Outline

Limitations of Maximum Likelihood

- Given a probabilistic model

$$P(x, y = k) = \pi_k f_k(x),$$

  we typically assume a parametric form for $f_k(x) = f(x|\, \phi_k)$ and compute the MLE $\widehat{\theta}$ of $\theta = (\pi_k, \phi_k)_{k=1}^n$ based on the training data $\{X_i, Y_i\}_{i=1}^n$.
- We then use a plug-in approach to perform classification

$$P\left(y = k|\, x, \widehat{\theta}\right) = \frac{\widehat{\pi}_k f\left(x|\, \widehat{\phi}_k\right)}{\sum_{j=1}^K \widehat{\pi}_j f\left(x|\, \widehat{\phi}_j\right)}.$$

Limitations of Maximum Likelihood

- Even for simple models, this can prove difficult; e.g. if $f(x|\,\phi_k) = \mathcal{N}(x; \mu_k, \Sigma)$ then the MLE estimate of Σ is not full rank for $p > n$.
- One possibility is to simplify even further the model as in Naïve Bayes; e.g.

$$f(x|\,\phi_k) = \prod_{l=1}^{p} \mathcal{N}\left(x^l; \mu_k^l, \left(\sigma_k^l\right)^2\right)$$

  but this might be too crude.
- Moreover, the plug-in approach does not take into account the uncertainty about the parameter estimate.

A Toy Example

- Consider a trivial case where $X \in \{0, 1\}$ and $K = 2$ so that
$$f\left(x|\,\phi_k\right) = \phi_k^x \left(1 - \phi_k\right)^{1-x}.$$
then the MLE estimates are given by
$$\widehat{\phi}_k = \frac{\sum_{i=1}^n \mathbb{I}\left(x_i = 1, y_i = k\right)}{n_k}, \ \widehat{\pi}_k = \frac{n_k}{n}$$
where $n_k = \sum_{i=1}^n \mathbb{I}\left(y_i = k\right)$.
- Assume that all the training data for class 1 are such that $x_i = 0$ then $\widehat{\phi}_1 = 0$ and
$$\begin{aligned} P\left(y = 1|\,x = 1, \widehat{\theta}\right) &= \frac{P\left(x = 1|\,y = 1, \widehat{\theta}\right) P\left(y = 1|\,\widehat{\theta}\right)}{P\left(y = 1|\,\widehat{\theta}\right)} \\ &= \frac{\widehat{\phi}_1 \widehat{\pi}_1}{P\left(y = 1|\,\widehat{\theta}\right)} = 0. \end{aligned}$$
- Hence if we have not observed such events in our training set, we predict that we will never observe them, ever!

Text Classification

- Assume we are interested in classifying documents; e.g. scientific articles or emails.
- A basic but standard model for text classification consists of considering a pre-specified dictionary of p words (including say physics, calculus.... or dollars, sex etc.) and summarizing each document by $X = \left(X^1, ..., X^p\right)$ where

$$X^l = \begin{cases} 1 & \text{if word } l \text{ is present in document} \\ 0 & \text{otherwise.} \end{cases}$$

- To implement a probabilistic classifier, we need to model $f_k(x)$ for $k = 1, ..., K$.
- A Naive Bayes approach ignores features correlations and assumes $f_k(x) = f(x|\,\phi_k)$ where

$$f(x|\,\phi_k) = \prod_{l=1}^{p} \left(\phi_k^l\right)^{x^l} \left(1 - \phi_k^l\right)^{1-x^l}$$

Maximum Likelihood for Text Classification

- Given training data, the MLE is easily obtained

$$\widehat{\pi}_k = \frac{n_k}{n}, \ \widehat{\phi}_k^l = \frac{\sum_{i=1}^n \mathbb{I}\left(X_i^l = 1, Y_i = k\right)}{n_k}$$

- If word l never appears in the training data for class k then $\widehat{\phi}_k^l = 0$ and

$$P\left(y = k | \, x = \left(x^{1:l-1}, x^l = 1, x^{l+1:p}\right), \widehat{\theta}\right) = 0;$$

  i.e. we will never attribute a new document containing word l to class k.
- In many practical applications, we have $p \gg n$ and this problem often occurs.

