

# Revue rapide de Statistiques

October 8, 2009

- Qu'est-ce que le MV?
- Méthode de Monte Carlo?
- Le Paradigme Bayésien.
- Melanges de Dirichlet.
- Chaîne de Markov?

	Codon	o/oo	count	p <sub>i</sub>
Proline est CC* (regular expression), il ya quatre facons de l'ecrire:	CCU	3.4	( 4457)	0.059
	CCC	17.0	( 22503)	0.294
	CCA	6.1	( 8085)	0.106
	CCG	31.4	( 41507)	0.542
	Total	57.9	76552	1.001

# Estimation par Maximum de Vraisemblance (MLE)

## Maximum Likelihood Estimation

Definition:

L'estimateur de maximum de vraisemblance de  $\theta$  est la valeur de  $\theta$  qui maximise  $\text{lik}(\theta)$ : c'est la valeur qui rend les données les plus probables.

**Observations Indépendantes:**

$X_1, X_2, X_3, \dots, X_n$  ont une densité conjointe notée

$$f_{\theta}(x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta) = \text{lik}(\theta)$$

Quand les observations sont indépendantes, nous faisons le produit des densités:

$$f(x_1, x_2, \dots, x_n | \theta) = \prod f(x_i | \theta)$$

Plutôt que d'utiliser le produit on utilise la somme en utilisant le fait que

la fn log est croissante

$$l(\theta) = \sum_{i=1}^n \log(f(x_i|\theta))$$

On peut dériver et résoudre:  $l'(\theta) = 0$

# Maximum Likelihood pour les cellules d'une Multinomiale

$X_1, X_2, \dots, X_m$  le nombre de boules tombés dans les boites de 1 à  $m$ . Chaque boite a une probabilité différente (petites ou grandes boites). Nous fixons le nombre total d'objets qui tombe à  $n$ :  $X_1 + X_2 + \dots + X_m = n$ . La probabilité de chaque boite est  $p_i$ , avec la contrainte  $p_1 + p_2 + \dots + p_m = 1$ , donc les  $X_i$ 's ne sont pas indépendantes, la probabilité conjointe multivariée est  $x_1, x_2, \dots, x_m$  s'appelle la multinomiale et prend la forme:

$$f(x_1, x_2, \dots, x_m | p_1, \dots, p_m) = \frac{n!}{\prod x_i!} \prod p_i^{x_i} = \binom{n}{x_1, x_2, x_m} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}$$

Chaque boite est binomiale si on la considère seule en groupant les autres.

Etudiant le log-vraisemblance:

$$l(p_1, p_2, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

Contrainte doit être prise en compte: multiplicateurs de Lagrange

$$L(p_1, p_2, \dots, p_m, \lambda) = l(p_1, p_2, \dots, p_m) + \lambda(1 - \sum_i^m p_i)$$

En forçant tous les dérivés à être égaux à 0, on obtient comme estimateur naturel:

$$\hat{p}_i = \frac{x_i}{n}$$

# Hardy-Weinberg

Trinomiale avec 3 boites, et un parametre libre. Les probabilités des trois types sont de la forme:

$$(1 - \theta)^2 \quad 2\theta(1 - \theta) \quad \theta^2$$

Le dérivé de la vraisemblance:

$$l'(\theta) = -\frac{2X_1 + X_2}{1 - \theta} + \frac{2X_3 + X_2}{\theta}$$

$$l''(\theta) = -\frac{2X_1 + X_2}{(1 - \theta)^2} + \frac{2X_3 + X_2}{\theta^2}$$

$$\hat{\theta}_{ML} = \frac{2x_3 + x_2}{2(x_1 + x_2 + x_3)}$$

Chacun des trois types est binomiales de probabilité respective:

$$E(X_1) = n(1 - \theta)^2$$

$$E(X_2) = 2n\theta(1 - \theta)$$

$$E(X_3) = n\theta^2$$

# Qu'est-ce que le raisonnement Bayésien en Statistiques?

C'est un pas de plus dans la paramétrization. On peut utiliser les modèles probabilistes pour inférer ou bien en utilisant des intervalles de confiance ou des tests d'hypothèses.

La population cible a souvent des aspects connus et des aspects inconnus.

On veut pouvoir attacher à la connaissance a priori une valeur quantitative, comme dans le cas des anniversaires.

C'est rare les situations où on ne connait rien du tout du problème, on sait souvent si on doit utiliser un mètre ou un "pied de coulisse" pour faire une mesure.

