Application of Bayesian Inference in Population Genetics

DS 102

February 27, 2020

Clara Wong-Fannjiang

The Bayesian Way

Problem set-up

- Setting: we choose a prior $\mathbb{P}(\theta)$, a model $\mathbb{P}(X \mid \theta)$, and observe data *X*.
- Goal of Bayesian inference: get $\mathbb{P}(\theta \mid X)$

Two approaches for getting $\mathbb{P}(\theta \mid X)$:

1. Sampling (asymptotically correct)
2. Variational inference

Bayesian inference in action

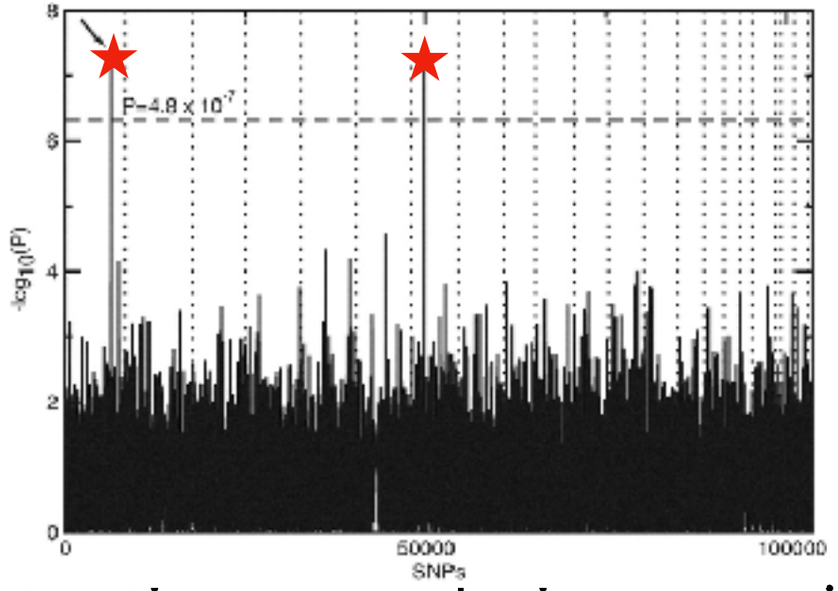
- Today, we will develop the Bayesian model and sampling method used in a population genetics method, STRUCTURE.
- This will expose some of the weaknesses of sampling, and motivate a light introduction to variational inference.

Application: Inferring population structure from genetic data

Problem Set-up

- Given genetic data from *N* individuals, how can we infer which of *K* populations they came from?
- We don't know what the populations "look like" beforehand.

Human migration patterns



Detecting population-specific bias in genetic studies

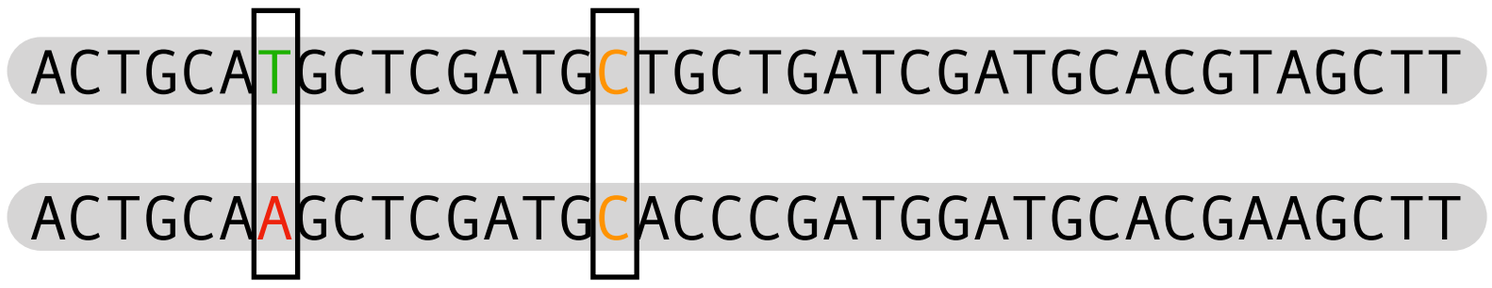
23andMe®

Personal ancestry

Application: Inferring population structure from genetic data

Genetics 101

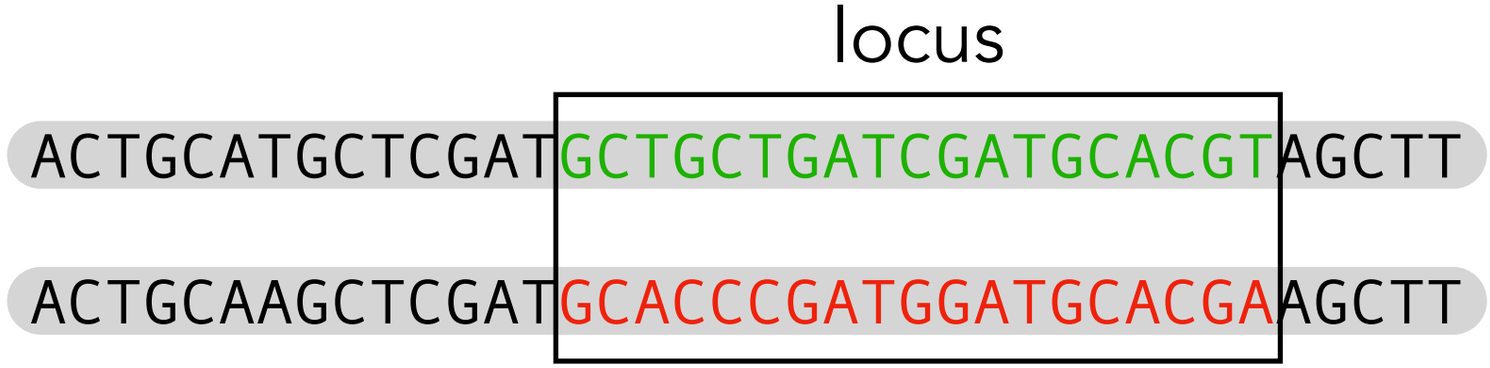
2 copies of genome



Application: Inferring population structure from genetic data

Genetics 101

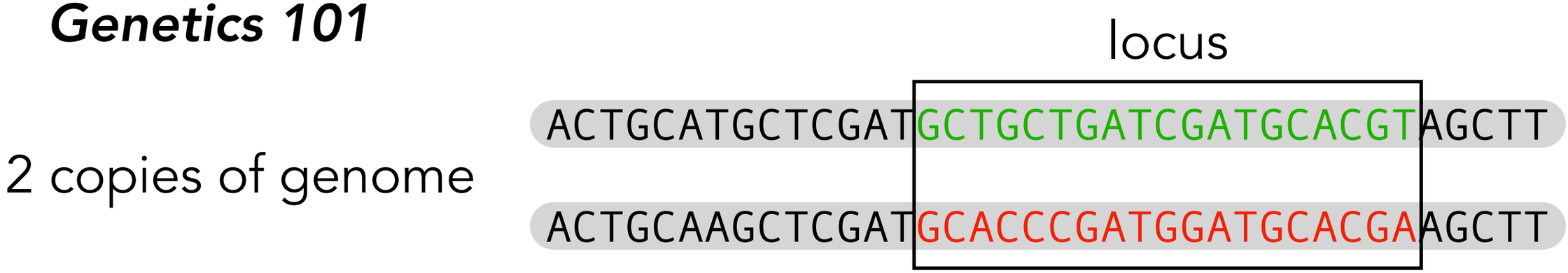
2 copies of genome



- **locus**: specific location on genome (single base, gene, etc.)
- **alleles**: states that a locus can take on
- **genotype** of an individual: the 2 alleles at each of *L* loci