A Bayesian Approach

- An elegant way to deal with the problem consists of using a Bayesian approach.
- We start with the very simple case where

$$f(x|\phi) = \phi^{x}(1-\phi)^{1-x}$$

and now set a Beta prior on $p(\phi)$ on ϕ

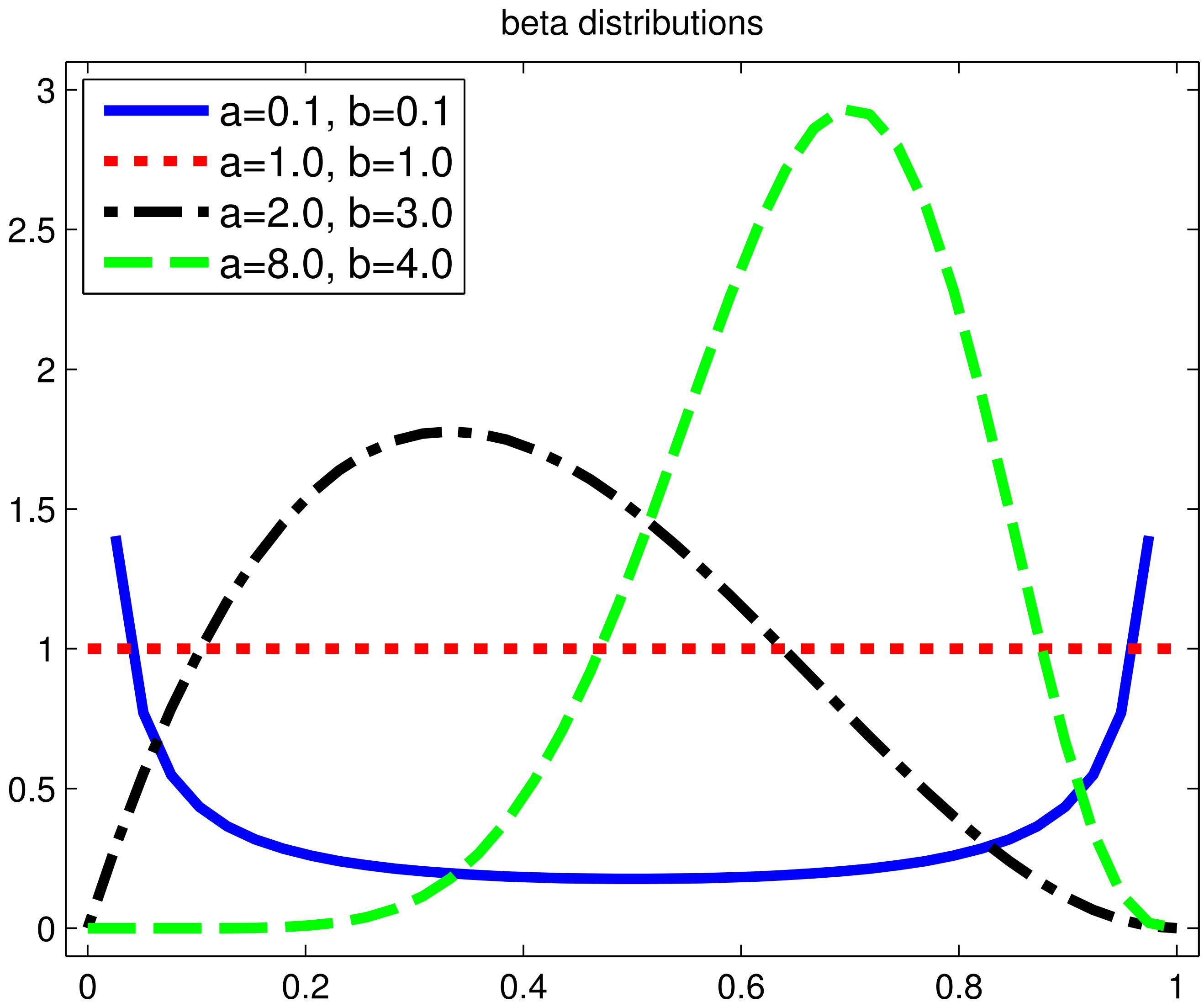
$$p(\phi) = Beta(\phi; a, b)$$

where

$$Beta(\phi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\phi^{a-1}(1-\phi)^{b-1}1_{[0,1]}(\phi)$$

with $\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}dt$. Note that $\Gamma(u) = (u-1)!$ for $u \in \mathbb{N}$.
(a, b) are *fixed* quantities called *hyperparameters*. For $a = b = 1$, the Beta density corresponds to the uniform density.

