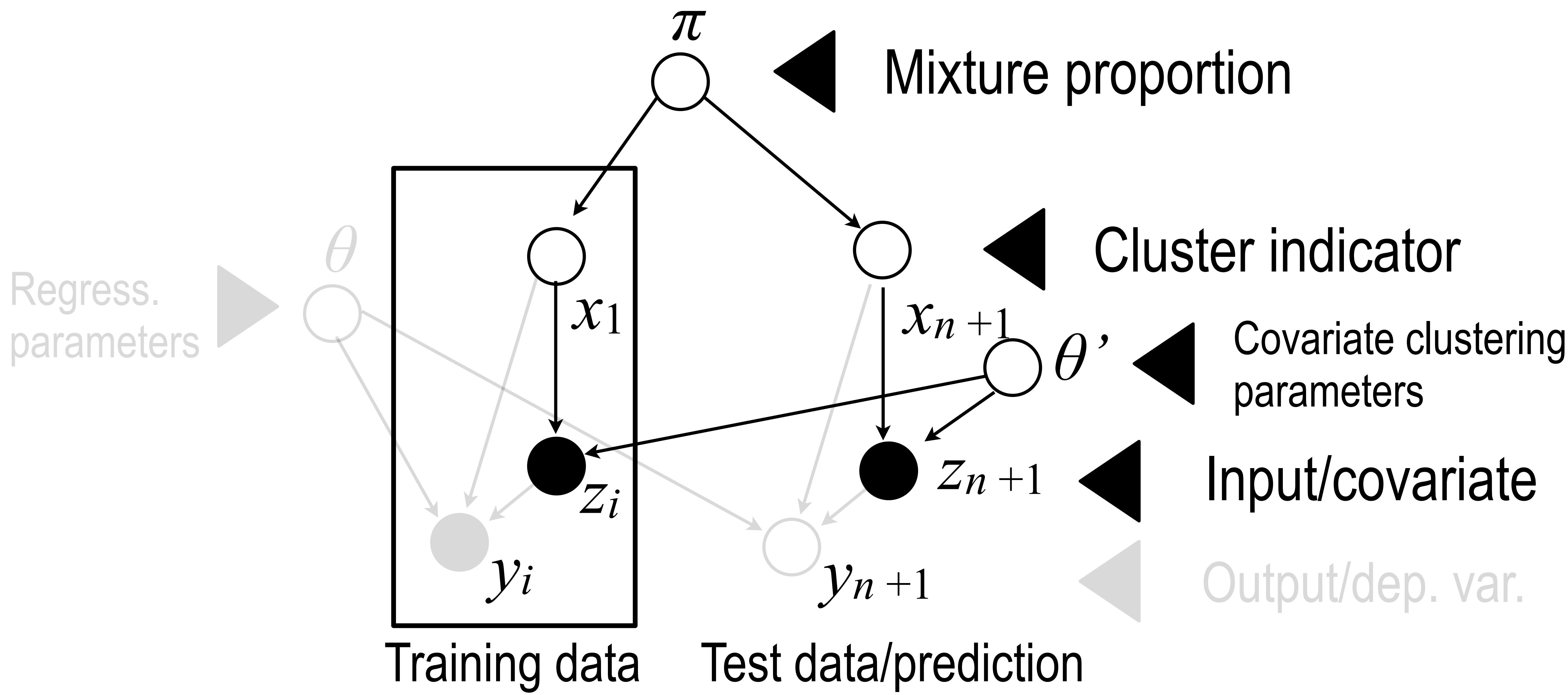


Note: in this basic setup, distribution on z_i does not affect prediction (but we will need dist on z later, so G-prior excluded)

