Bayesian Methods for Default Estimation in Low-Default Portfolios

Nicholas M. Kiefer
Cornell University
Departments of Economics and Statistical Sciences
and
Office of the Comptroller of the Currency (OCC)
Risk Analysis Division

March 8, 2007

Introduction

The Basel II framework provides for (some) banks to determine parameters necessary to calculate minimum capital requirements.
Banks using Internal Ratings Based (IRB) methods must calculate default probabilities (PD), and other quantities
For safe assets, calculations based on historical data may "not be sufficiently reliable" to form a probability of default estimate
This has caused concern (Newsletter 6, BBA, LIBA, ISDA paper, panicky articles).
I argue for a probability approach incorporating expert information explicitly.

The Setting and Program

Estimation of the default probability θ for a portfolio of safe assets.

Modeling of uncertainty through probabilities: The specification of the likelihood function
The role of expert information about the unknown default probability
Combination of expert and data information
Elicitation of an expert's information and representation in a probability distribution
Inference issues and supervisory issues

Preview of Results

For a low-default portfolio (described below) with 100 observations (asset-years),
The estimated PDs for 0,1,2 or 5 observed defaults are:
Maximum Likelihood (frequency estimator) 0.0, 0.01, 0.02, 0.05
Bayesian: 0.0036, 0.0052, 0.0067, 0.0109.
We will show the Bayesian estimators have a sound logical basis.

The Characterization of Uncertainty

Consider default configurations, E_i,
E_1 might be that asset 1 and only asset 1 defaults.
E_k might be a more complicated event, like "assets 3, 4 and 22-30 default."
You wish to assign numbers to describe the likelihood of the events.
Let $E_i = 1$ if event E_i occurs and $E_i = 0$ if not.
Choose numbers x_i to minimize your forecast error:

$$s(x_1, ...x_n | E_1, ..., E_n) = \sum_{i=1}^{n} (x_i - E_i)^2$$

Definition

Coherence: A set of predictions, $\{x_i^*\}$ is *coherent* if there is no alternative set of predictions $\{x_i\}$ such that

$$\begin{aligned} s(x_1, ...x_n|E_1, ..., E_n) &\leq s(x_1^*, ...x_n^*|E_1, ..., E_n) \\ &\quad \forall \{E_i\} \textit{ configurations} \\ s(x_1, ...x_n|E_1, ..., E_n) &< s(x_1^*, ...x_n^*|E_1, ..., E_n) \\ &\quad \textit{for some } \{E_i\} \end{aligned}$$

Theorem

Convexity: $0 \leq x \leq 1$.

Proof.

Suppose $x < 0$. This x only appears as $(x-1)^2$ and x^2. Both of these can be reduced by increasing x to 0. The same logic establishes that $x \leq 1$.

Additivity

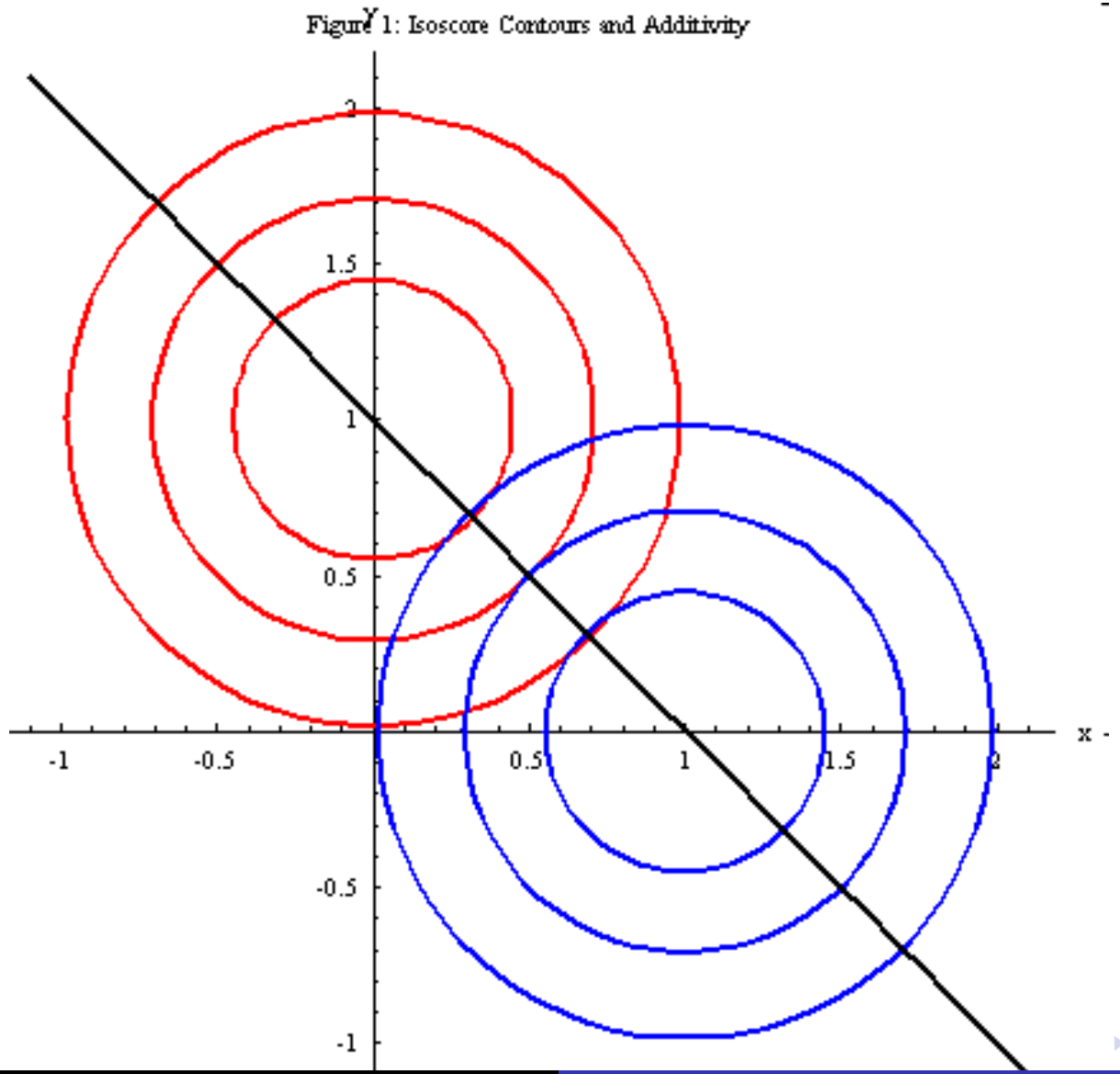
Theorem

Additivity: Let x refer to the event E and y the event $\sim E$. Then $x + y = 1$.

Proof.