Le modèle a priori est souvent très schématique, ça n'a pas d'importance (on l'appelle souvent subjective, car ça peut dépendre de la personne, de son niveau de connaissance, de son expérience personnelle).

Quand l'incertitude du modèle peut se réduire à un paramètre  $\theta$  les Bayésiens traitent  $\theta$  comme si c'était une variable aléatoire  $\Theta$  dont la distribution décrit l'incertitude.

Toute une distribution peut être difficile et dans ce cas on doit utiliser des événements successifs du type  $\Theta \leq \theta$ , et n'a pas besoin d'être trop précis.

La raison la probabilité a priori subjective n'est pas trop importante est que c'est assujéti à modification aussitôt qu'on se retrouve avec de vraies données. Supposons que l'état actuel de nos connaissances se traduit par une densité  $g(\theta)$  et la densité des données est donnée par  $f(x|\theta)$ .

# Theoreme de Bayes

$$P(H|\text{data}) = \frac{P(\text{data}|H)P(H)}{P(\text{data})}$$

La probabilité de H étant des données s'appelle la probabilité à posteriori de H, c'est à dire apres avoir vu les données.

La probabilité inconditionnelle de H :  $P(H)$  est la probabilité a priori de H.

Pour des données observées données la  $P(\text{data}|H)$  est la vraisemblance de H.

$$P(H|\text{data}) \propto P(\text{data}|H)P(H)$$

La probabilité. a posteriori est proportionnelle à la vraisemblance multiplié par l'a priori.

$$\frac{P(H|\text{data})}{P(H^c|\text{data})} \propto \frac{P(\text{data}|H) P(H)}{P(\text{data}|H^c) P(H^c)}$$

Posterior odds = Prior odds times likelihood ratio.

Représenter les données par une variable aléatoire  $Y$ :

$$h(\theta) \propto L(\theta)g(\theta)$$

$L(\theta)$  est proportionnelle à la densité de probabilité de  $Y$  étant donné  $\theta$ . En fait on peut considérer qu'on étudie la distribution conjointe de deux variables  $\Theta$  et  $Y$ .

La distribution marginale de  $Y$  n'est pas évidente c'est dans le facteur de proportionnalité.

$$m(y) = \int f(y|\theta)g(\theta)d\theta$$

Remarque: Ne pas se soucier du choix de probabilité a priori.

Two people with divergent prior opinions but reasonably open-minded will be forced into arbitrarily close agreement about future observations by a sufficient amount of data. We will see an example of this later on.

# About Priors

Les distributions 'molles' reflètent une faiblesse d'information a priori. Quand on est allée dans la lune, avant on ne savait pas l'épaisseur de la poussière. La première approximation peut être rejetée dès qu'on recueille un peu d'information.

Quand l'information a priori est disponible la méthode Bayésienne donne un moyen facile de mettre à jour l'information quand les données nouvelles arrivent.

Plusieurs étapes pour construire la distribution a priori.

# Calibrer les degrés de croyance

Supposez qu'on veuille découvrir votre probabilité que l'adulte male moyen penguin empereur pese plus de 30 kg? On fait des experiences de comparaison

1. Preferez vous parier que A est vrai ou que vous obteniez un jeton R entre 1R/1V? Supposons que vous preferez le dernier?

2. Est-ce que vous preferez parier sur les vert avec 3V/1R ou bien sur A?

....etc... Cela permet de mettre des bornes sur les probabilites inconnues.

Another type of thought experiment could be used to build  $P[\Theta \leq \theta]$  for an increasing sequence of  $\theta$ 's.

This is not usually how priors are built though because it seems quite an exhaustive process to build up a whole density prior, instead we are going to use families of priors who have easy updating processes with regards to the specific likelihoods at hand.

# Distributions Conjuguées

Souvent on peut définir la distribution a priori pour la distribution a posteriori peut etre dans la famille, ca donne une distribution a posteriori facile a definir. C'est une facon objective de construire une distribution a priori.

# Binomial-Beta

**Beta a priori pour le parametre d'une Binomiale** La petite histoire: Laplace et Bayes.

La balle de billiard blanche est roulee sur la table et on regarde la ou elle s'arrete de 0 a 1. On suppose qu'elle ait une distribution uniforme entre 0 et 1, elle s'arrete a un point  $p$ .