Application: Inferring population structure from genetic data

Genetics 101



- **locus**: specific location on genome (single base, gene, etc.)
- **alleles**: states that a locus can take on
- **genotype** of an individual: the 2 alleles at each of *L* loci

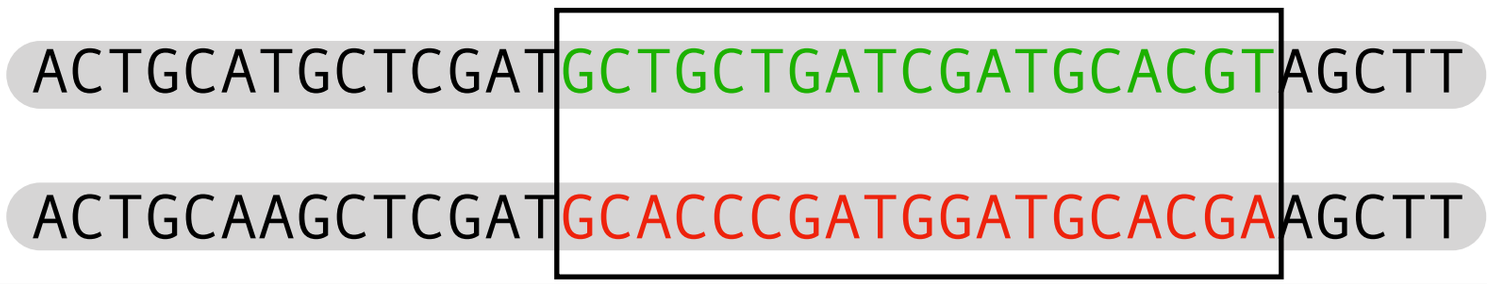
Problem: Given the genotypes of *N* individuals, can we infer which of *K* populations they came from, and what those populations look like?

Observation: Different populations of people tend to have distinctive genotypes.

We'll model a **population** as a set of *L* vectors, each of which gives the allele frequencies of a locus.

Application: Inferring population structure from genetic data

Problem Set-up



A **population** is modeled as a set of *L* vectors, each of which is a vector of allele frequencies at a locus.

p_{kl} : vector of allele frequencies at locus *l* in population *k*

p_{klj} : frequency of allele *j* at locus *l* in population *k*

The **population of origin** of an individual is a categorical variable.

$z^{(i)}$: population of origin of individual *i*

A **genotype** consists of the 2 alleles at each of *L* loci.

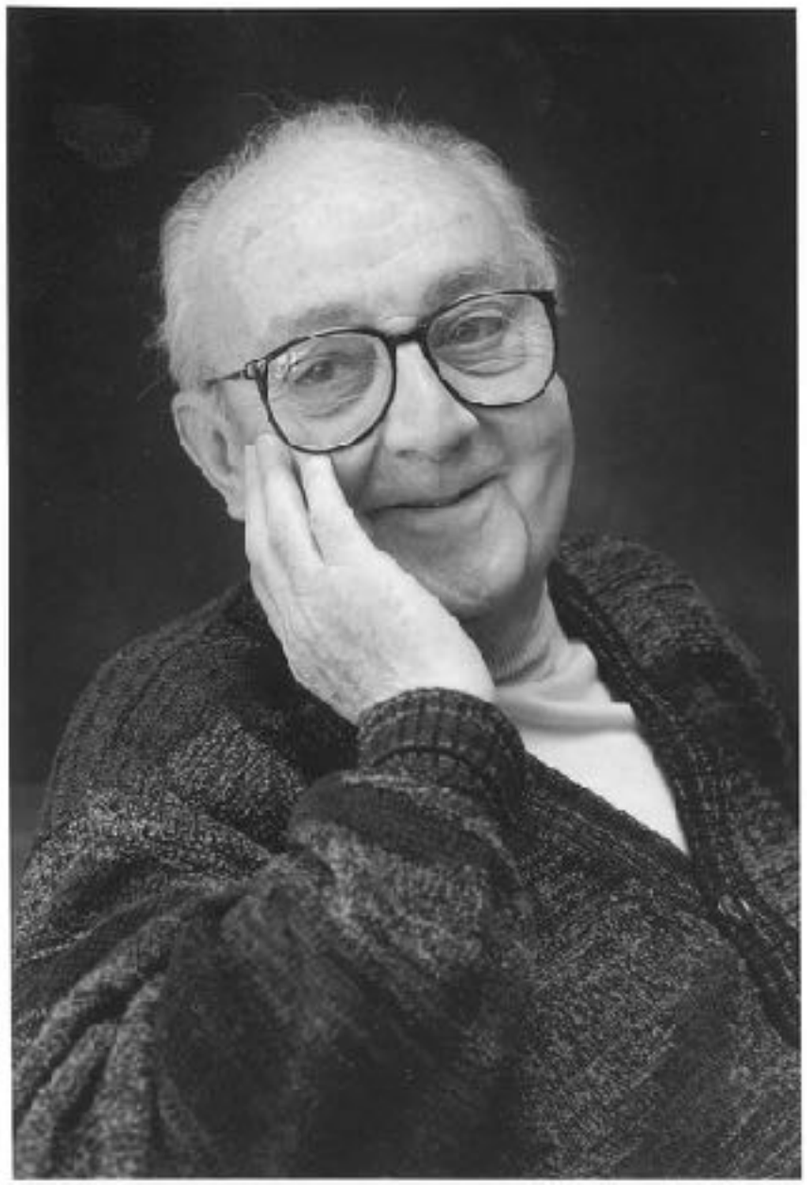
$(x_l^{(i,1)}, x_l^{(i,2)})$: 2 alleles (categorical variables) at locus *l* of individual *i*

$$P = \{p_{klj}\}_{k,l,j}, Z = \{z^{(i)}\}_i, X = \{(x_l^{(i,1)}, x_l^{(i,2)})\}_{i,l}$$

$$i = 1, \ldots, N, \quad j = 1, \ldots, J_l, \quad k = 1, \ldots, K, \quad l = 1, \ldots, L$$

Application: Inferring population structure from genetic data

How should we model this?



George Box (1919 - 2013)

"All models are wrong,
but some are useful."

"Statisticians, like
artists, have the bad
habit of falling in love
with their models."

Application: Inferring population structure from genetic data

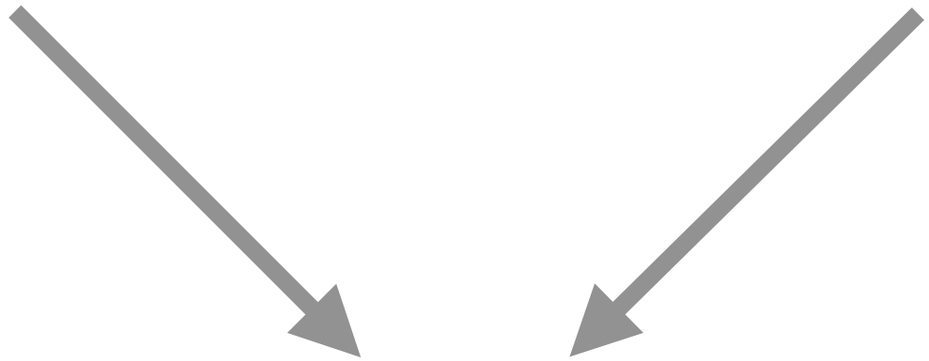
What does the dependency structure look like?

population allele frequencies

P

population of origin of individuals

Z

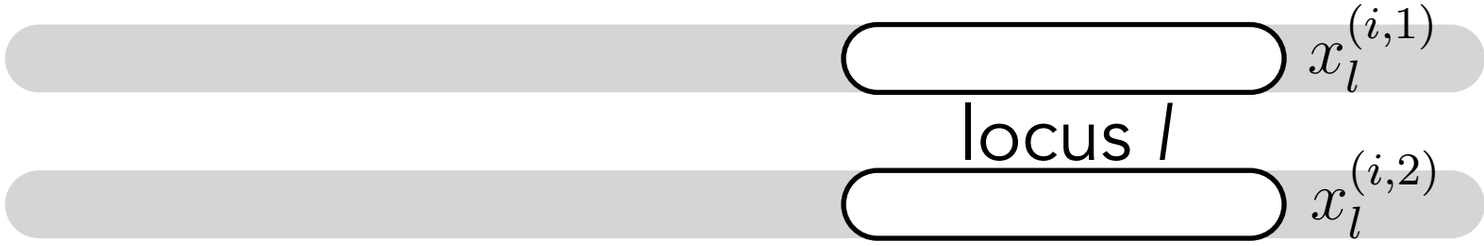


X

genotypes of individuals

Application: Inferring population structure from genetic data

How should we model the genotypes X?