Consider only terms involving x and y, The isoscore sets corresponding to the event E are spheres centered on (1,0) in the x, y plane, and the sets corresponding to $\sim E$ are spheres centered on (0,1). The coherent choices occur at tangencies, which lie on the line segment connecting the centers of the spheres. Thus, $x + y = 1$. Conditioning shows that if x is the event E, y the event F, and z the event E or F, and E and F are mutually exclusive, then $x + y = z$. □

Graph for Proof



Multiplication

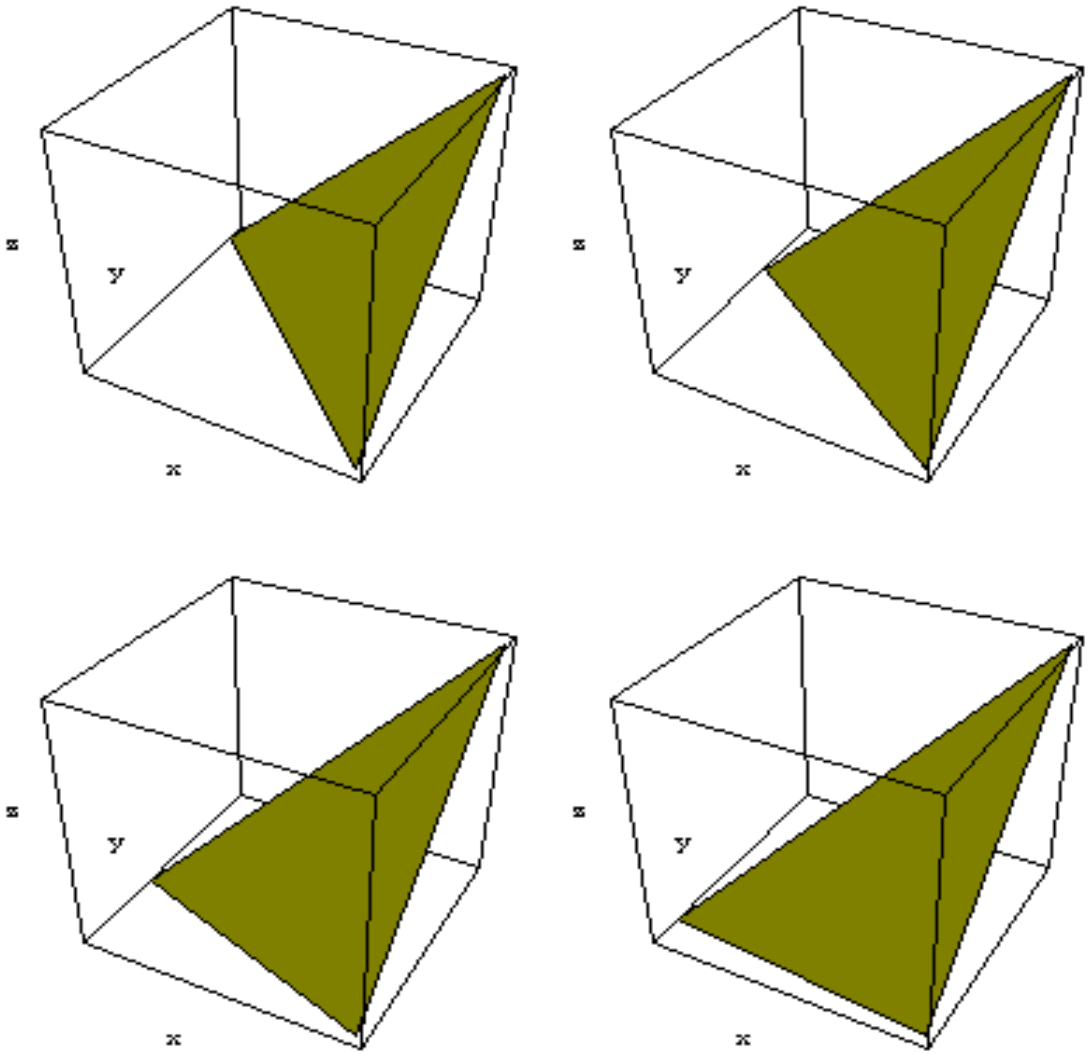
Theorem

Multiplication: Let x correspond to E, y to F given E, and z to E and F. Then $z = xy$.

Proof.

There there are 3 configurations: EF, in which case the score is $(x-1)^2 + (y-1)^2 + (z-1)^2$; $E(1-F)$, giving $(x-1)^2 + y^2 + z^2$, and $(1-E)(1-F)$, giving $x^2 + z^2$. The isoscore sets are the spheres centered on (1,1,1) and (1,0,0) and the cylinders centered on the y axis in the (x, y, z) coordinate system. The coherent triplets (x, y, z) must lie in the convex hull of (1,1,1), (0,y,0) and (1,0,0). Thus, $(x, y, z) = \alpha(1, 1, 1) + \beta(0, y, 0) + (1 - \alpha - \beta)(1, 0, 0)$, implying $z = xy$. □

Figure 2: Multiplication



Characterization (Finished)

These three properties are often taken as defining a system of probabilities.
We have obtained them from coherence
The coherent way to measure uncertainty is by probability.
This development is due to DeFinetti
The quadratic specification is inessential to give probability as the coherent measure of uncertainty (Lindley)
This development does not say what the probabilities are (numerically) or where they come from.

The likelihood function

Expert judgement e is crucial at every step of a statistical analysis. Data are asset/years In each year, there is either a default or not. We model the problem as independent Bernoulli sampling conditional on θ.
d_i indicates default ($d_i = 1$) or not ($d_i = 0$). The cond. distribution is $p(d_i|\theta, e) = \theta^{d_i}(1-\theta)^{1-d_i}$.
Let $D = \{d_i, i = 1, ..., n\}$ $r = r(D) = \sum_i d_i$ Then

$$\begin{aligned} p(D|\theta, e) &= \prod \theta^{d_i}(1-\theta)^{1-d_i} \\ &= \theta^r(1-\theta)^{n-r} \end{aligned}$$

The likelihood function 2

As a function of θ for given data D this is the likelihood function $L(\theta|D, e)$. Since this depends on D only through r, we can focus on the distribution of r

$$p(r|\theta, e) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$$

As a function of θ for given data r , this is the likelihood function $L(\theta|r, e)$.Since $r(D)$ is a sufficient statistic, no other function of the data is informative about θ given $r(D)$. The *strict* implication is that no amount of data massaging or processing can increase the data evidence on θ.

Likelihood Functions $\overline{L}(\theta|r,e) = L(\theta|r,e)/\max_\theta L(\theta|r,e)$