La balle rouge est ensuite roulee  $n$  fois avec la meme distribution et  $r$  est le nombre de fois que R va moins loin que W

On voit  $r$  que peut on dire de  $p$ ?

En le disant autrement, on aurait pu avoir rouler les  $n$  balles blanches puis la rouge et regarder le nombre de boules blanches avant la rouge.

ON aurait pu les rouler ensemble et regarder a la probabilité que la rouge soit la aieme avec probabilitite  $1/(n+1)$ .

Let's say this again in our terminology: We are looking for the posterior distribution of  $p$  given  $X$ .

$p$  is a number between 0 and 1

The prior distribution of  $p$  is  $\text{Uniform}(0,1)=\text{Beta}(1,1)$ .

# Beta family

$$f(p|r, s) \propto p^{r-1}(1-p)^{s-1}$$

$$B(r, s) = \int_0^1 p^{r-1}(1-p)^{s-1} dp = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

$$X \sim \mathcal{B}(n, p)$$

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

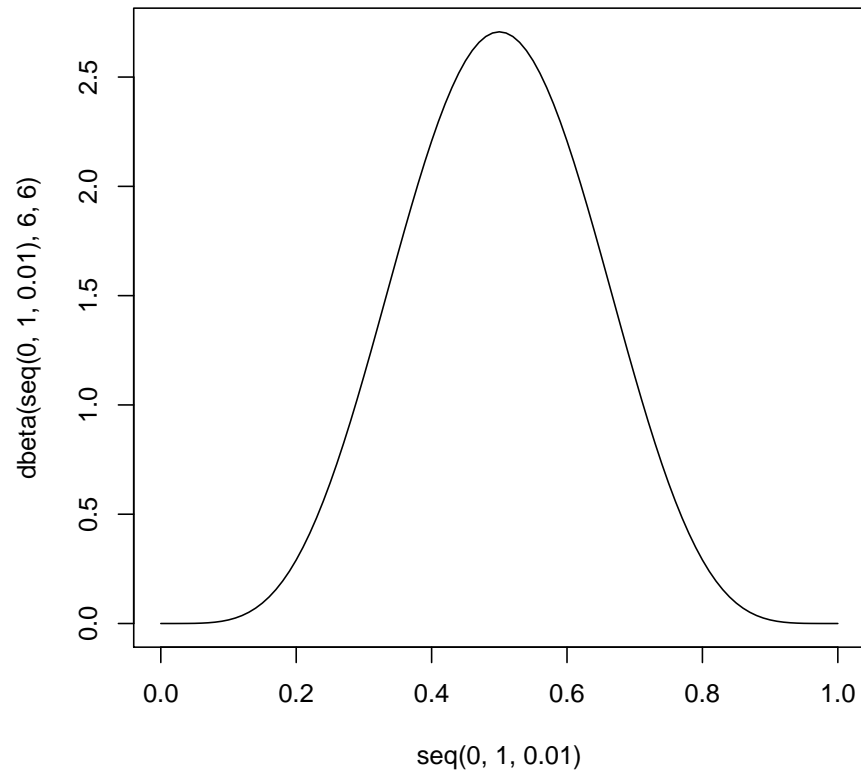
$$P(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$$

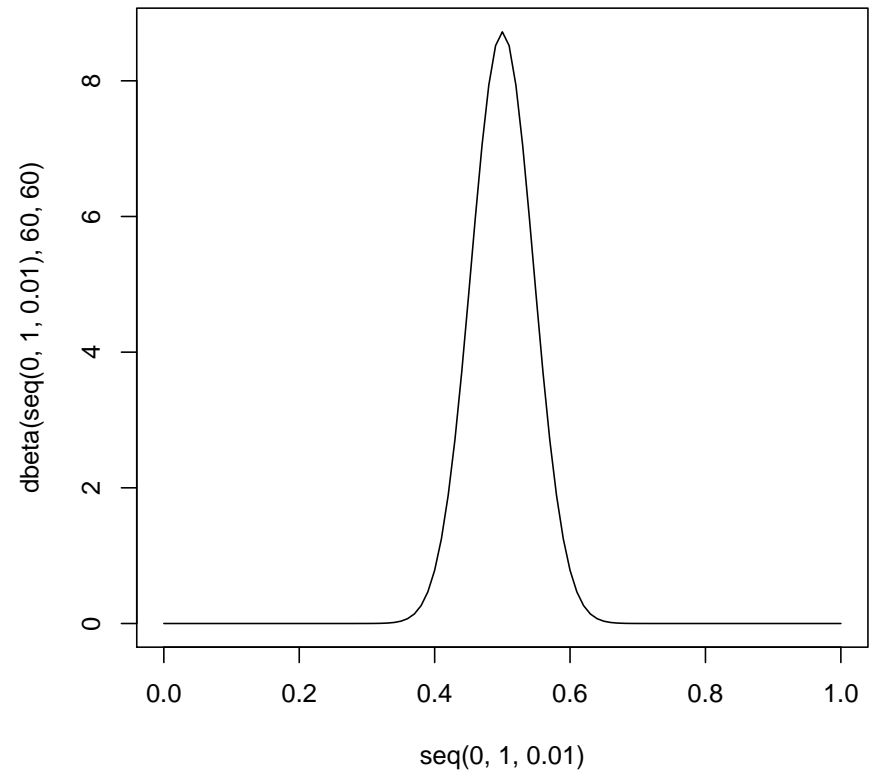
$$P(a < p < b|X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