- Conditioned on the populations of origin Z and population allele frequencies *P*, draw individuals' genotypes randomly based on the populations of origins' allele frequencies.

$$x_l^{(i,1)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$
$$x_l^{(i,2)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$

$$\mathbb{P}(x_l^{(i,1)} = j \mid Z, P) = p_{z^{(i)}lj},\ j = 1, \ldots, J_l$$
$$\mathbb{P}(x_l^{(i,2)} = j \mid Z, P) = p_{z^{(i)}lj},\ j = 1, \ldots, J_l$$

- Is it reasonable to model alleles as distributed independently across different loci? (Nope. Linkage disequilibrium.)

Application: Inferring population structure from genetic data

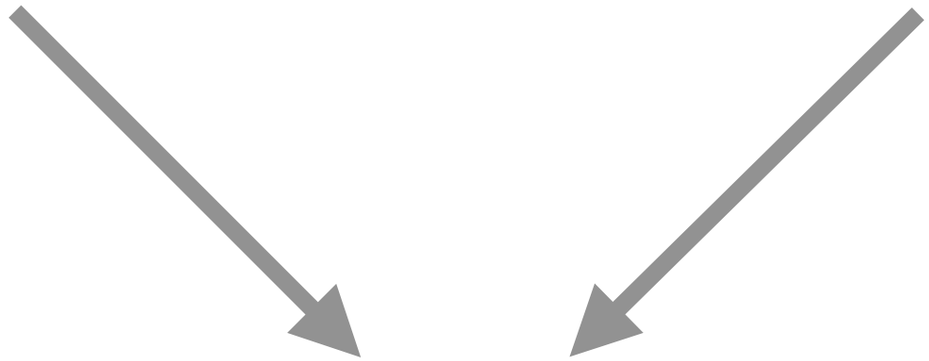
What does the dependency structure look like?

population allele frequencies

P

population of origin of individuals

Z



X

genotypes of individuals

Application: Inferring population structure from genetic data

What prior should we put on individuals' populations of origin Z?

- Assume individuals' populations of origin are drawn independently and identically as:

$$z^{(i)} \sim \text{Multinomial}\left(1, \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right)$$

$$\mathbb{P}(z^{(i)} = k) = \frac{1}{K},\ k = 1, \ldots, K$$

- When might this modeling decision be inappropriate?

Application: Inferring population structure from genetic data

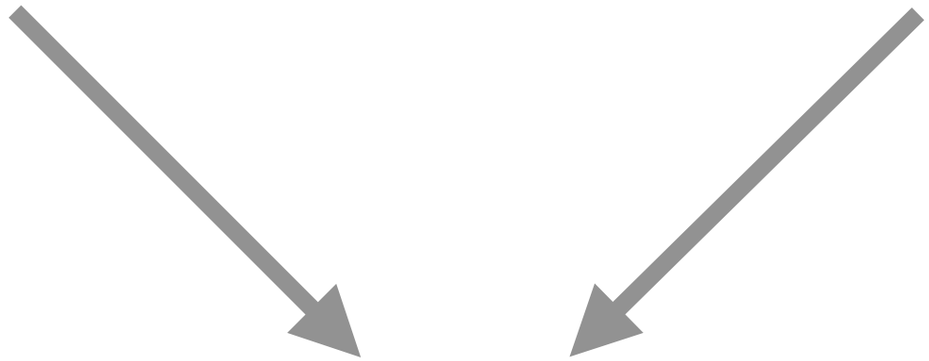
What does the dependency structure look like?

population allele frequencies

P

population of origin of individuals

Z



X

genotypes of individuals

Application: Inferring population structure from genetic data

What prior should we put on population allele frequencies P?

p_{kl} = 

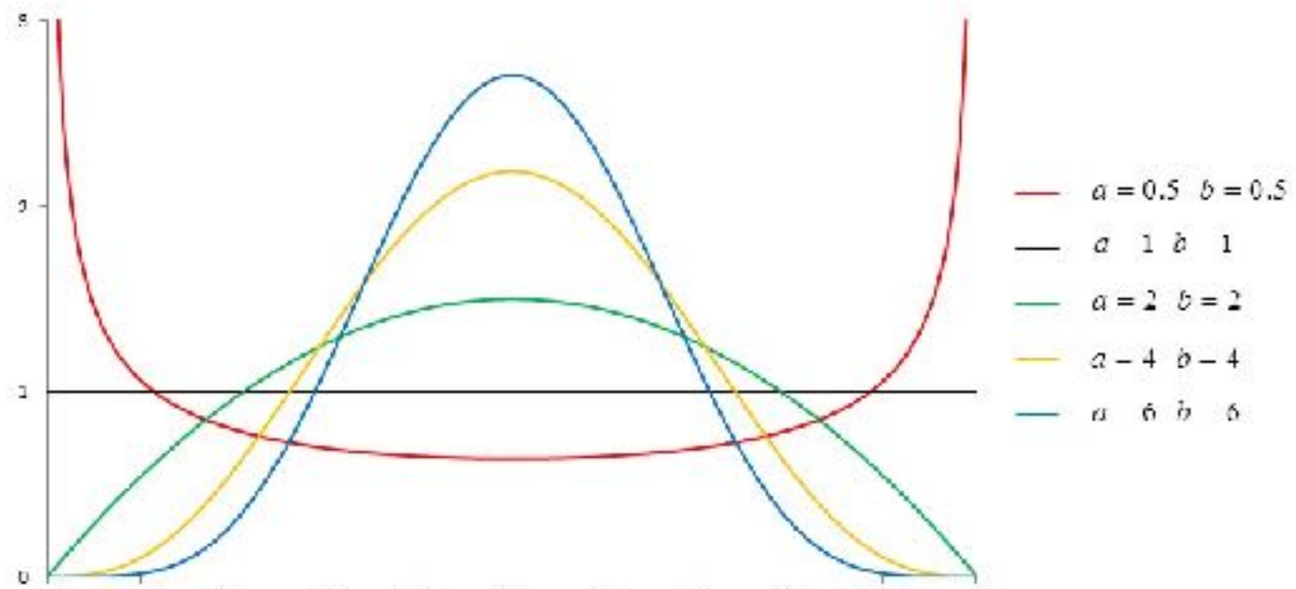
- Assume the p_{kl} are drawn independently and identically from a **Dirichlet distribution**.

$$p_{kl} \sim \text{Dirichlet}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l})$$