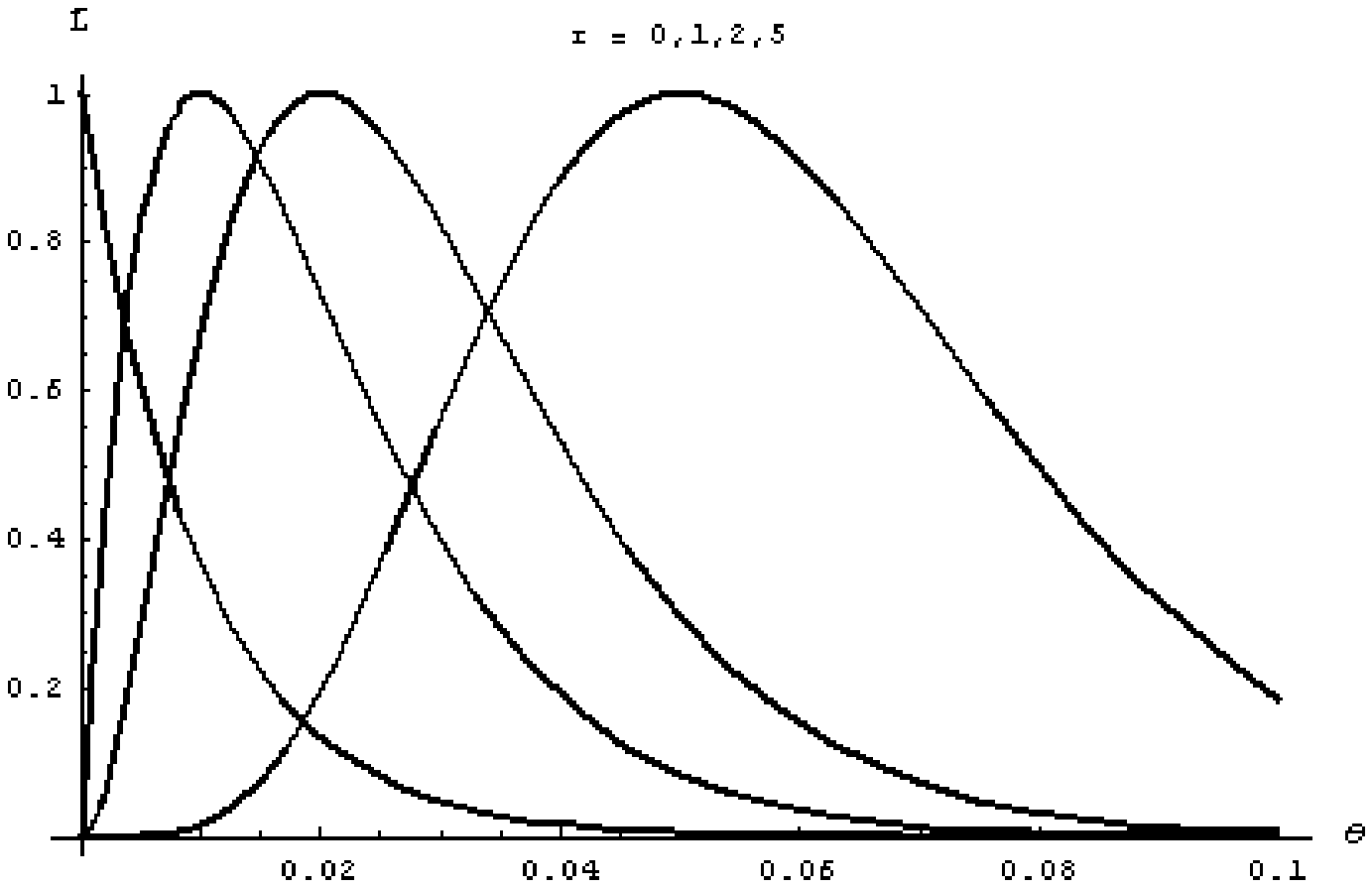


Figure 1: Likelihood Functions, n=100

Expected likelihood functions for given θ (Explain)

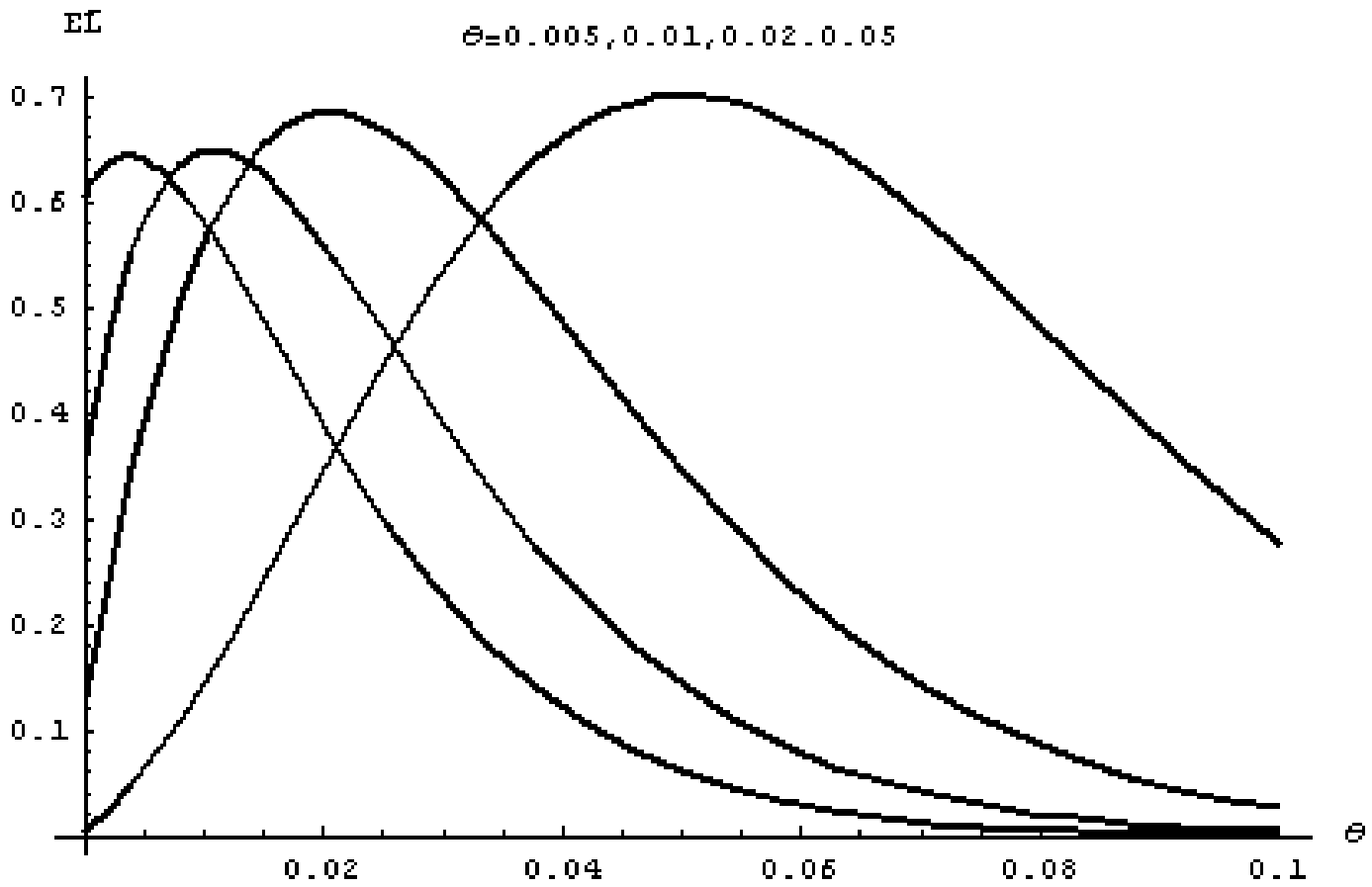


Figure 3: Expected Likelihood, n=100

Expert Information

There is some information available about θ in addition to the data.
We expect that the portfolio in question is a low-default portfolio.
Thus, we would be surprised if θ turned out to be, say, 0.2.
Further, there is a presumption that no portfolio has default probability 0.
Can this information be organized and incorporated in the analysis?
Yes.