# The Beta Family

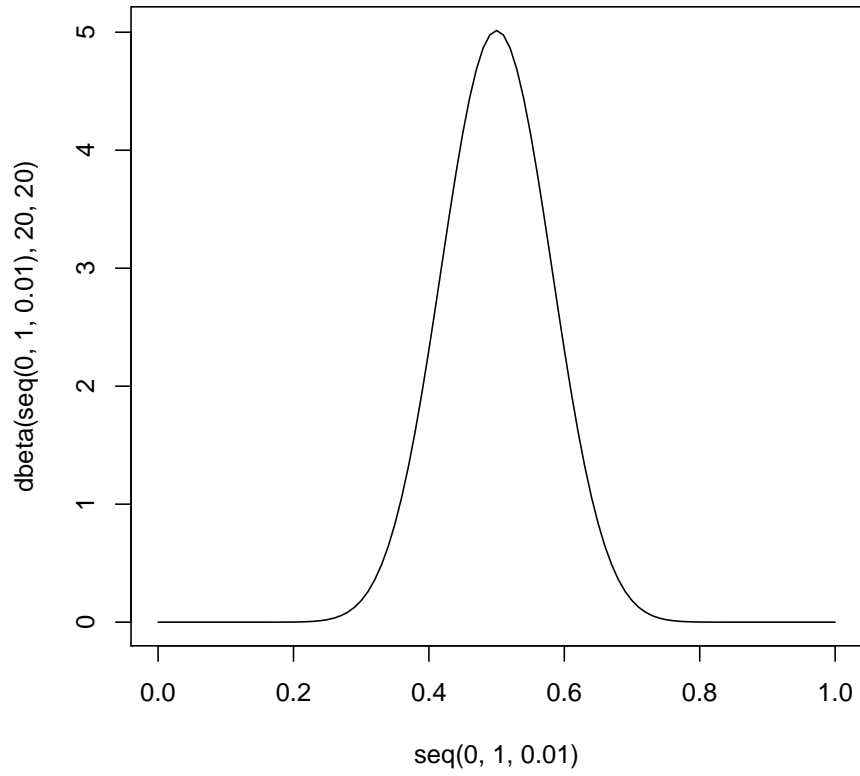
**Beta(6,6)**



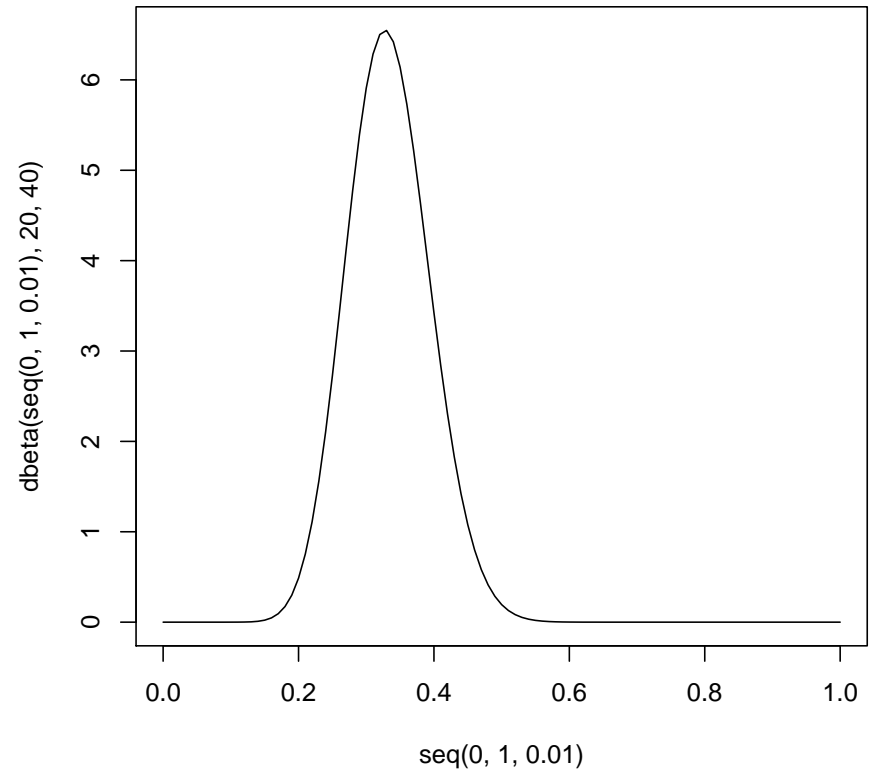
**Beta(60,60)**



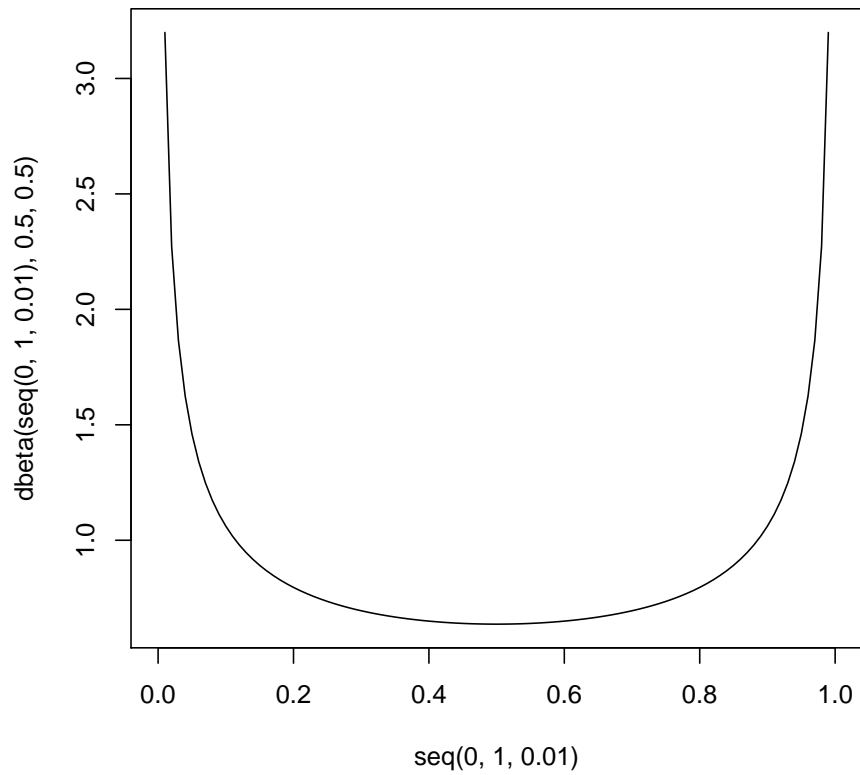
**Beta(20,20)**



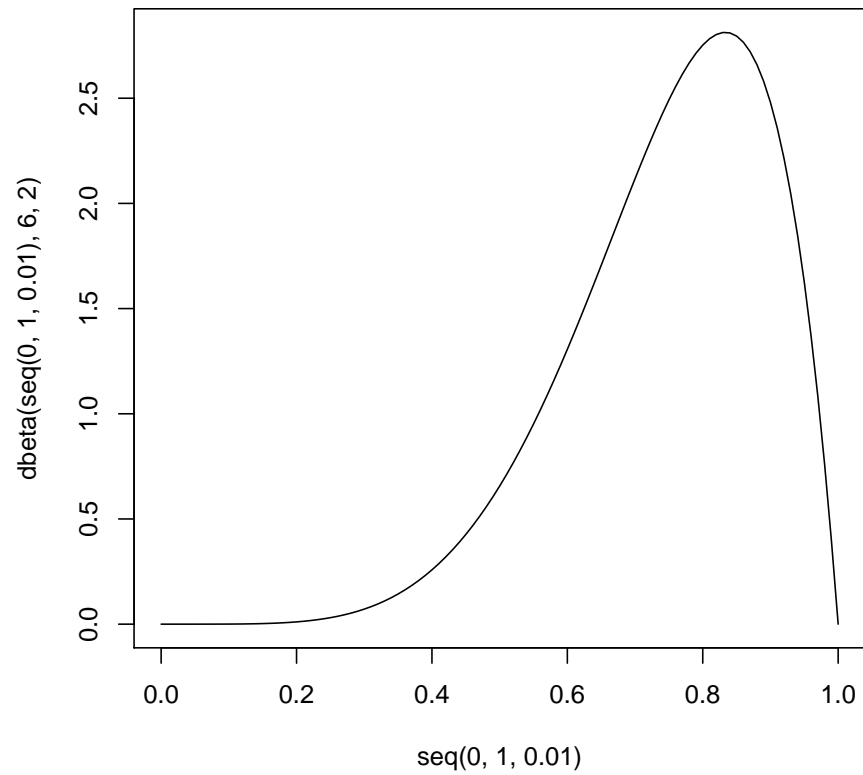
**Beta(20,40)**



**Beta(.5,.5)**



**Beta(6,2)**



# Normal-Normal

La distribution conjuguee de la Normale est normale. La precision est l'inverse de la variance:

$$\xi = \frac{1}{\sigma^2} \text{ and } \xi_0 = \frac{1}{\sigma_0^2}$$

## Normal Conjugates:

Supposons  $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ . La distribution a posteriori distribution de  $\mu$  est normale de moyenne

$$\mu_1 = \frac{\xi_0 \mu_0 + \xi x}{\xi_0 + \xi}$$

and precision

$$\xi_1 = \xi_0 + \xi$$

La moyenne a posterior est la moyenne ponderee de la moyenne a priori et des donnees avec des poids proportionnelles aux precisions.

Avec une distribution a priori tres molle, on aura une precision basse  $\xi_0$ , une a priori tres plate et la moyenne sera surtout celle des donnees.