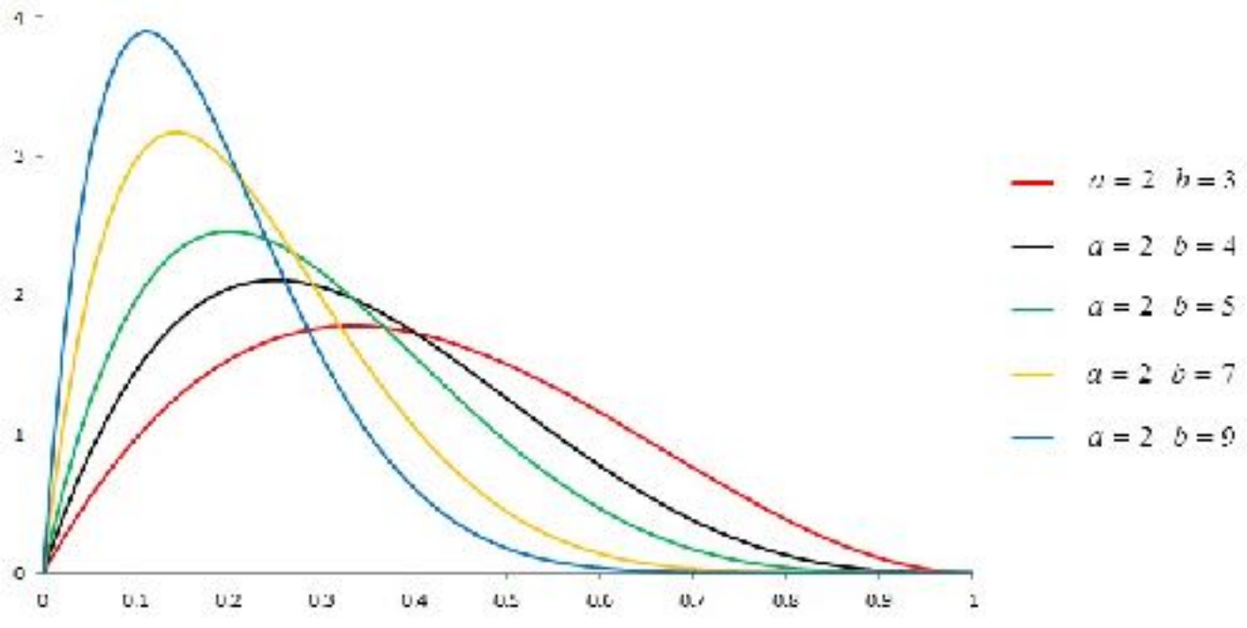
- Dirichlet distribution is a generalization of a Beta distribution to more than 2 categories.

Aside: the Dirichlet distribution!

- $\mathrm{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l})$ is a distribution over valid probability vectors.
- Generalization of a $\mathrm{Beta}(\alpha, \beta)$ to more than 2 categories.



$\alpha = \beta$
symmetric distribution



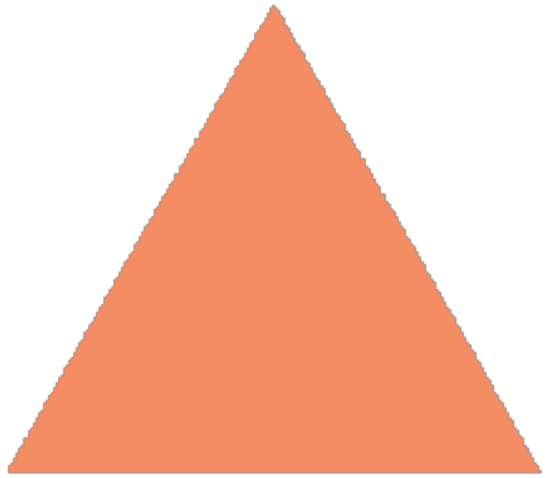
greater α, β :
more concentration

Aside: the Dirichlet distribution!

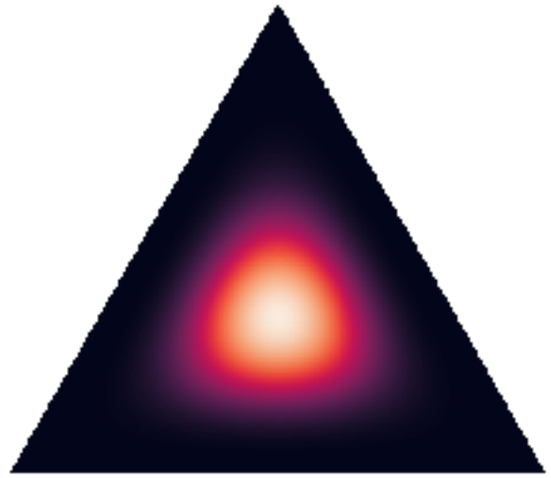
- $\mathrm{Dirichlet}(\lambda_1, \ldots, \lambda_{J_l})$ is a distribution over valid probability vectors.
- Generalization of a $\mathrm{Beta}(\alpha, \beta)$ to more than 2 categories.

(0.85, 0.85, 0.85)

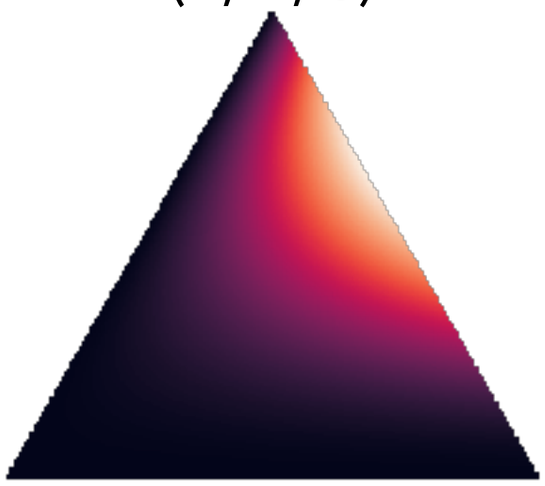
(1, 1, 1)



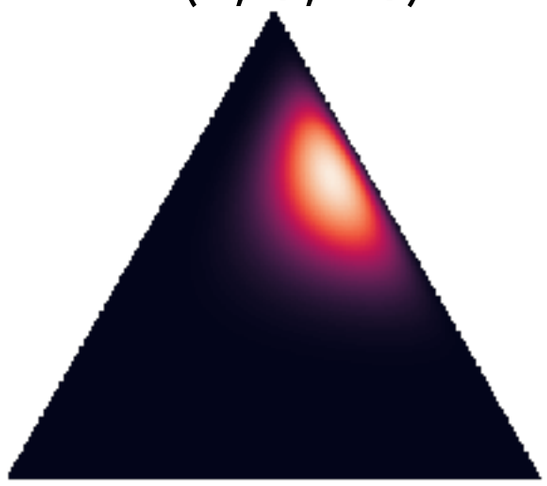
(5, 5, 5)



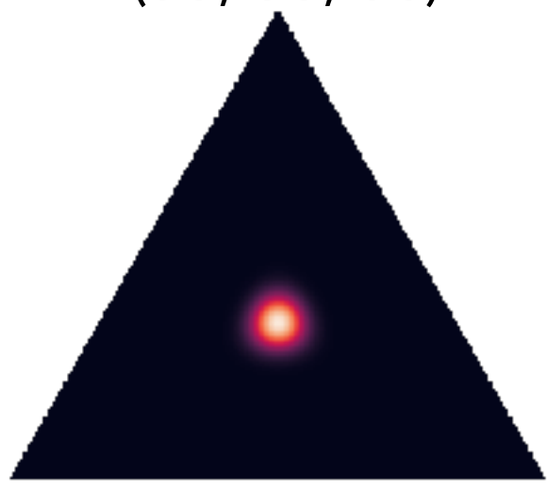
(1, 2, 3)



(2, 5, 10)



(50, 50, 50)