Quantification of uncertainty about θ

We know uncertainty should be measured by probabilities
Quantification of a physical property (length, weight), involves comparison. with a standard (meter, kilogram).
Uncertainty: comparison with a simple experiment: drawing balls from an urn, sequences of coin flips.
Events: $A = "\theta \leq 0.005;"$ $B = "\theta \leq 0.01;"$ $C = "\theta \leq 0.015,"$ etc..
Assign probabilities by comparison; A is about as likely as seeing 3 heads in 50 throws of a fair coin. This requires some thought and some practice, but is feasible

Statistical Model for uncertain θ

As in the case of defaults, uncertainty is modeled with a probability distribution, $p(\theta|e)$.
Unlikely that $p(\theta|e)$, will be an exact and accurate description of beliefs.
Judgement: Does our statistical model capture the essential features of the problem?
Practical matter: functional form for the prior distribution $p(\theta|e)$.
The beta distribution for the random variable $\theta \in [0,1]$ with parameters (α, β) is

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Generalizations

The beta distribution with support $[a, b]$. This distribution has mean $E\theta = (b\alpha + a\beta)/(\alpha + \beta)$. Examples:

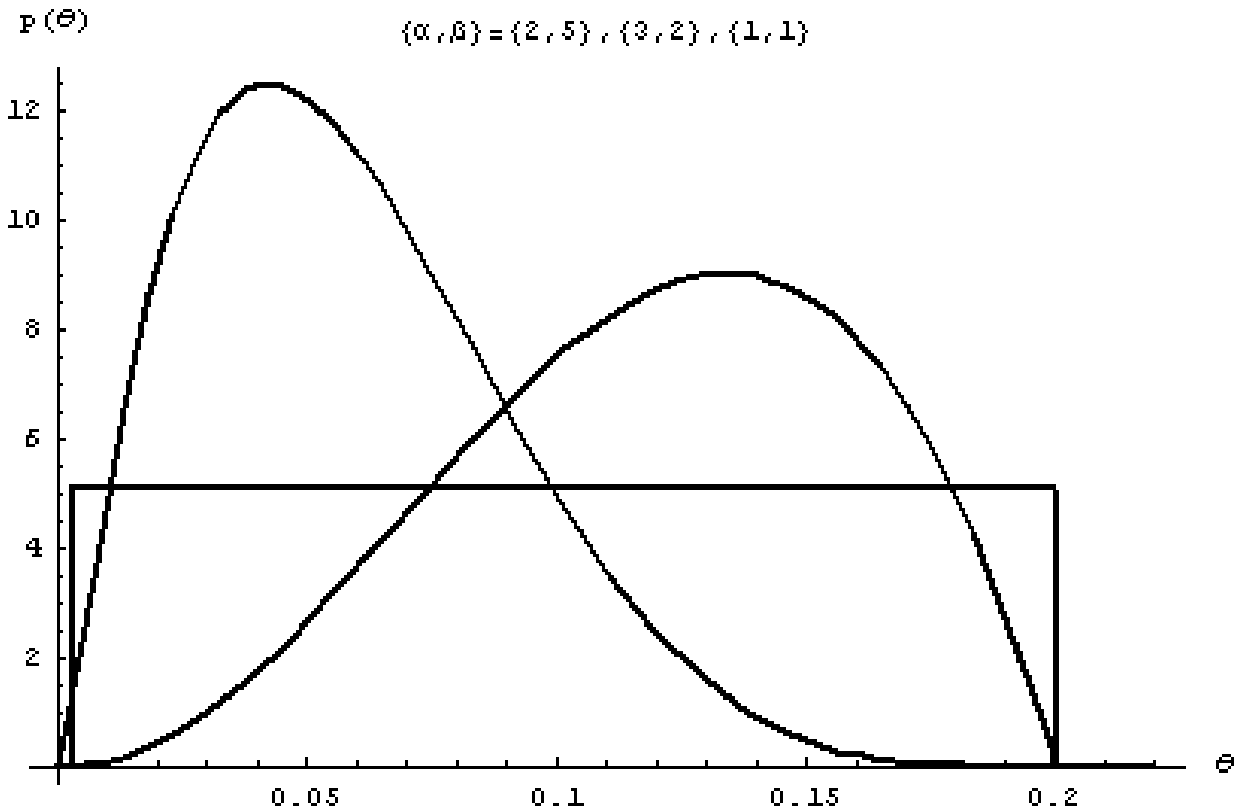


Figure 6: Examples of 4-parameter Beta Distributions

Representation of Prior Information

Beta distributions cannot represent arbitrary coherent beliefs. For example, Betas are either unimodal or have modes at the endpoints.
Mixtures give much greater flexibility,

$$p(\theta|\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda) = \lambda p(\theta|\alpha_1, \beta_1) + (1 - \lambda)p(\theta|\alpha_2, \beta_2)$$

An arbitrary continuous distribution can be approximated arbitrarily well with enough mixture components.

Updating (learning)

Ingredients: $p(\theta|e)$ (expert opinion) and $p(r|\theta, e)$ (data)
The rules for combining probabilities imply
$P(A|B)P(B) = P(A \text{ and } B) = P(B|A)P(A)$, or

$$P(B|A) = P(A|B)P(B)/P(A)$$

Applying this rule gives Bayes' rule for updating beliefs

$$p(\theta|r, e) = p(r|\theta, e)p(\theta|e)/p(r|e)$$

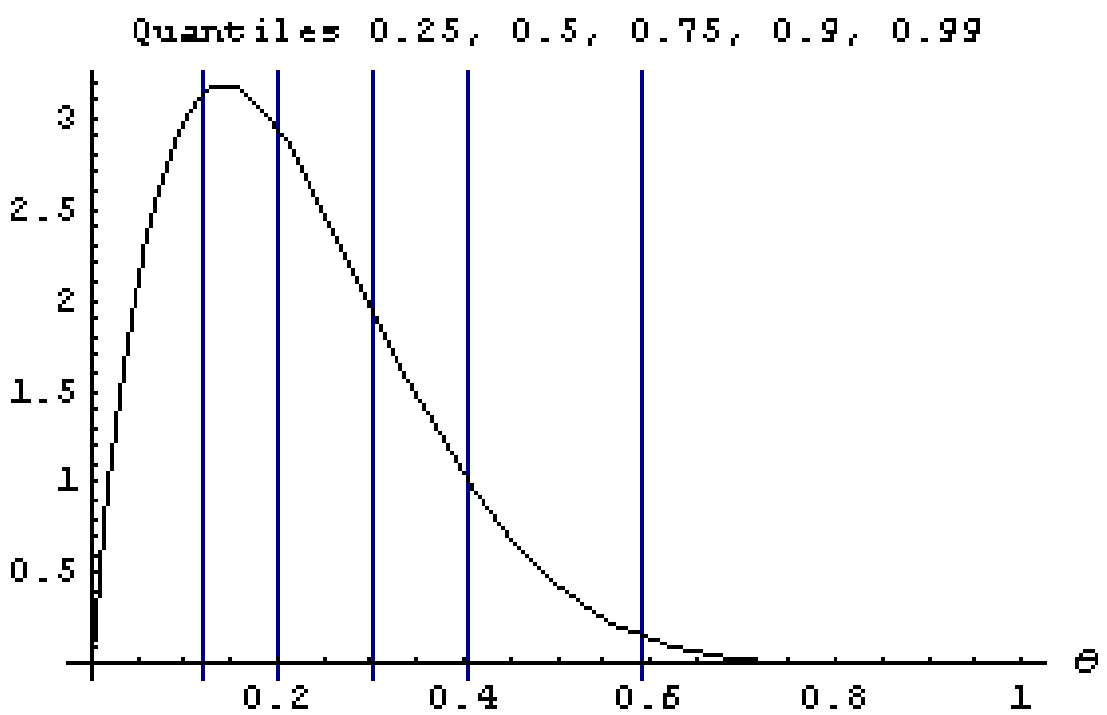
the posterior distribution describing the uncertainty about θ after observation of r defaults in n trials.
$p(r|e)$ is the marginal (in θ) distribution of the number of defaults