Beta Distribution



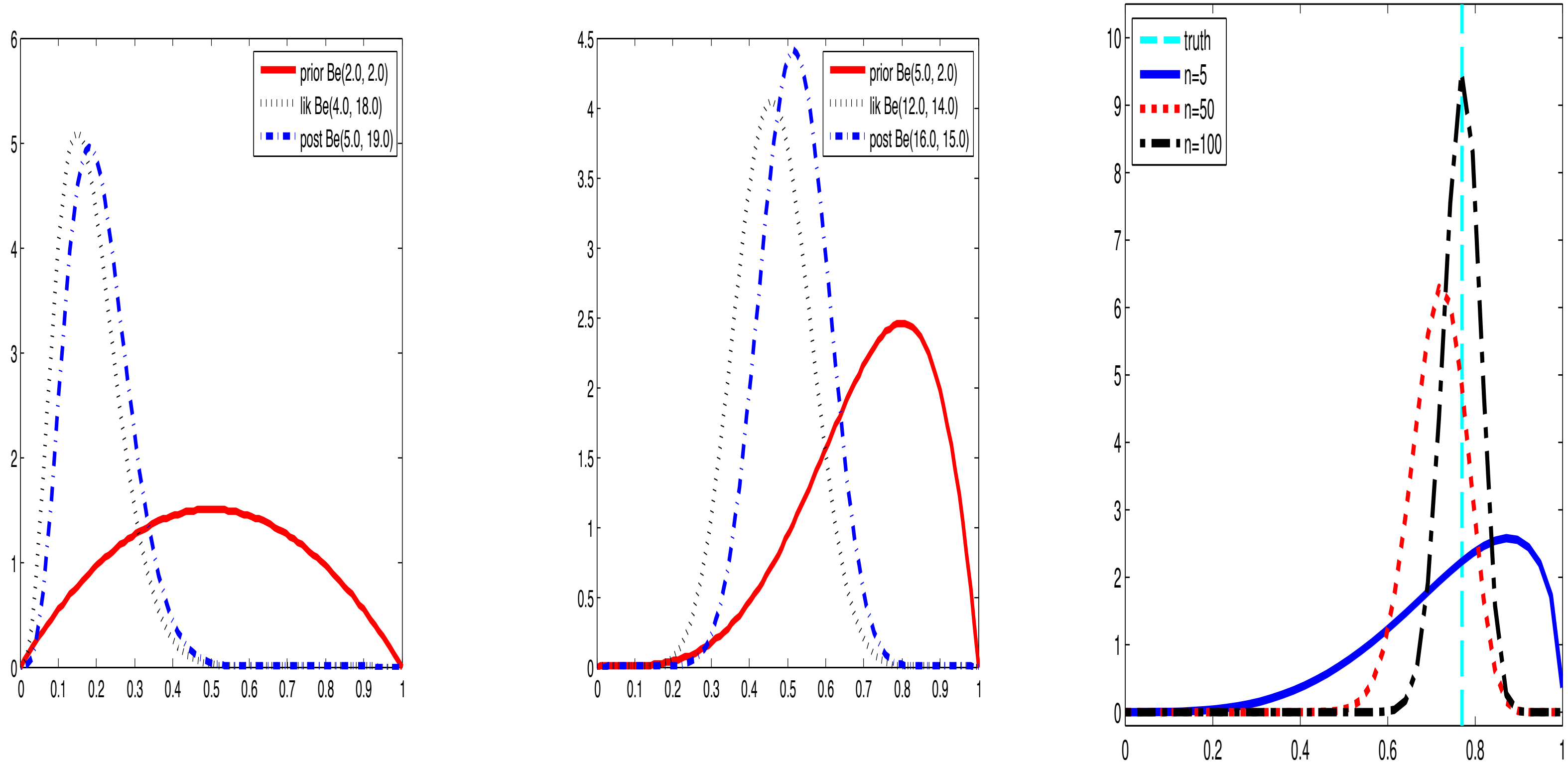
A Bayesian Approach

- Given a realization of $X_{1:n} = (X_1, ..., X_n)$, inference on ϕ is based on the posterior

$$
\begin{aligned}
p(\phi | x_{1:n}) &= \frac{p(\phi) \prod_{i=1}^{n} f(x_i | \phi)}{\pi(x_{1:n})} \\
&= Beta(\theta; a + n_s, b + n - n_s)
\end{aligned}
$$

  with $n_s = \sum_{i=1}^{n} \mathbb{I}(x_i = 1)$.
- The prior on θ can be conveniently reinterpreted as an imaginary initial sample of size $(a + b)$ with a observations "1" and b observations "0". Provided that $(a + b)$ is small with respect to n, the information carried by the data is prominent.

Beta Posteriors



(left) Updating a Beta(2,2) prior with a Binomial likelihood with $n_s = 3$, $n = 20$ to yield a Beta(5,19); (center) Updating a Beta(5,2) prior with a Binomial likelihood with $n_s = 11$, $n = 24$ to yield a Beta(16,15) posterior. (right) Sequentially updating a Beta distribution starting with a Beta(1,1) and converging to a delta function centered on the true value.