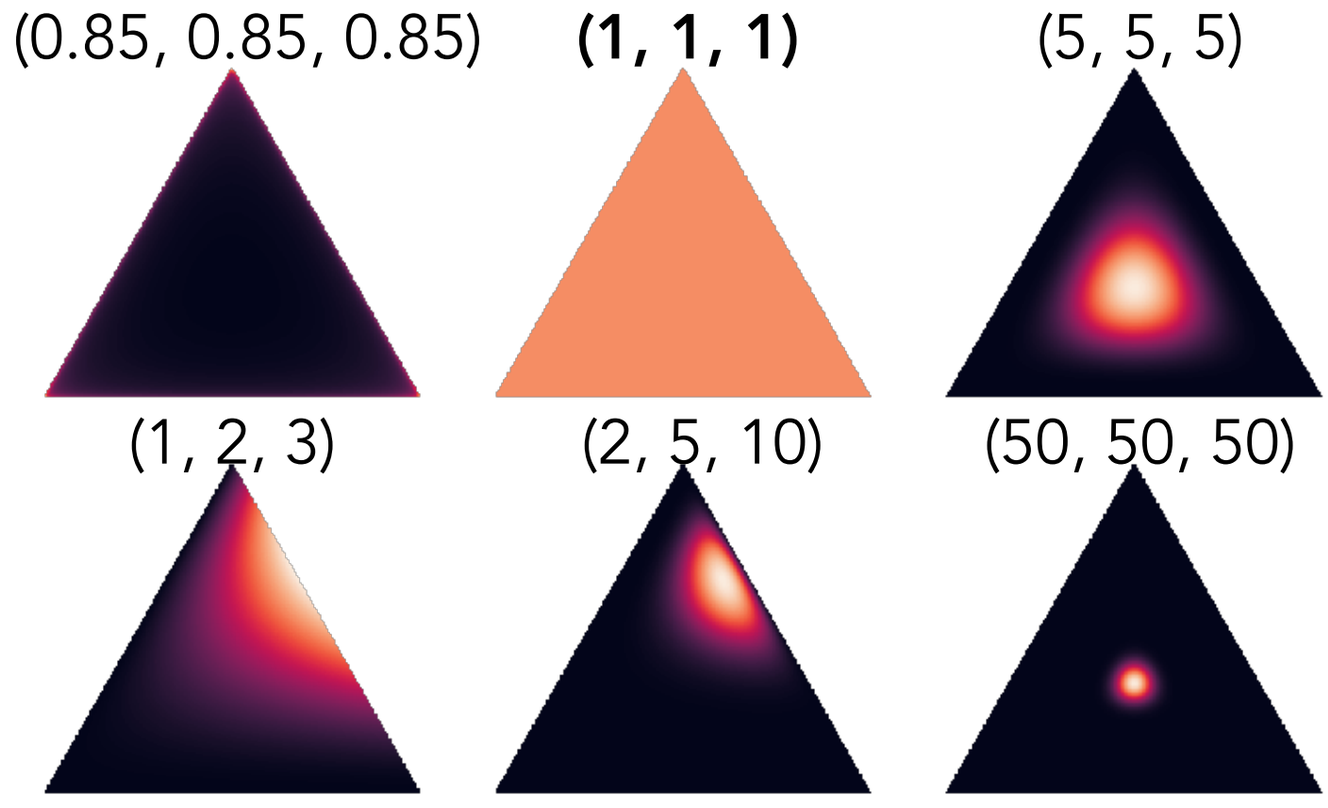
Application: Inferring population structure from genetic data

What prior should we put on population allele frequencies P?

$$p_{kl} = \boxed{0.3}\boxed{0.4}\boxed{0.2}\boxed{0.1}$$

- Assume the p_{kl} are drawn independently and identically from a **Dirichlet distribution**.

$$p_{kl} \sim \text{Dirichlet}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l})$$



- Uniform over all probability vectors

$$\mathbb{P}(p_{kl} = (1/J_l, \ldots, 1/J_l)) = \mathbb{P}(p_{kl} = (1, 0 \ldots, 0))$$

Application: Inferring population structure from genetic data

Putting everything together...

population allele frequencies

$$p_{kl} \sim \text{Dirichlet}(1, \ldots, 1)$$

population of origin of individual *i*

$$z^{(i)} \sim \text{Multinomial}\left(1, \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right)$$

P

Z

X

genotype of individual *i*

$$x_l^{(i,1)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$

$$x_l^{(i,2)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$

Application: Inferring population structure from genetic data

***Given X, how do we get* P(*P*, *Z* | *X*)?**

- Gibbs sampling! Markov Chain Monte Carlo (MCMC) method, meaning it produces a sequence of samples $\theta^{(1)}, \ldots, \theta^{(t)}$
- For sufficiently large *t*, these samples will be (approximately) independent random samples from the posterior.

Algorithm 1 Gibbs sampler for $\mathbb{P}(\theta|X)$, $\theta \in \mathbb{R}^d$

Initialize $\theta^{(0)}$.

for $t = 1, \ldots$ **do**

$$\theta_1^{(t)} \sim \mathbb{P}(\theta_1|X, \theta_2 = \theta_2^{(t-1)}, \ldots, \theta_d = \theta_d^{(t-1)})$$
$$\theta_2^{(t)} \sim \mathbb{P}(\theta_2|X, \theta_1 = \theta_1^{(t)}, \theta_3 = \theta_3^{(t-1)}, \ldots, \theta_d = \theta_d^{(t-1)})$$
$$\ldots$$
$$\theta_d^{(t)} \sim \mathbb{P}(\theta_d|X, \theta_1 = \theta_1^{(t)}, \theta_2 = \theta_2^{(t)}, \ldots, \theta_{d-1} = \theta_{d-1}^{(t)})$$

end for

Application: Inferring population structure from genetic data

***Given X, how do we get* P(*P*, *Z*| *X*)?**

Algorithm 2 Gibbs sampler for $\mathbb{P}(P, Z|X)$

Initialize $P^{(0)}, Z^{(0)}$.

for $t = 1, \ldots$ **do**

independent

$$p_{11}^{(t)} \sim \mathbb{P}(p_{11}|X, p_{12} = p_{12}^{(t-1)}, \ldots, z^{(1)} = z^{(1)(t-1)}, \ldots)$$

$$\ldots$$

$$p_{KL}^{(t)} \sim \mathbb{P}(p_{KL}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(1)} = z^{(1)(t-1)}, \ldots)$$

$$\ldots$$

independent

$$z^{(1)(t)} \sim \mathbb{P}(z^{(1)}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(n)} = z^{(n)(t-1)}, \ldots)$$

$$\ldots$$

$$z^{(n)(t)} \sim \mathbb{P}(z^{(1)}|X, p_{11} = p_{11}^{(t)}, \ldots, z^{(n-1)} = z^{(n-1)(t)}, \ldots)$$

end for

Application: Inferring population structure from genetic data

Given X, how do we get P(*P, Z*| *X*)?

Algorithm 3 Block Gibbs sampler for $\mathbb{P}(P, Z|X)$

Initialize $P^{(0)}, Z^{(0)}$.