$$p(r|e) = \int p(r|\theta, e)p(\theta|e)d\theta.$$

Prior Distribution

I have asked an expert to specify a portfolio and give me some aspects of his beliefs about the unknown default probability.
Portfolio: loans to highly-rated large banks.
There are 50 or fewer banks in this highly-rated category
A sample period over the last 7 years or so might include 300 observations max.
Def of default is an issue. Probability of "insolvency?"
Consistent definitions in expert information and in data.

Thinking about Quantiles



Thinking about Quantiles 2

Guidance if necessary (can be ignored): 10% is a little less than the probability of getting 3 heads in 3 throws of a fair coin, 5% less than 4 heads in 4 throws, 1% a little more than 7 heads in 7 throws.
"Equally likely" is easy to deal with (though perhaps a little more subtle than it seems).

Prior Distribution 2

We considered a sample of 300 asset/years. also a "small" sample of 100 and a "large" sample of 500
Expert Information: The modal value at 300 obs was zero defaults. The expert was comfortable thinking about probabilities over probabilities. The minimum value was 0.0001 (one basis point). $\Pr(\theta > 0.035) < 0.1$, and an upper bound was 0.05. The median value was 0.0033; mean 0.005. Quartiles were assessed.
The lower was 0.00225 ("between 20 and 25 basis points"). The upper, 0.025. This set of answers is more than enough information to determine a 4-parameter Beta distribution.

Prior Distribution 3

I used a method of moments to fit parametric probability statements to the expert assessments.

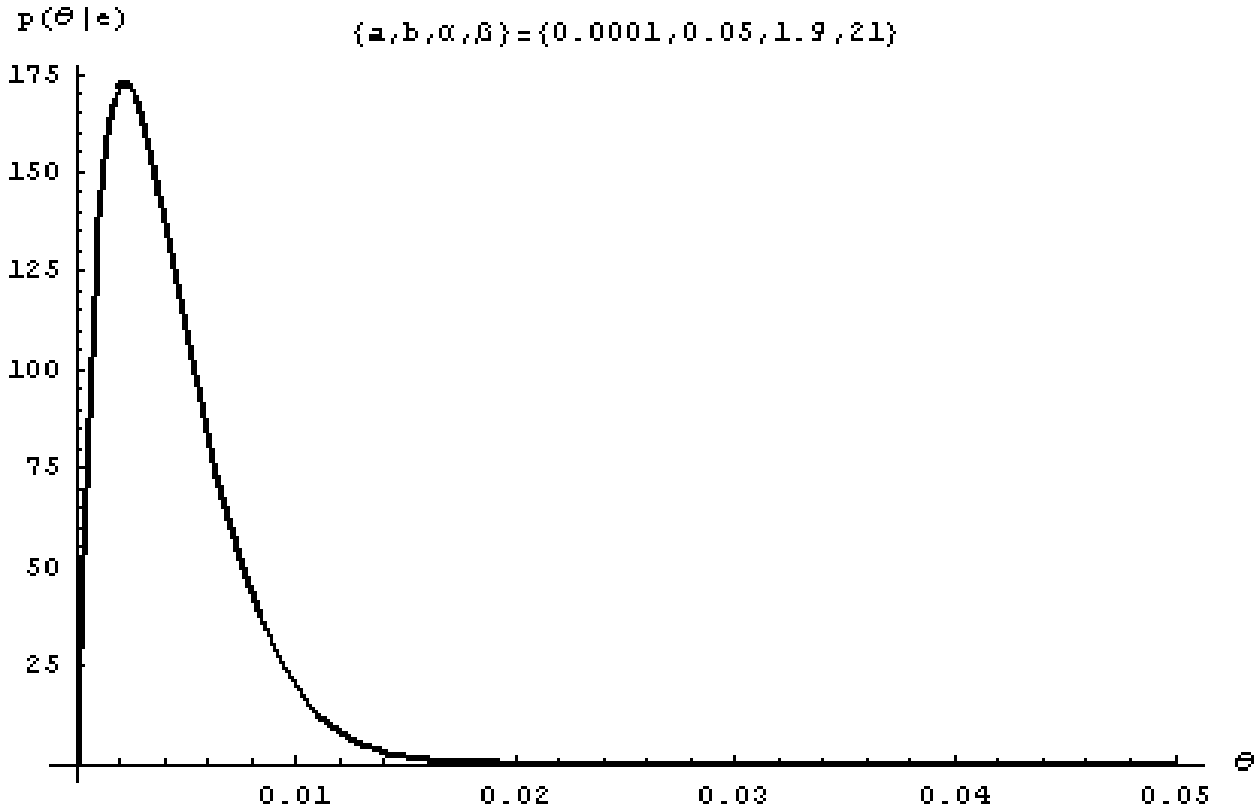


Figure 8: Distribution Reflecting Expert Information

Assessment Comments

The median of this distribution is 0.0036, the mean is 0.0042.
After the information is aggregated into an estimated probability distribution, the expert should be reconsulted. Here there was one round of feedback.
Further rounds were omitted for two reasons.
First, we are doing a hypothetical example here, Thus the prior should be realistic, but it need not be as painstakingly assessed and refined as in an application.
Second, I did not want to annoy the expert beyond the threshold of participation.

Diversion (?) Another motivation for probability

Scoring provided a motivation for quantifying uncertainty in terms of probabilities.
Another motivation can be provided by a betting argument.
Betting terms avoiding sure losses are consistent with a probability distribution.
Another reason for insisting on coherence.

Betting

Suppose a bookie (you) chooses to post betting terms on a set of events $E_k, k = 1, ..., K$.
The set of potential "states of the world" is S.
In the default application, with n asset-years there are 2^n states of the world,
An event, for example "there are exactly 10 defaults," may occur in many of these states (in this case $\binom{n}{10}$ of them).
The terms posted for a bet on E_k are the net gains in each state of the world from betting 1 on the event.

Betting 2

Arrange these terms in a matrix A $(|S| \times K)$.
x is the vector of amounts bet on each event. Ax is the vector of payoffs in each state.
Since you do not like to give away money, no column of A lies in the positive cone.
A combination of bets guaranteeing a positive payoff is called a "Dutch Book."
The requirement that Dutch Book is impossible implies coherency: You choose terms as though you had a probability distribution over states of the world.