Posterior Statistics

- We have

$$\mathbb{E}\left(\phi|\, x_{1:n}\right) = \frac{a+n_s}{a+b+n}$$

  and the posterior means behave asymptotically like n_s/n (the ‘frequentist’ estimator) and converge to ϕ^*, the ‘true’ value of ϕ.

- We have

$$\begin{aligned}\mathbb{V}\left(\phi|\, x_{1:n}\right) &= \frac{(a+n_s)(b+n-n_s)}{(a+b+n)^2(a+b+n+1)} \\ &\approx \frac{\widehat{\phi}\left(1-\widehat{\phi}\right)}{n} \text{ for large } n\end{aligned}$$

- The posterior variance decreases to zero as $n \to \infty$, at rate n^{-1}: the information you get on ϕ gets more and more precise.
- For n large enough, the prior is washed out by the data. For a small n, its influence can be significant.

Prediction Plug in vs Bayesian Approaches

- Assume you have observed $X_1 = \cdots = X_n = 0$, then the plug-in prediction is

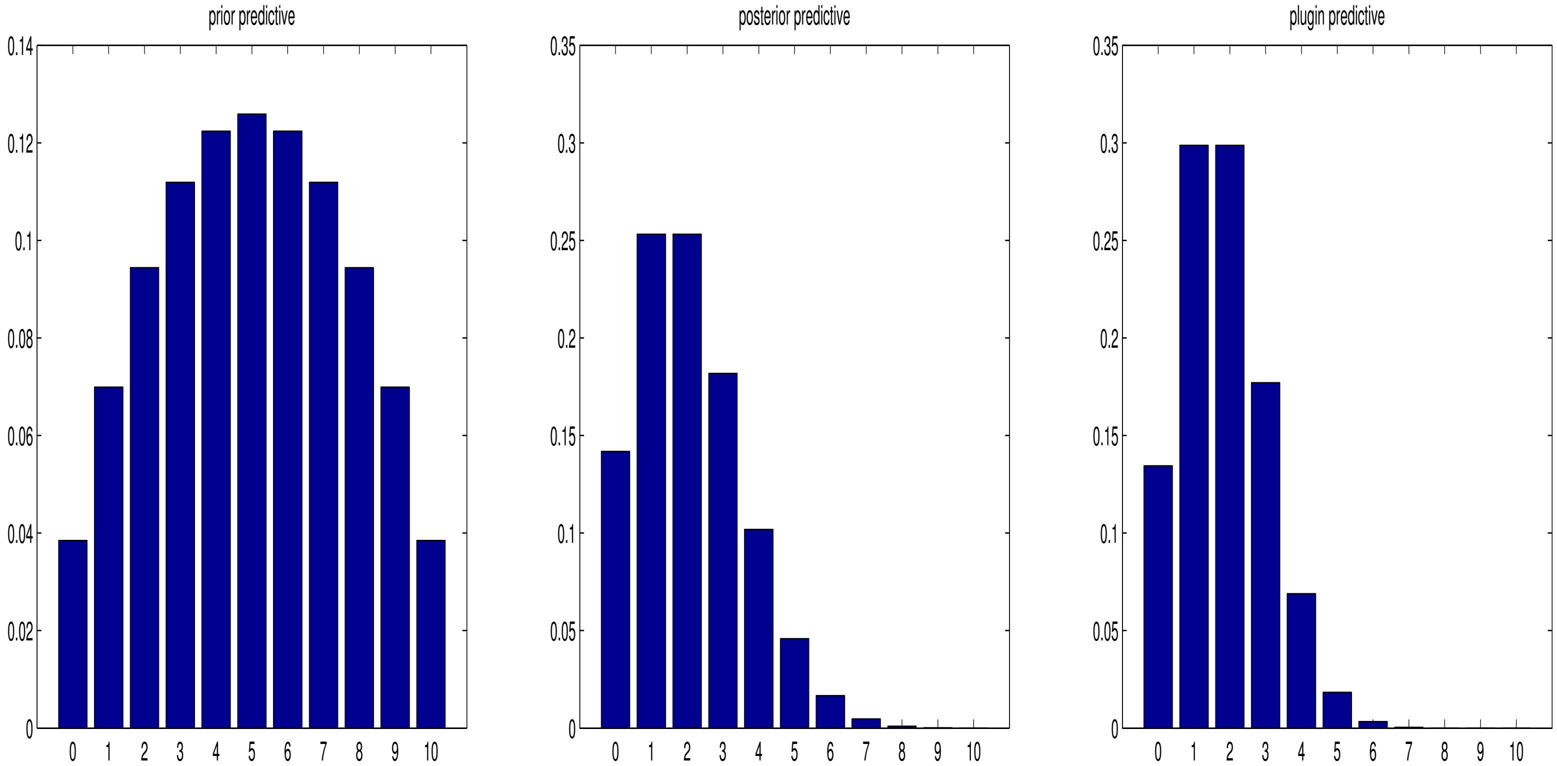
$$P\left(x = 1 | \widehat{\phi}\right) = \widehat{\phi}$$