for $t = 1, \ldots$ **do**

$$P^{(t)} \sim \mathbb{P}(P|X, Z = Z^{(t-1)})$$
$$Z^{(t)} \sim \mathbb{P}(Z|X, P = P^{(t)})$$

end for

$$p_{kl} \sim \text{Dirichlet}(1, \ldots, 1)$$

$$z^{(i)} \sim \text{Multinomial}\left(1, \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right)$$

$$x_l^{(i,1)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$
$$x_l^{(i,2)} \sim \text{Multinomial}(1, (p_{z^{(i)}l1}, p_{z^{(i)}l2}, \ldots, p_{z^{(i)}lJ_l}))$$

posterior of *P*: Dirichlet

$$p_{kl} \mid X, Z \sim \text{Dirichlet}(1 + n_{kl1}, \ldots, 1 + n_{klJ_l})$$
$$n_{klj} = |\{(i, a) : x_l^{(i,a)} = j, z^{(i)} = k\}|$$

posterior of Z: multinomial

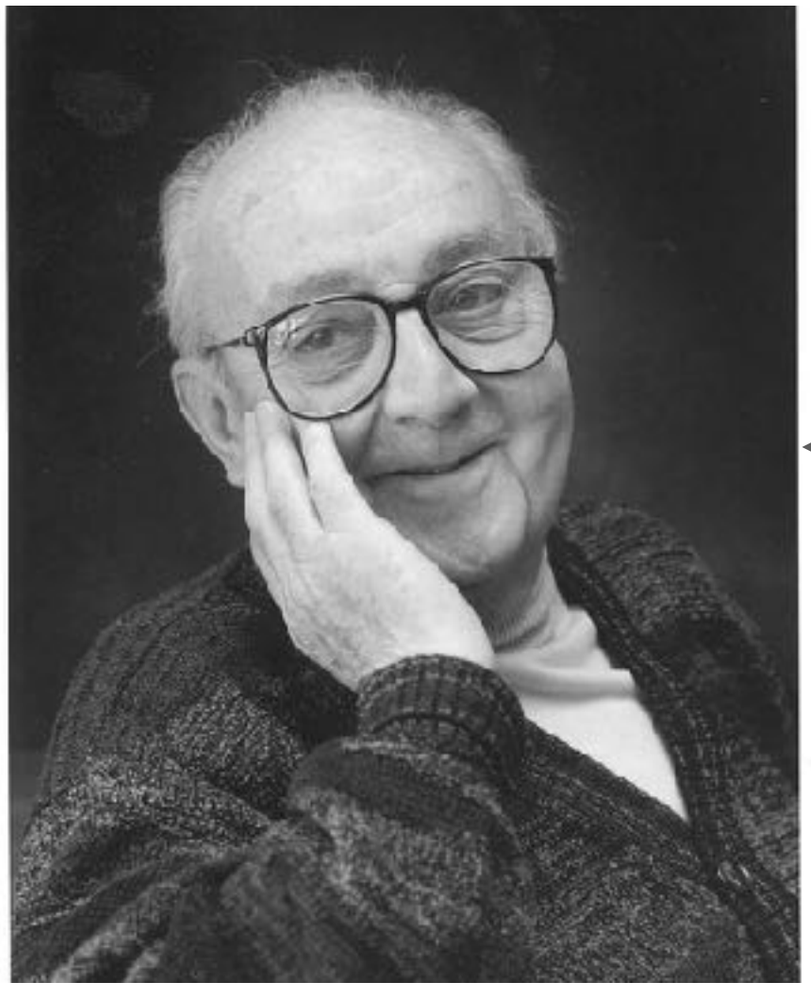
$$\mathbb{P}(z^{(i)} = k \mid X, P) = \frac{\mathbb{P}(x^{(i)} \mid P, z^{(i)} = k)}{\sum_{k'=1}^{K} \mathbb{P}(x^{(i)} \mid P, z^{(i)} = k')}$$
$$\mathbb{P}(x^{(i)} \mid P, z^{(i)} = k) = \prod_{l=1}^{L} p_{klx^{i,1}} p_{klx^{i,2}}$$

Application: Inferring population structure from genetic data

How should we model the individuals' populations of origin?

- Assume individuals' populations of origin are drawn as

$$z^{(i)} \sim \text{Multinomial}\left(1, \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right)$$



Does an individual really just have **one** population of origin?

Application: Inferring population structure from genetic data

How should we model the individuals' populations of origin?

Admixture is when a genotype has multiple populations of origin.

$q_k^{(i)}$: proportion of individual *i*'s genotype that originate in population *k*

$q^{(i)}$: vector of population frequencies in individual *i*'s genotype

$Q = \{q_k^{(i)}\}_{k,i}$

Population of origin is now assigned to each allele of each locus of each individual, instead of to each individual.

previous: $z^{(i)}$ population of origin of individual *i*

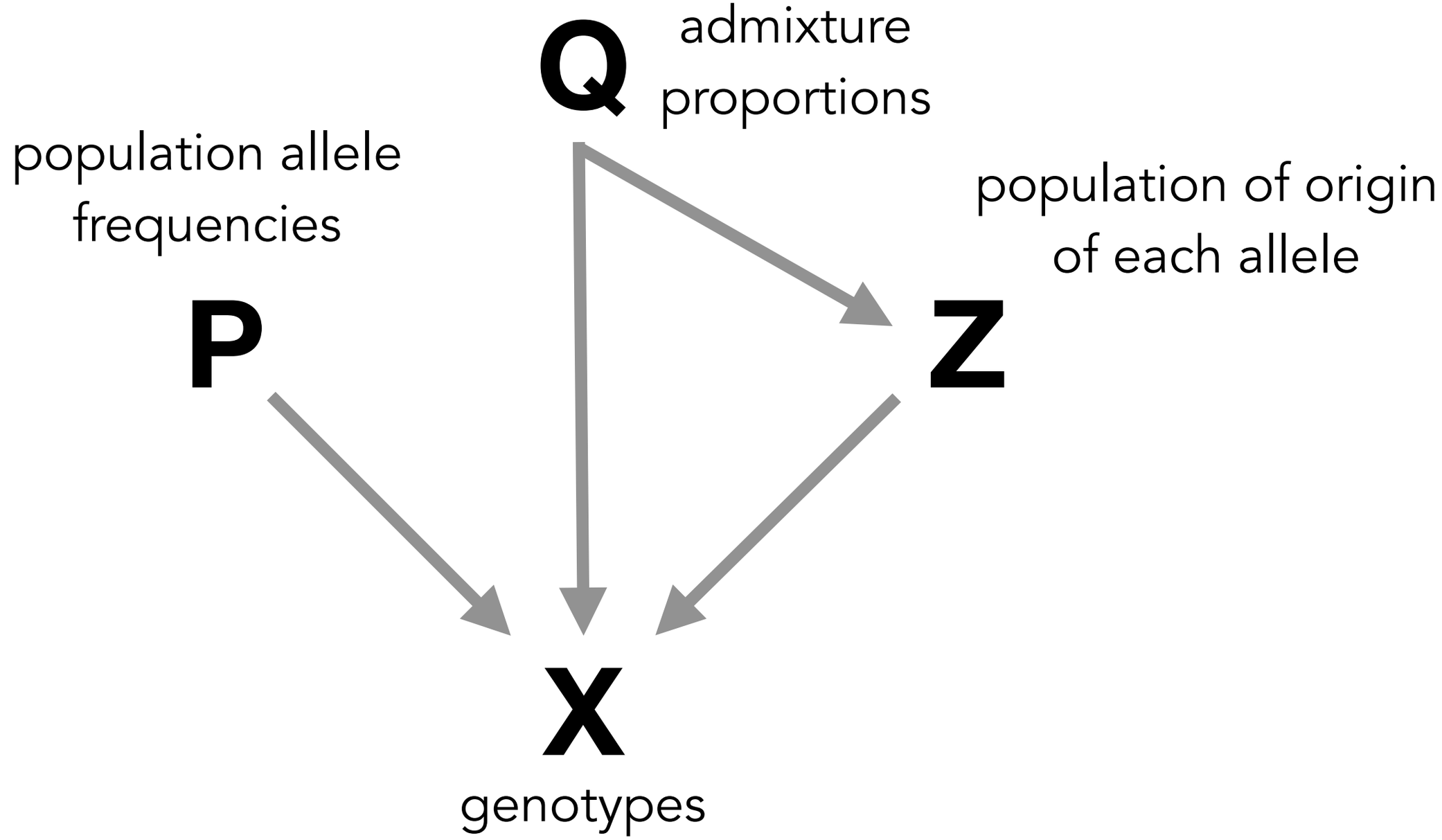
now: $(z_l^{(i,1)}, z_l^{(i,2)})$ population of origin of each allele at locus *l* of individual *i*

$Z = \{(z_l^{(i,1)}, z_l^{(i,2)})\}_{i,l}$

Application: Inferring population structure from genetic data

Introducing admixture…

$$q^{(i)} \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$



Application: Inferring population structure from genetic data

Given X, how do we get* P(*P, Q, Z | X*)*?