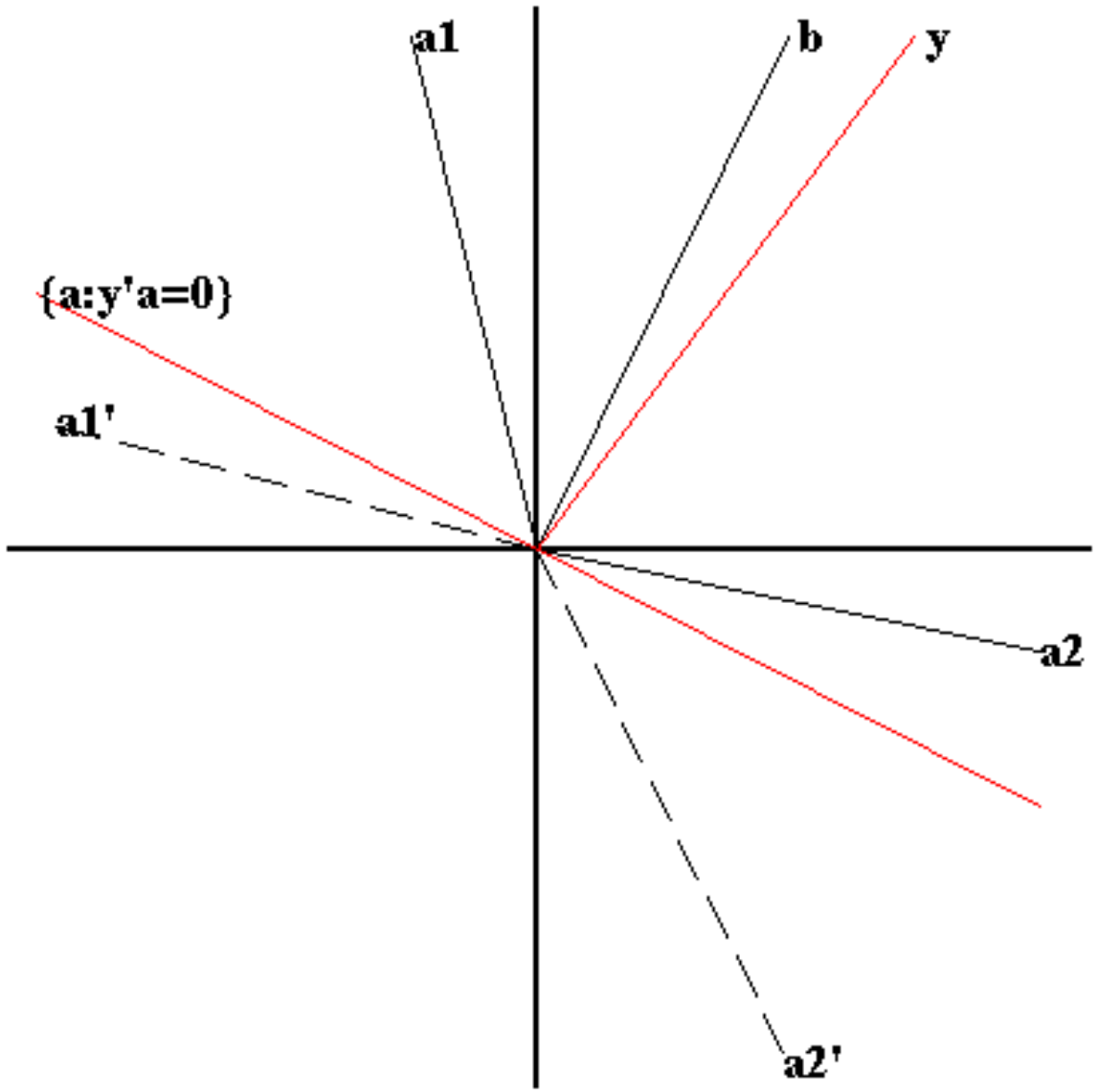
Farkas' Lemma

Lemma

Farkas' lemma: $Ax = b, x \geq 0$, *has no solution iff the system* $A'y \leq 0$, $b'y > 0$ *has a solution.*

This is a slight restatement so that y will be nonnegative in our application. The geometry is illustrated in Figure 3 for $|S| = K = 2$.

Figure 3: Farkas' Lemma and Probability



Theorem

Suppose the bookie posts terms for a set of bets including all simple events $s \in S$. Either there exists a probability distribution $\pi = (\pi_1, ..., \pi_{|S|})$ such that $\pi A = 0$ (the expected payout for each bet is zero) or there exists a Dutch Book such that sure losses are incurred.

Farkas' lemma implies that if there is no Dutch Book there exists a y with $A'y \leq 0$ and $b'y > 0$.
For any $b > 0$, y can be taken so that $y \geq 0$ and $y'1 = 1$. We now allow both positive and negative bets, by also posting terms $-A$. The interpretation is that the bookie posts terms and accepts all bets, for or against. Thus $A* = [A| - A]$. Then $A^{*'}y \leq 0$ implies $A\prime y = 0$. Then all terms offered must lead to zero expected payoffs according to the probability distribution y'.

A Familiar Argument?

This is exactly the same mathematical argument that leads to the "state space distribution" for pricing by expected value in financial markets.
If there are no arbitrage opportunities, there is a distribution of prices which values claims (usually derivatives) according to expected values.
This may not be (is not) the distribution actually generating prices.

The Predictive Distribution (Likely Data Sets)