Of course what we are usually interested in is the posterior given an iid sample of size  $n$ , what you could expect happens it is equivalent to adding one observation  $\bar{x}$  from a distribution that has variance  $\sigma^2/n$ .

# Multinomial-Dirichlet

You are given a set  $\mathcal{X}$  (here taken as finite) and a probability density  $p(x)$ , ( $p(x) \geq 0, \sum p(x) = 1$ ). Also given is a set  $A$  in  $\mathcal{X}$ . The problem is to compute or approximate  $p(A)$ .

In order to go further we need to extend what we did before for the binomial and its Conjugate Prior to the multinomial and the the Dirichlet Prior. This is a probability distribution on the  $n$  simplex

$$\Delta_n = \{ \tilde{p} = (p_1, \dots, p_n), p_1 + \dots + p_n = 1, p_i \geq 0 \}$$

It is a  $n$ -dimensional version of the beta density. The Dirichlet has a parameter vector:  $\tilde{\mathbf{a}} = (a_1, \dots, a_n)$ . Throughout we write  $A = a_1 + \dots + a_n$ .

$\Delta_n$  is normalised to have total mass 1 the Dirichlet has density:

$$D_{\tilde{\mathbf{a}}}(\tilde{\mathbf{x}}) = \frac{\Gamma(A)}{\prod \Gamma(a_i)} x_1^{a_1-1} x_2^{a_2-1} \dots x_n^{a_n-1}$$