STRUCTURE Block Gibbs sampler for $\mathbb{P}(P, Q, Z|X)$

Initialize $P^{(0)}, Z^{(0)}$.

for $t = 1, \ldots$ **do**

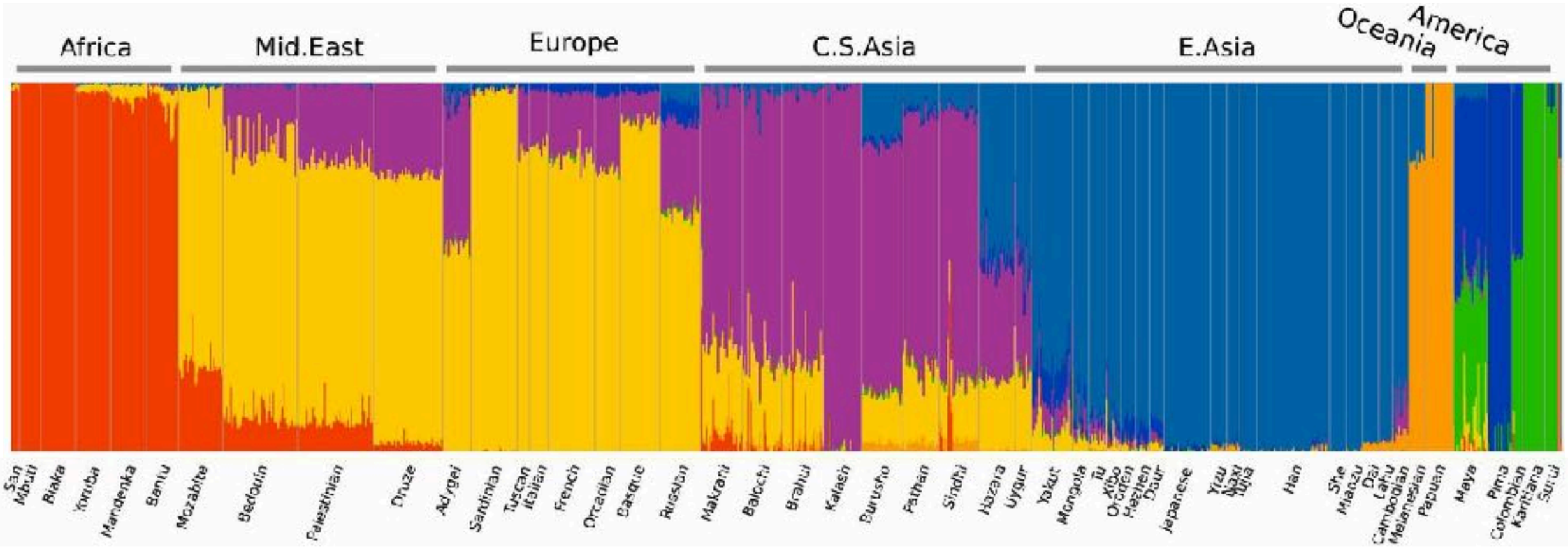
$$P^{(t)}, Q^{(t)} \sim \mathbb{P}(P, Q \mid X, Z = Z^{(t-1)})$$

$$Z^{(t)} \sim \mathbb{P}(Z \mid X, P = P^{(t)}, Q = Q^{(t)})$$

end for

Application: Inferring population structure from genetic data

What does STRUCTURE output?



STRUCTURE became the gold standard in genetics and evolutionary biology for inferring population structure from genetic data.

Application: Inferring population structure from genetic data

However, the slow Gibbs samplers quickly became a bottleneck for massive datasets (100K loci instead of 100s of loci).

Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There may be important population structure that is not well captured by current geographical region of residence. Present implementations of strongly model-based approaches such as STRUCTURE[11,12] are impracticable for data sets of this size, and we reverted to the classical method of principal components[13,14], using a subset of 197,175 SNPs chosen to reduce inter-locus linkage disequi-

nature genetics

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

differentiation, leading to a loss in power. Structured association uses a program such as STRUCTURE[15] to assign the samples to discrete subpopulation clusters and then aggregates evidence of association within each cluster. If fractional membership in more than one cluster is allowed, the method cannot currently be applied to genome-wide association studies because of its intensive computational cost on large data sets. Furthermore, assignments of individuals

OPEN ACCESS Freely available online PLoS GENETICS

Population Structure and Eigenanalysis

Nick Patterson[1*], Alkes L. Price[1,2], David Reich[1,2]

Our implementation of PCA has three major features. 1) It runs extremely quickly on large datasets (within a few hours on datasets with hundreds of thousands of markers and thousands of samples), whereas methods such as STRUCTURE can be impractical. This makes it possible to extract

Solution: Variational Inference

Basic idea behind variational inference

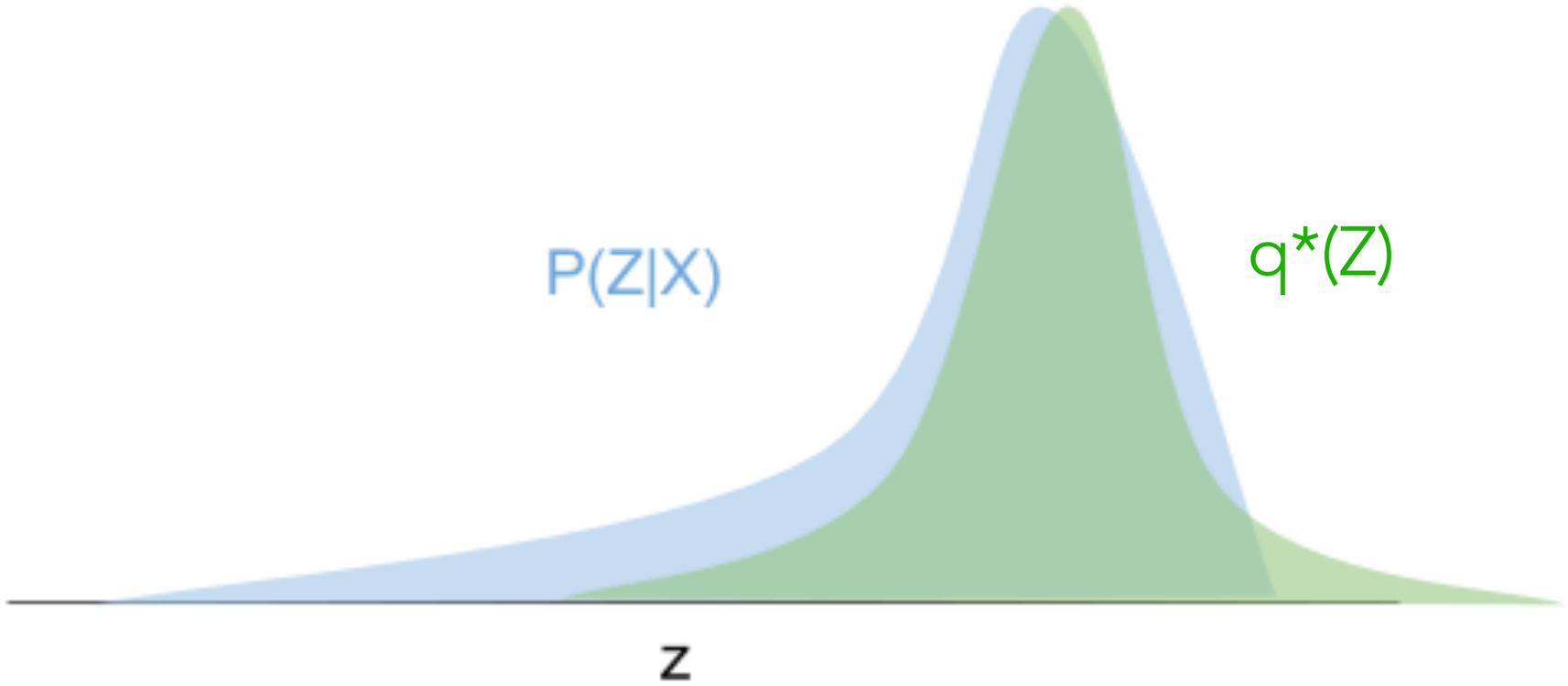
- Sampling from the posterior is difficult. Samplers can take a long time to mix in practice, and it's not always clear how to assess when they have mixed.
- Instead, let's **fit a good approximation of the posterior**.
- The posterior can be arbitrarily complicated. Instead, let's pick **some class Q of simpler distributions** we know how to work with.
- The goal of variational inference is to **find the distribution in Q that is "closest" to the posterior**, called the variational approximation. This is an optimization problem:

$$q^* = \arg\min_{q \in Q} D_{KL}(q(Z) \,||\, \mathbb{P}(Z \mid X))$$

Solution: Variational Inference

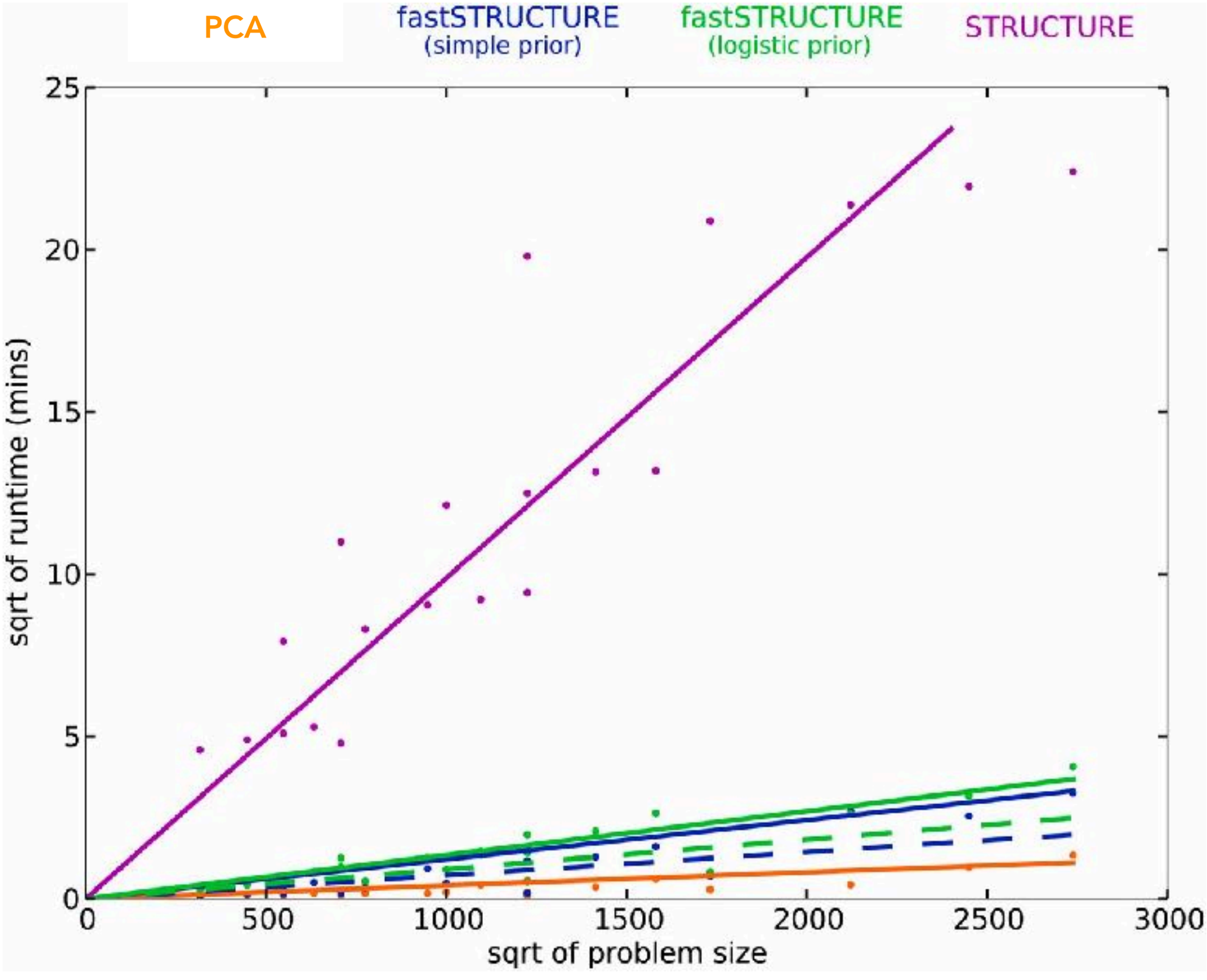
Ingredients to variational inference:

1. Pick class Q of simpler distributions.
2. Solve optimization problem to find best approximation q* in Q.



Revamping STRUCTURE with variational inference = fastSTRUCTURE!

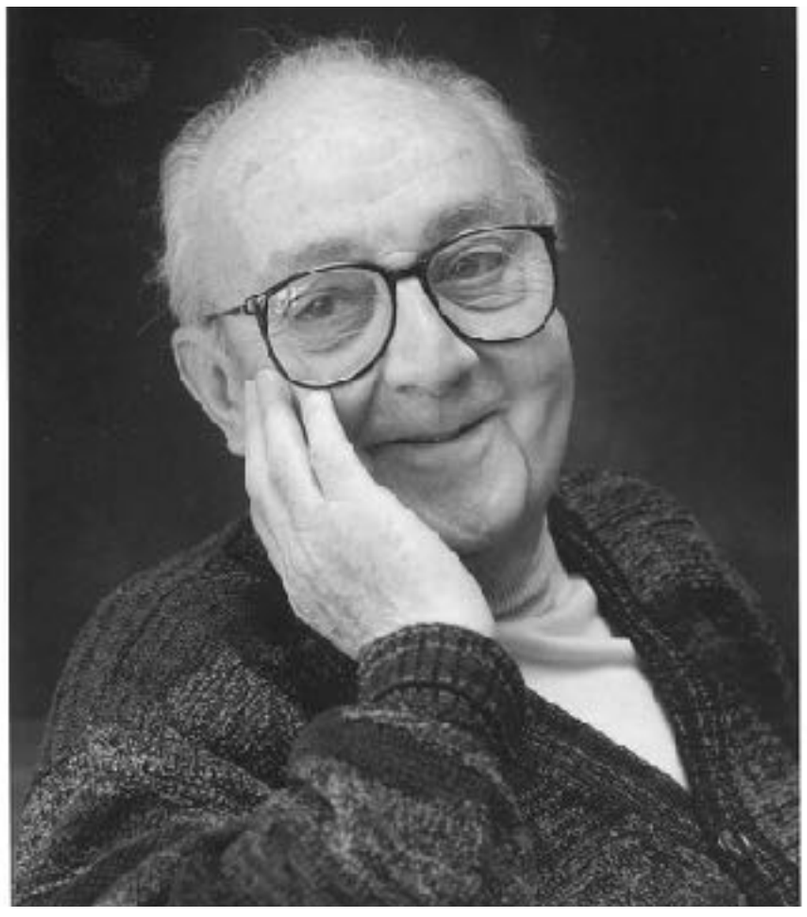
STRUCTURE (sampling-based)
vs. fastSTRUCTURE (variational inference)



Conclusions

Two general approaches for being Bayesian:

1. Sampling (asymptotically correct, but slow)
2. Variational inference (wrong, but fast)



"All models are wrong,
but some are useful."