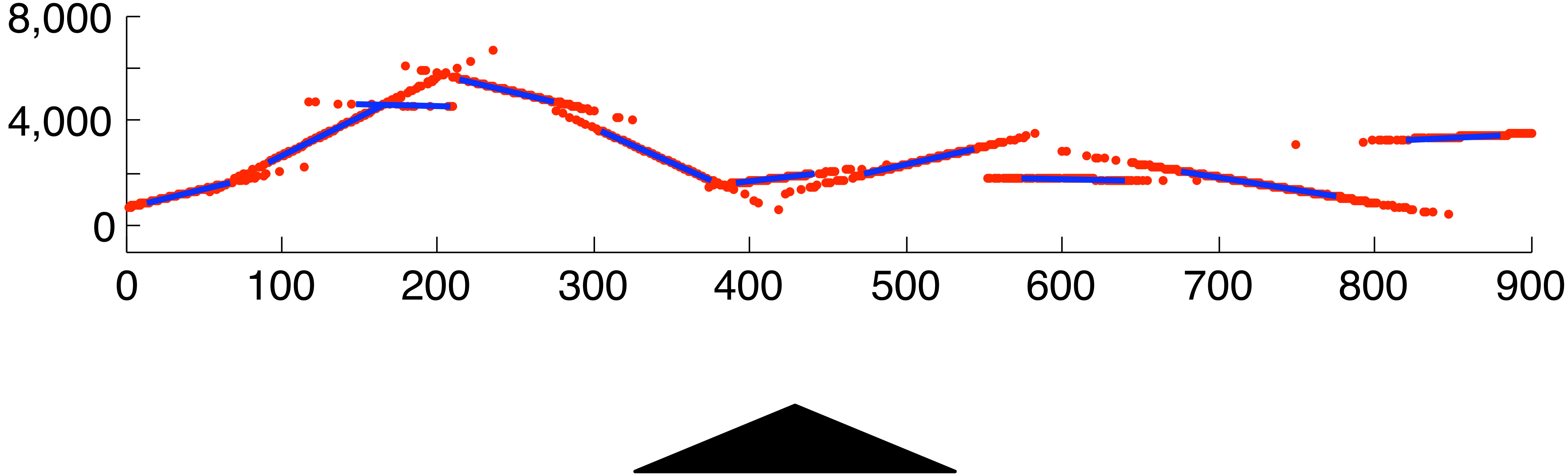
Nonparametric Bayesian regression



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

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

Applications of Dirichlet Processes in NLP

Language models

Shannon's game: guess the next word...

I have lived in San ______

I am not going to go ______

there or their?

Application: finding which sentence is more likely

Example: Speech recognition

Language models: first approach

Fix a certain prefix length, and estimate one categorical distribution for each prefix from a text dataset (*n*-gram)

Distribution over what follows after the prefix Fix ___

Guess	Pr
a	1.0

Distribution over what follows after the prefix a ___

Guess	Pr
certain	0.5
text	0.5

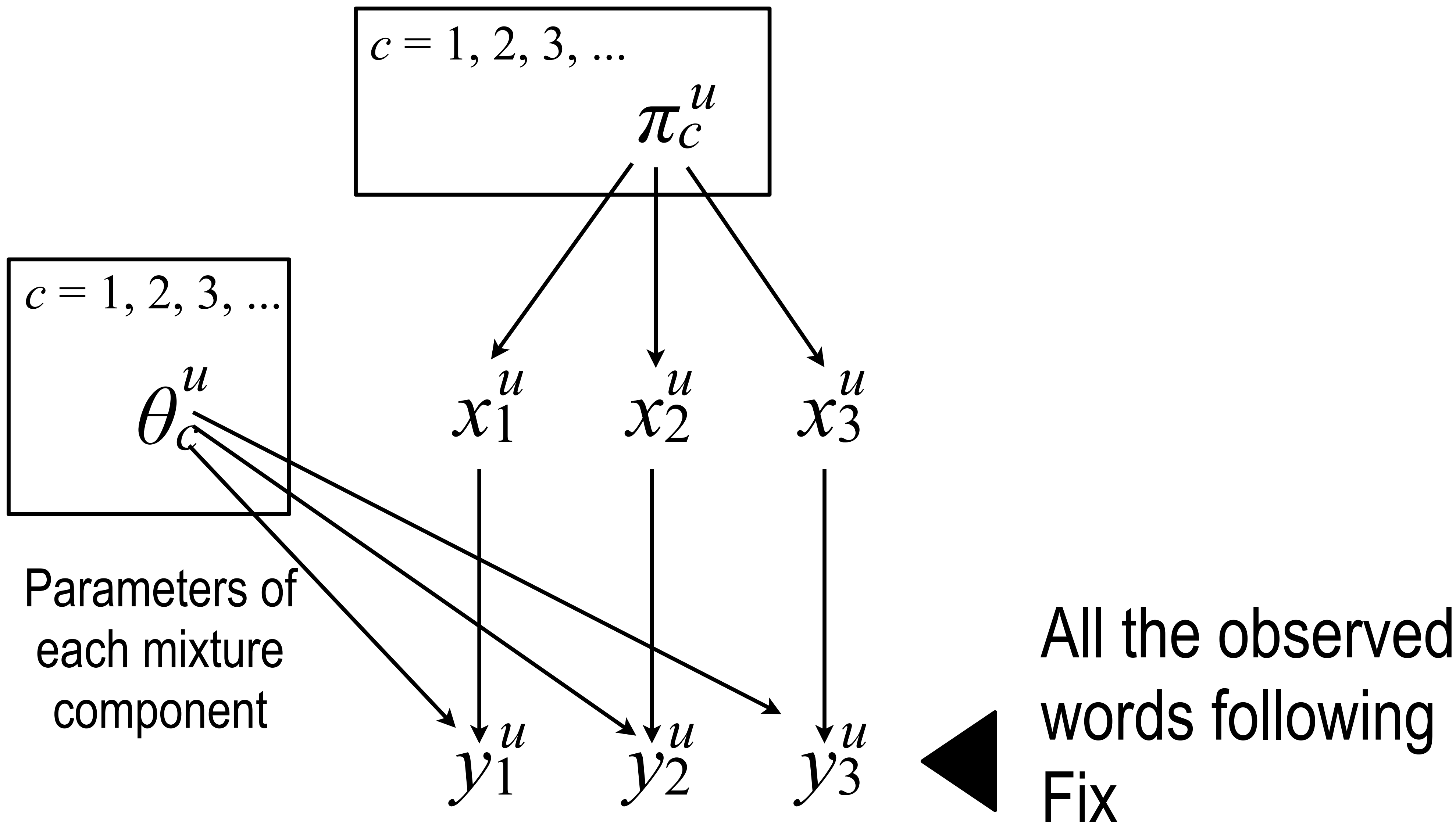
...