The uniform distribution on  $\Delta_n$  results from choosing all  $a_i = 1$ . The multinomial distribution corresponding to  $k$  balls dropped into  $n$  boxes with fixed probability  $(p_1, \dots, p_n)$  (with the  $i$ th box containing  $k_i$  balls) is

$$\binom{k}{k_1 \dots k_n} p_1^{k_1} \dots p_n^{k_n}$$

If this is averaged with respect to  $D_{\tilde{\mathbf{a}}}$  one gets the marginal (or Dirichlet/Multinomial):

$$P(k_1, \dots, k_n) = \frac{(a_1)_{(k_1)} (a_2)_{(k_2)} \dots (a_n)_{(k_n)}}{A_{(k)}}$$

where  $m_{(j)} \stackrel{\text{def}}{=} m(m+1)\cdots(m+(j-1))$

From a more practical point of view there are two simple procedures worth recalling here:

- To pick  $\tilde{p}$  from a Dirichlet prior; just pick  $X_1, X_2, \dots, X_n$  independent from gamma densities

$$\frac{e^{-x} x^{a_i-1}}{\Gamma(a_i)} \text{ and set } p_i = \frac{X_i}{X_1 + \cdots + X_n}, 1 \leq i \leq N$$

- To generate sequential samples from the marginal distribution use **Polya's Urn**:

Consider an urn containing  $a_i$  balls of color  $i$  (actually fractions are allowed).