  which does not account whatsoever for the uncertainty about ϕ.
- In a Bayesian approach, we will use the predictive distribution

$$\begin{aligned} P\left(x = 1 | x_{1:n}\right) &= \int P\left(x = 1 | \phi\right) p\left(\phi | x_{1:n}\right) d\phi \\ &= \frac{a + n_s}{a + b + n} \end{aligned}$$

  so even if $n_s = 0$ then $P\left(x = 1 | x_{1:n}\right) > 0$ and our prediction takes into account the uncertainty about ϕ.

Beta Posteriors



(left) Prior predictive dist. for a Binomial likelihood with $n = 10$ and a Beta(2,2) prior. (center) Posterior predictive after having seen $n_s = 3, n = 20$. (right) Plug-in approximation using $\widehat{\phi}$.

Bayesian Inference for the Multinomial

- Assume we have $Y_{1:n} = (Y_1, ..., Y_n)$ where $Y_i = \left(Y_i^1, ..., Y_i^K\right) \in \{0, 1\}^K$, $\sum_{k=1}^K Y_i^k = 1$ and

$$P(y|\,\pi) = \prod_{k=1}^{K} \pi_k^{y^k}$$

for $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$.

- We have seen that the MLE estimate is

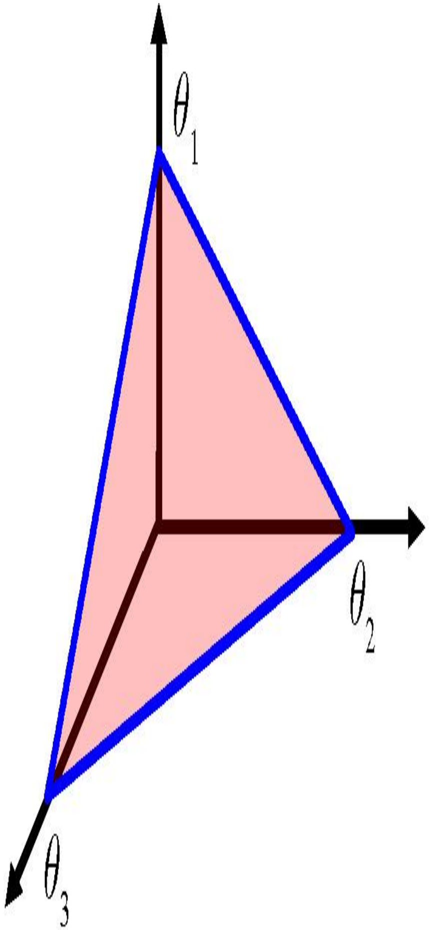
$$\widehat{\pi}_k = \frac{\sum_{i=1}^n \mathbb{I}\left(y_i^k = 1\right)}{n} = \frac{n_k}{n}$$