Problem with the maximum likelihood estimator?

First try: language model using DPs

Fix a prefix, e.g. u = (Fix __)

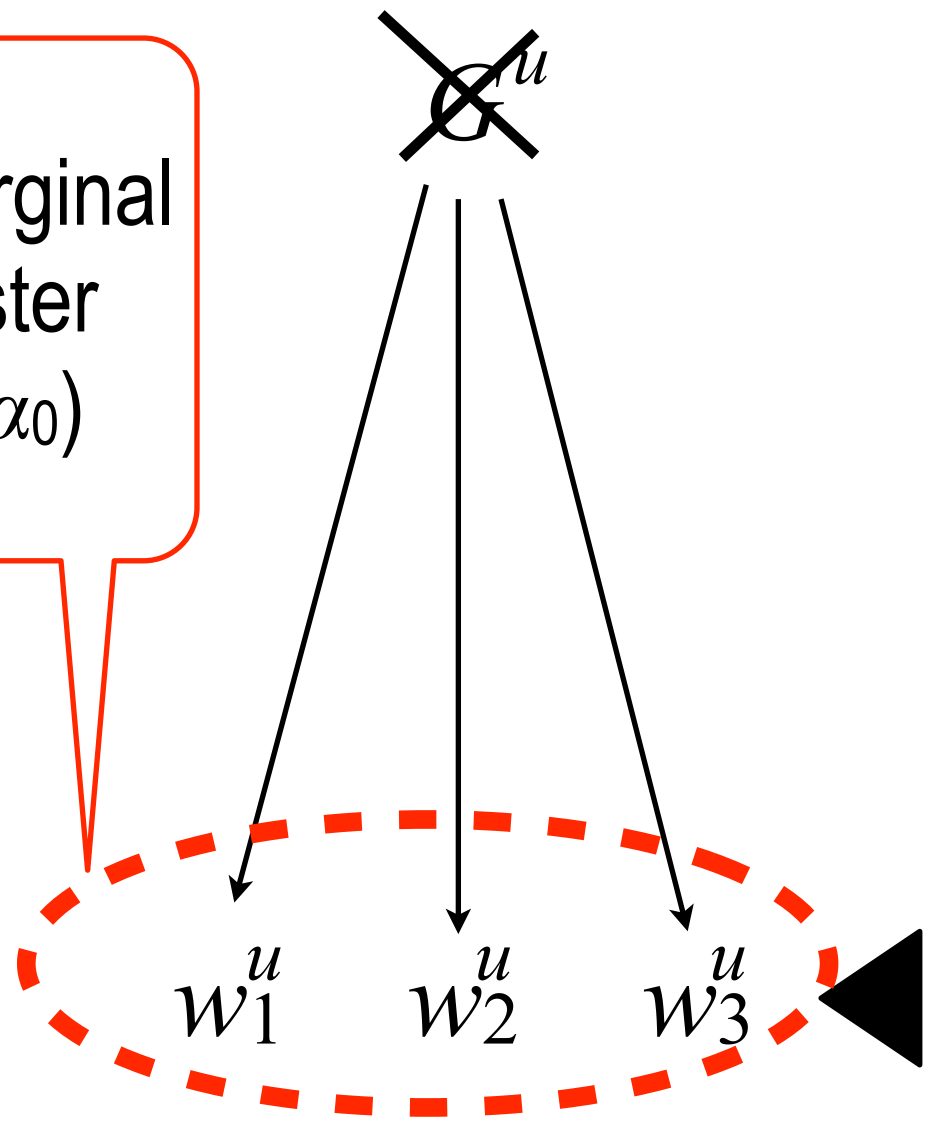
Model:



Alternative view to the CRP: cache model

Fix a prefix, e.g. u = (Fix __)

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

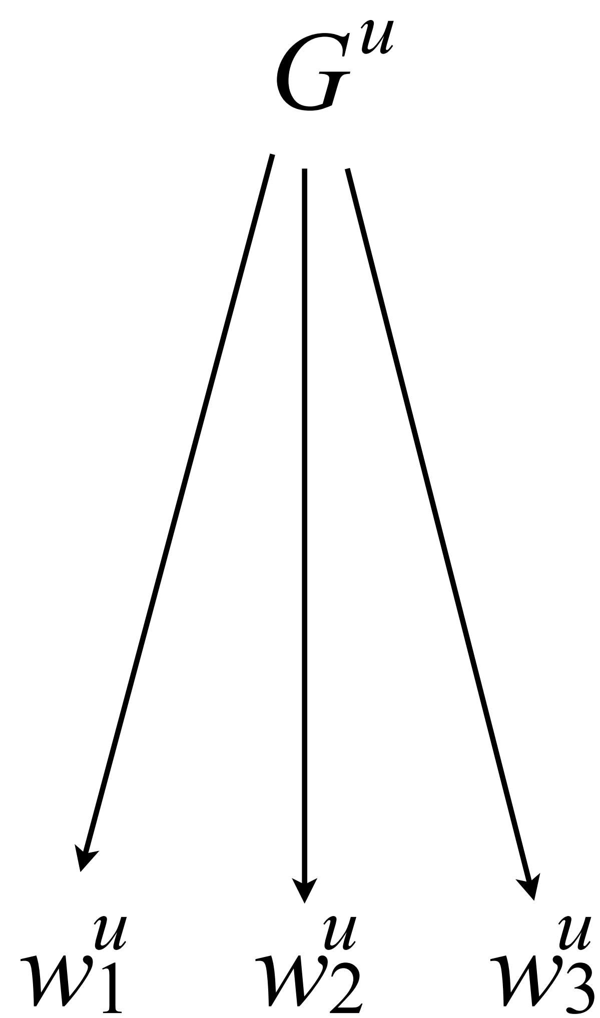


All the observed words following Fix ____

First try: language model using DPs

Fix a prefix, e.g. u = (Fix __)

Simplified model:



All the observed words following Fix ____

Problem...

Prior for prefix 1

Distribution over what follows after the prefix Fix ___

Guess	Pr
a	0.92
...	...
...	...

Prior for prefix 2

Distribution over what follows after the prefix a ___

Guess	Pr
certain	0.46
text	0.46
...	...

...

...

Some prefixes are rare. Is that a problem?

Solution: hierarchical model

Hyper-prior over words---not specific to a prefix

Distribution over words
in text dataset

Guess	Pr
the	0.04
a	0.02
...	...

Prior for prefix 1

Distribution over what follows after
the prefix
Fix ___

Guess	Pr
a	0.92
...	...
...	...

Prior for prefix 2

Distribution over what follows after
the prefix
a ___

Guess	Pr
certain	0.46
text	0.46
...	...

...

...

Another problem...

Dirichlet process does not have the right tail behavior!

Empirical observation: number of unique words (word types) in a natural language corpus containing n words tokens is O(n^s) for $s \in [1/2, 1)$

A simple asymptotic result

Expected number of tables t as number of customers n goes to infinity?

Note: the probability of creating a new table for a new customer $n+1$ does not depend on the previous sitting arrangement:

$$\mathbb{P}(\text{customer } n \text{ starts a new table}) = \frac{\alpha_0}{\alpha_0 + n}$$

Therefore: the number of tables is an harmonic sum, so the asymptotic number of tables is O($\log n$)

Soon: Pitman-Yor, a process that has O(n^d) table...