Each time, choose a color  $i$  with probability proportional to the number of balls of that color in the urn. If  $i$  is drawn, replace it along with another ball of the same color.

The Dirichlet is a convenient prior because the posterior for  $\tilde{p}$  having observed  $(k_1, \dots, k_n)$  is Dirichlet with probability  $(a_1 + k_1, \dots, a_n + k_n)$ . An important characterization of the Dirichlet: it is the only prior that predicts outcomes linearly in the past. One frequently used special case is the symmetric Dirichlet when all  $a_i = c > 0$ . We denote this prior as  $D_c$ .

(the  $a_i$  are often called pseudocounts, and help to get around the paradox that that ML method leads to for “unseen species”)

Reading: See Chapter 11 of Durbin et al.

# Outside of Conjugate Families

'Non parametric' methods, high dimensional methods. Modern Bayesians all use 'The Gibbs sampler' we will do a special example, see Chapter 5 of the reader and lectures next week.

Criteria for building good families: They should generalize well.

Use weights based on "information/entropy" ( $H = E(-\log p)$ ). The goal is to maximize entropy of training sequence, so that the statistical spread of the model should be as broad as possible, thus minimize overfitting.

# Dirichlet Mixtures

$$P(\mathbf{p}|\alpha^1, \dots, \alpha^m) = \sum_k^m q_k \mathcal{D}(\mathbf{p}|\alpha^k)$$

$q_k$  is the prior of the  $k$ th component of the mixture.

Suppose we observe a vector of counts  $v_1, \dots, v_n$  for an  $n$  dimensional problem.

The posterior probabilities are then

$$P(\mathbf{p}|\mathbf{v}) = \sum_k^m q_k P(\mathbf{p}|\alpha^k, \mathbf{v}) P(\alpha^k|\mathbf{v}) = \sum_k^m P(\alpha^k|\mathbf{v}) \mathcal{D}(\mathbf{p}|\mathbf{v} + \alpha^k)$$

Dirichlet mixtures are dense in the space of distributions on probability vectors.

# Markov Chains

**Most important type of dependency**

# Markov Chains

$$\begin{array}{l} \text{Dry} \\ \text{Wet} \end{array} \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

Product of two markov chain matrices:

$$\mathbf{PP} = \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix} \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix} = \begin{bmatrix} 0.44 & 0.56 \\ 0.28 & 0.72 \end{bmatrix}$$

Notice that the probability  $P(X_2 = W | X_0 = W)$  is exactly the formula for the  $(W, W)$  entry in  $\mathbf{PP}$

General case. Define

$$P_{i,j}^{(n)} = P(X_n = j | X_0 = i)$$

Then

$$P(X_{m+n} = j | X_m = i, X_{m-1} = i_{m-1}, \dots) = P(X_{m+n} = j | X_m = i)$$

$$= P(X_n = j | X_0 = i) = P_{i,j}^{(n)}$$

Assume  $P(Y = y | X = x, U = u, V = v) = P(Y = y | X = x)$  for any  $x, y, u, v$ .

$$\text{Then } P(Y = y | X = x, U = u) = P(Y = y | X = x)$$

In words, if knowing both  $U$  and  $V$  doesn't change the conditional probability then knowing  $U$  alone doesn't change the conditional probability.