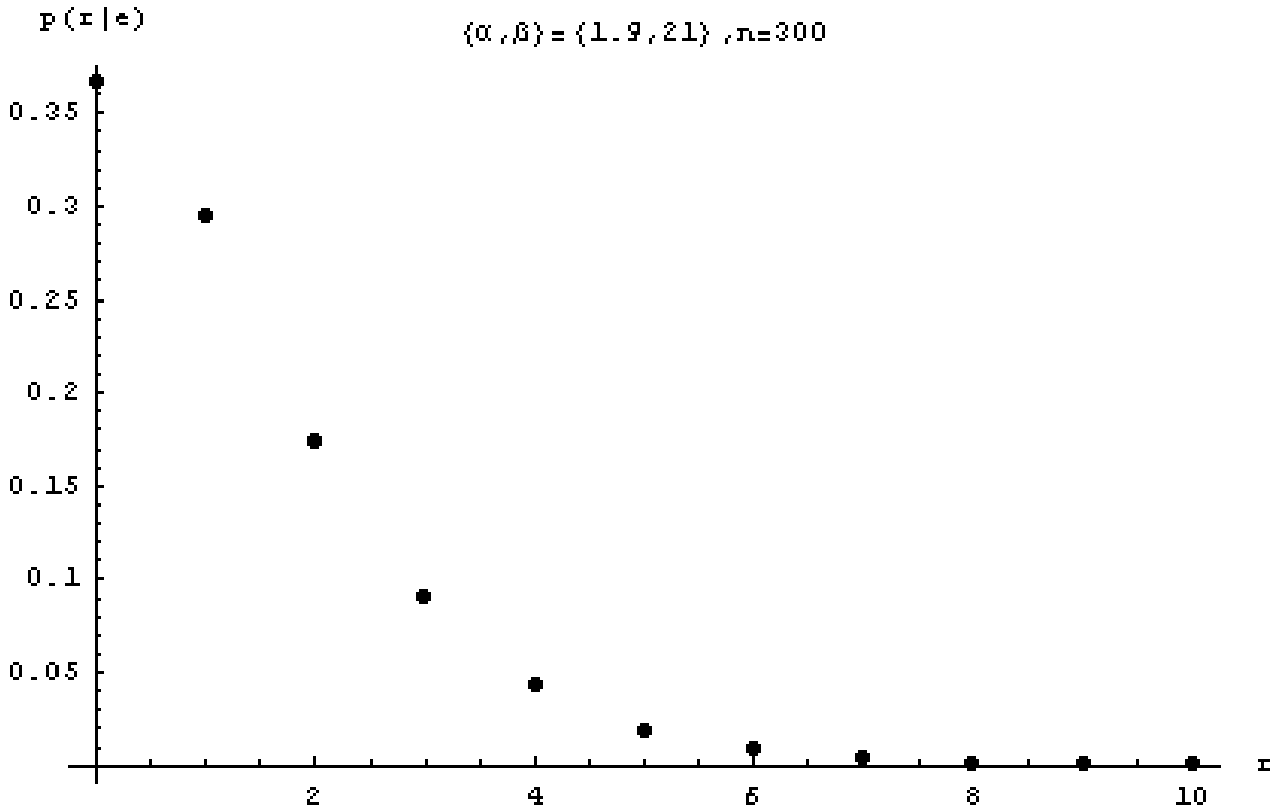


Figure 9: Predictive Distribution $p(r|\theta, e)$

$E(r|e) = \sum_{k=0}^{n} kp(k|e)$ =0.424 for n=100, 1.27 for n=300 and 2.12 for n=500. Defaults are expected to be rare events.