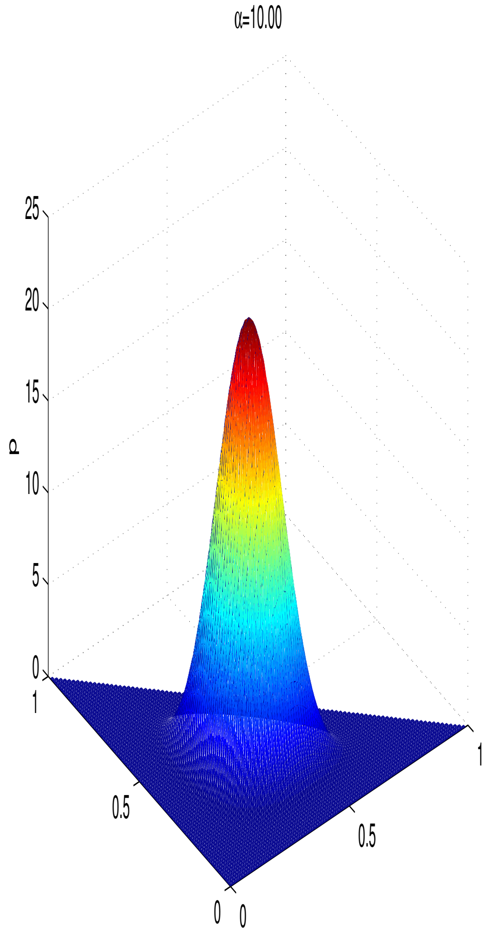
- We introduce the Dirichlet density

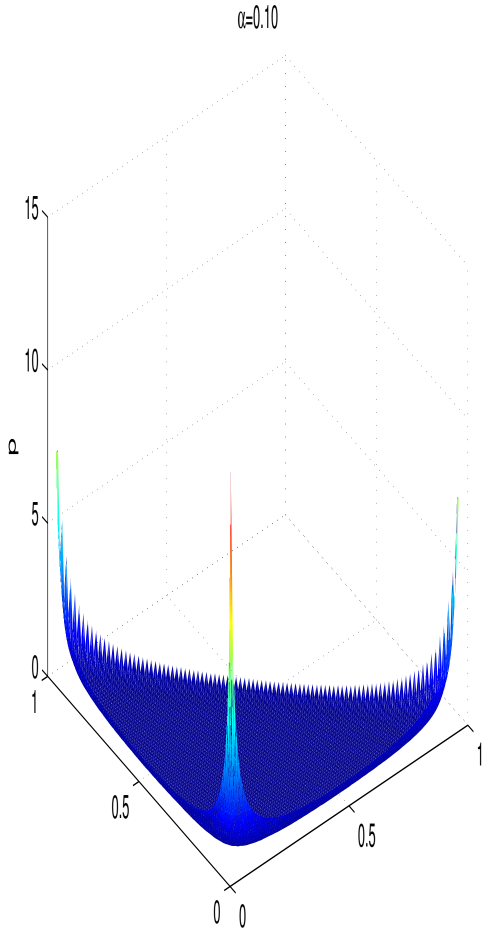
$$p(\pi) = \text{Dir}(\pi; \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

for $\alpha_k > 0$ defined on $\left\{\pi : \pi_k > 0 \text{ and } \sum_{k=1}^K \pi_k = 1\right\}$.