**Proof of claim: Take  $A = \{X = x, U = u\}$ , then**

$$\begin{aligned} P(Y = y \mid X = x, U = u) &= \frac{P(Y = y, A)}{P(A)} \\ &= \frac{\sum_v P(Y = y, A, V = v)}{P(A)} \\ &= \frac{\sum_v P(Y = y \mid A, V = v) P(A, V = v)}{P(A)} \\ &= \frac{\sum_v P(Y = y \mid X = x) P(A, V = v)}{P(A)} \\ &= \frac{P(Y = y \mid X = x) \sum_v P(A, V = v)}{P(A)} \\ &= \frac{P(Y = y \mid X = x) P(A)}{P(A)} \\ &= P(Y = y \mid X = x) \end{aligned}$$

Second step: consider

$$\begin{aligned} P(X_{n+2} = k | X_n = i) &= \sum_j P(X_{n+2} = k, X_{n+1} = j | X_n = i) \\ &= \sum_j P(X_{n+2} = k | X_{n+1} = j, X_n = i) \\ &\quad \times P(X_{n+1} = j | X_n = i) \\ &= \sum_j \mathbf{P}_{i,j} \mathbf{P}_{j,k} \end{aligned}$$

This shows that  $P(X_{n+2} = k | X_n = i) = (\mathbf{P}^2)_{i,k}$  where  $\mathbf{P}^2$  means the matrix product  $\mathbf{P}\mathbf{P}$ . Notice both that the quantity does not depend on  $n$  and that we can compute it by taking a power of  $\mathbf{P}$ . More general version  $P(X_{n+m} = k | X_n = j) = (\mathbf{P}^m)_{j,k}$ . Since  $\mathbf{P}^n \mathbf{P}^m = \mathbf{P}^{n+m}$  we get the

Chapman-Kolmogorov equations:

$$P(X_{n+m} = k | X_0 = i) = \sum_j P(X_{n+m} = k | X_n = j) P(X_n = j | X_0 = i) \quad (1)$$

**Summary: A Markov Chain has  $n$  step transition probabilities which are the  $n$  th power of the 1 step transition probabilities.**

Here is R output for the 1,2,4,8 and 16 step transition matrices for our weather example:

```
> P%*%P
      [,1] [,2]
[1,] 0.44 0.56
[2,] 0.28 0.72
> P2= P%*%P
> P4= P2%*%P2
> P8= P4%*%P4
> P16= P8%*%P8
> P16
```

```
      [,1]      [,2]
[1,] 0.3333336 0.6666664
[2,] 0.3333332 0.6666668
```

This computes the powers of P.

Fact:

$$\lim_{n \rightarrow \infty} \mathbf{P}^n = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} .$$

Suppose we toss a coin  $P(H) = \alpha_D$  and start the chain with Dry if we get heads and Wet if we get tails. Then

$$P(X_0 = x) = \begin{cases} \alpha_D & x = \text{Dry} \\ \alpha_W = 1 - \alpha_D & x = \text{Wet} \end{cases}$$

and

$$P(X_1 = x) = \sum_y P(X_1 = x | X_0 = y) P(X_0 = y) = \sum_y \alpha_y P_{y,x}. \quad (2)$$

Notice the last line is a matrix multiplication of the row vector  $\alpha$  by matrix multiplication of the row vector  $\alpha$  by matrix  $\mathbf{P}$ . A special  $\alpha$ : if we put  $\alpha_D = 1/3$  and  $\alpha_W = 2/3$  then

$$\begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 3 \end{bmatrix}.$$

$P(X_0 = D) = 1/3$  then  $P(X_1 = D) = 1/3$  and analogously for  $W$ . This means that  $X_0$  and  $X_1$  have the same distribution.

A probability vector  $\alpha$  is called an initial distribution for the chain if  $P(X_0 = i) = \alpha_i$ .

A Markov Chain is **stationary** if  $P(X_1 = i) = P(X_0 = i)$  for all  $i$ .

An initial distribution is called *stationary* if the chain is stationary. We find that  $\alpha$  is a stationary initial distribution if  $\alpha\mathbf{P} = \alpha$ .

Suppose  $\mathbf{P}^n$  converges to some matrix  $\mathbf{P}^\infty$ . Notice that  $\lim_{n \rightarrow \infty} \mathbf{P}^{n-1} = \mathbf{P}^\infty$  and

$$\mathbf{P}^\infty = \lim \mathbf{P}^n = [\lim \mathbf{P}^{n-1}] \mathbf{P} = \mathbf{P}^\infty \mathbf{P}.$$

This proves that each row  $\alpha$  of  $\mathbf{P}^\infty$  satisfies  $\alpha = \alpha\mathbf{P}$ .

Definition: A row vector  $x$  is a left eigenvector of  $A$  with eigenvalue  $\lambda$  if  $xA = \lambda x$ .

So each row of  $\mathbf{P}^\infty$  is a left eigenvector of  $\mathbf{P}$  with eigenvalue 1.