Dirichlet Distributions

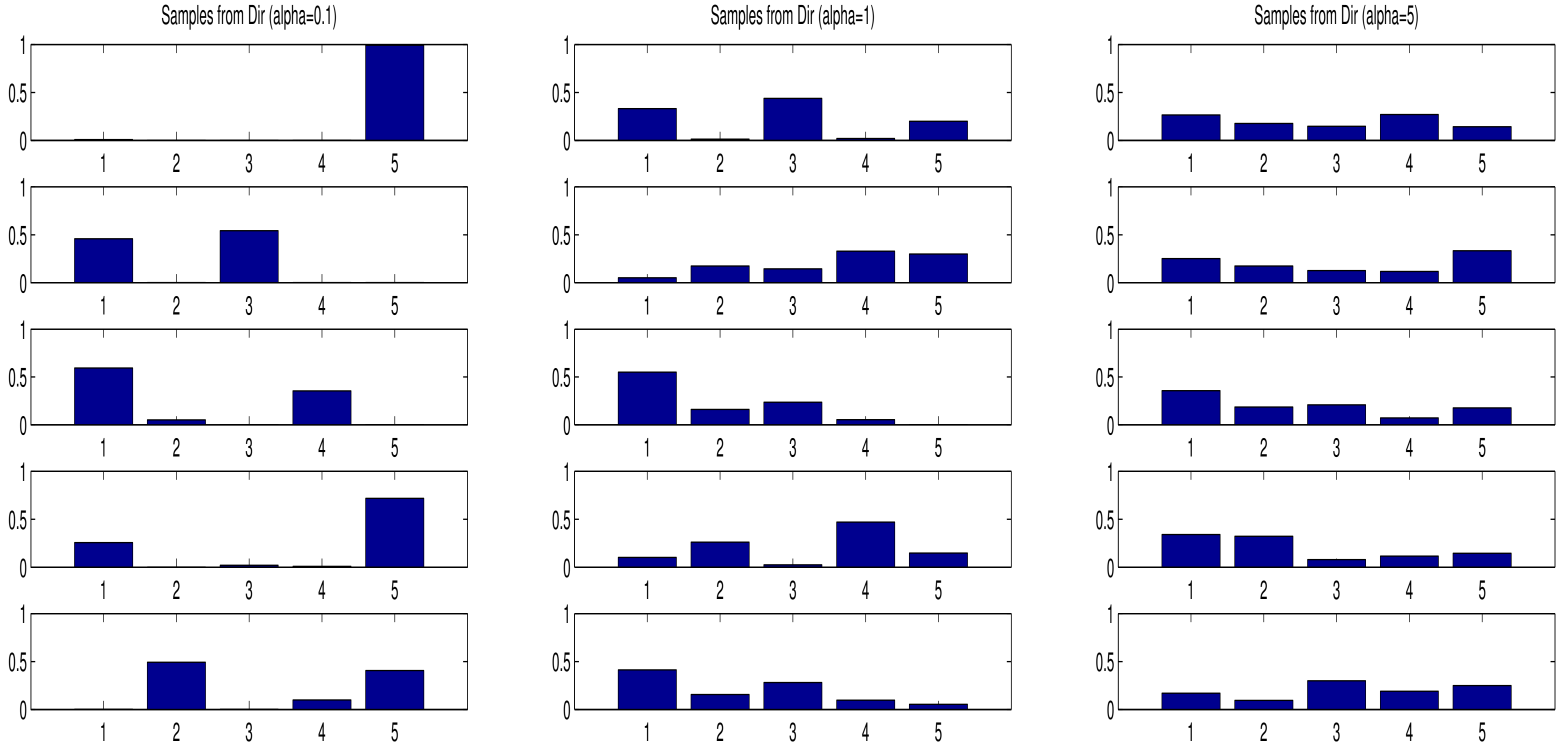






(left) Support of the Dirichlet density for $K = 3$ (center) Dirichlet density for $\alpha_k = 10$ (right) Dirichlet density for $\alpha_k = 0.1$.

Samples from Dirichlet Distributions



Samples from a Dirichlet distribution for $K = 5$ when $\alpha_k = \alpha_l$ for $k \neq l$.

Bayesian Inference

- We obtain

$$
\begin{aligned}
p\left(\pi \mid y_{1:n}\right) &= \frac{p\left(\pi\right) \prod_{i=1}^{n} P\left(y_i \mid \pi\right)}{p\left(y_{1:n}\right)} \\
&= Dir\left(\pi; \alpha_1 + n_1, \ldots, \alpha_K + n_K\right)
\end{aligned}
$$

- We have

$$
\begin{aligned}
P\left(y = k \mid y_{1:n}\right) &= \int P\left(y = k \mid \pi\right) p\left(\pi \mid y_{1:n}\right) d\pi \\
&= \frac{\alpha_k + n_k}{\sum_{j=1}^{K} \alpha_j + n}.
\end{aligned}
$$

Bayesian Text Classification

- We have $\theta = \left(\pi_k, \left(\phi_k^1, \ldots, \phi_k^p\right)\right)_{k=1,\ldots,K}$ with $\pi \sim$Dir(α) and $\phi_k^l \sim Beta\,(a, b)$.
- Given data $D = (x_i, y_i)_{i=1,\ldots,n}$, classification is performed using

$$P\left(y = k|\, D, x\right) = \frac{P\left(x|\, D, y = k\right) P\left(y = k|\, D\right)}{P\left(y = k|\, D\right)}$$

  where

$$P\left(y = k|\, D\right) = \frac{\alpha_k + n_k}{\sum_{j=1}^K \alpha_j + n}$$

  and $P\left(x|\, D, y = k\right) = \prod_{l=1}^p P\left(x^l \middle|\, D, y = k\right)$ with

$$P\left(x^l \middle|\, D, y = k\right) = \frac{a + \sum_{i=1}^n \mathbb{I}\left(x_i^l = 1, y_i = k\right)}{a + b + n_k}.$$

- A popular alternative for text data consists of using as features the number of occurrences of words in document and using a multinomial model for $P\left(x|\, \phi_k\right)$.

Bayesian QDA

- Let us come back to the QDA model where

$$f\left(x\middle|\,\phi_k\right)=\mathcal{N}\left(x;\mu_k,\Sigma_k\right).$$

- We set improper priors on $\left(\mu_k,\Sigma_k\right)$ where

$$p\left(\mu_k,\Sigma_k\right)\propto\frac{\exp\left(-\frac{1}{2}tr\left(\Sigma_k^{-1}B_k\right)\right)}{\left|B_k\right|^{q/2}}$$

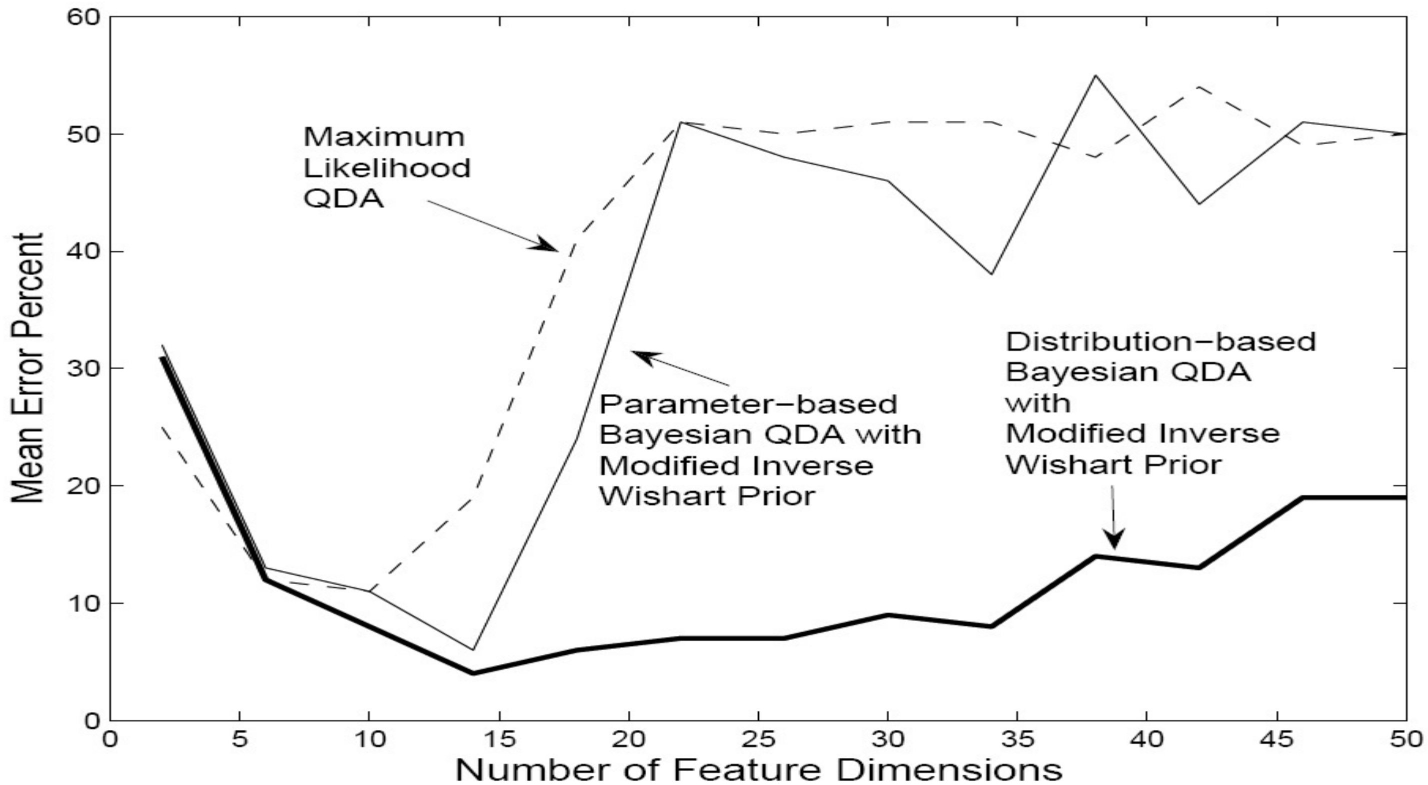
where $B_k>0$ (e.g. $B_k=\lambda I_p$ with $\lambda>>1.$) ; i.e. flat prior on μ_k and inverse-Wishart on Σ_k. Unimodal prior on Σ_k with mode B_k/q.

- It follows that

$$\begin{aligned}f\left(x\middle|\,D,y=k\right) &= \int\mathcal{N}\left(x;\mu_k,\Sigma_k\right)p\left(\mu_k,\Sigma_k\middle|\,D\right)d\mu_k d\Sigma_k\\ &= \left(\frac{n_k}{n_k+1}\right)^{p/2}\frac{\Gamma\left(\frac{n_k+q+1}{2}\right)}{\Gamma\left(\frac{n_k+q-p+1}{2}\right)}\frac{\left|\frac{S_k+B_k}{2}\right|^{\frac{n_k+q}{2}}}{\left|A_k\right|^{\frac{n_k+q+1}{2}}},\end{aligned}$$

$$A_k=\frac{1}{2}\left(S_k+\frac{n_k(x-\mu_k)(x-\mu_k)^T}{n_k+1}+B_k\right),$$
$$S_k=\sum_{i=1}^{n}I\left(y_i=k\right)\left(x_i-\widehat{\mu}_k\right)\left(x_i-\widehat{\mu}_k\right)^T.$$

Bayesian QDA



Mean error rates are shown for a two-class problem where the samples from each class are drawn from a Gaussian distribution with the same mean but different, highly ellipsoidal covariance matrices. 40 training examples, 100 test samples.