Posterior Analysis

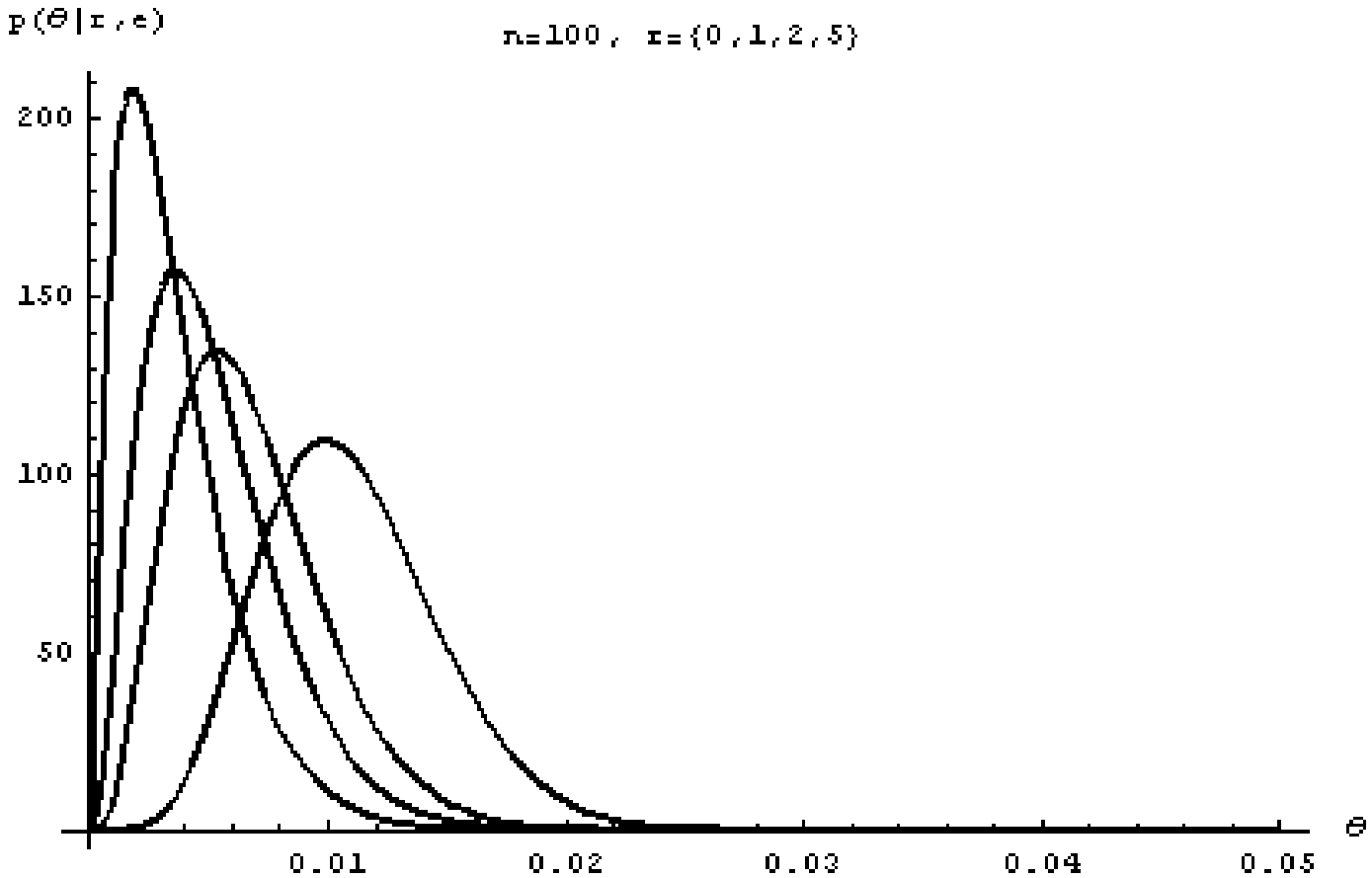


Figure 11: Posterior Distributions $p(\theta|r,e)$ for n=100

Posterior Analysis 2

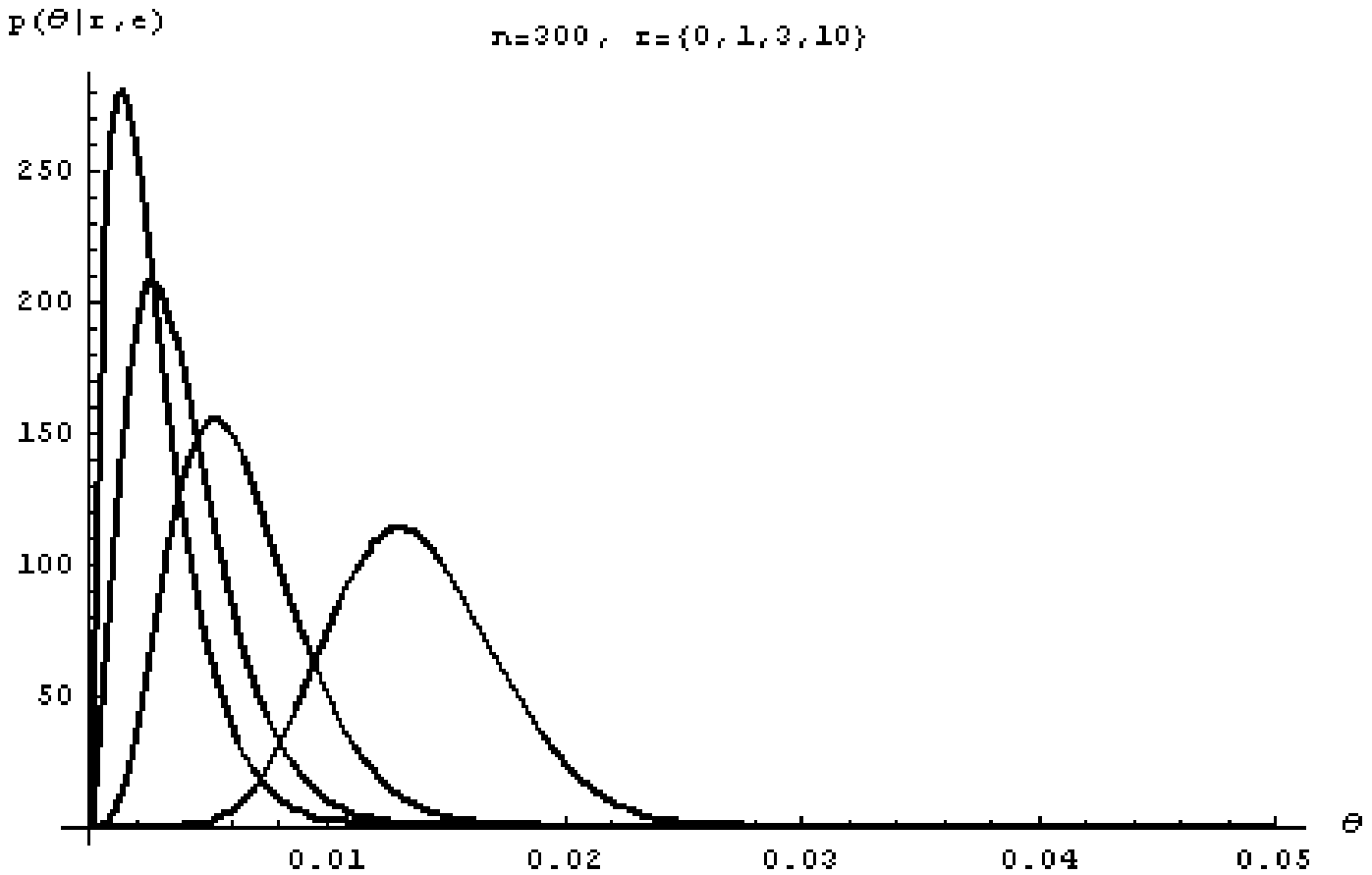


Figure 12: Posterior Distributions $p(\theta|r,e)$ for n=300

Posterior Analysis 3

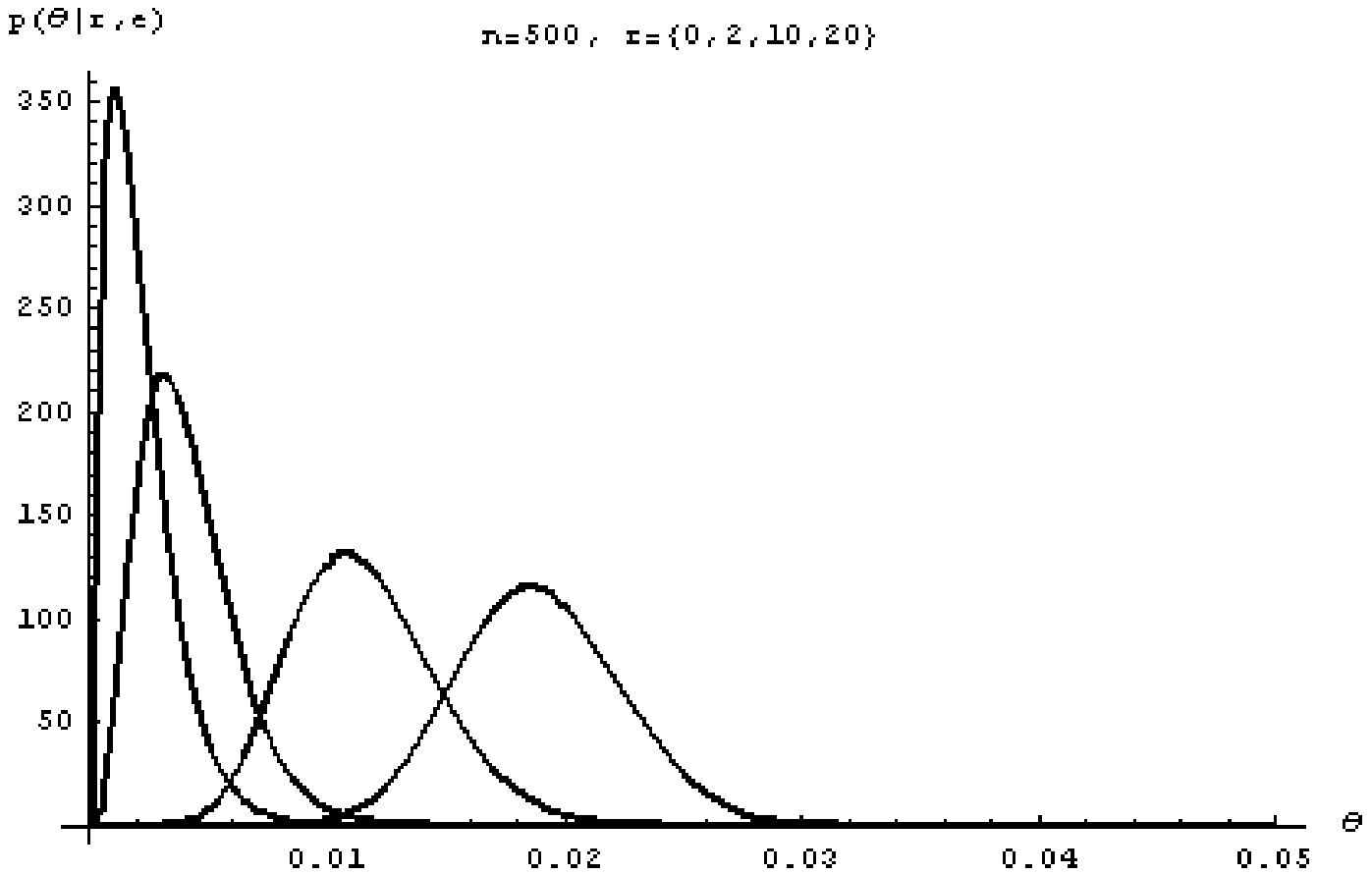


Figure 13: Posterior Distributions $p(\theta|r,e)$ for n=500

PD's for the capital model

Given the distribution $p(\theta|r,e)$, we might ask for a suitable estimator for plugging into the required capital formulas
A natural value to use is the posterior expectation, $\overline{\theta} = E(\theta|r,e)$.
An alternative, by analogy with the maximum likelihood estimator $\widehat{\theta}$, is the posterior mode $\dot{\theta}$.
As a measure of our confidence we would use the posterior standard deviation $\sigma_\theta = \sqrt{E(\theta - \overline{\theta})^2}$.
The approximate standard deviation of the maximum likelihood estimator is $\sigma_{\widehat{\theta}} = \sqrt{\widehat{\theta}(1-\widehat{\theta})/n}$.

Estimators

n	r	$\overline{\theta}$	$\dot{\theta}$	$\widehat{\theta}$	σ_θ	$\sigma_{\widehat{\theta}}$
100	0	0.0036	0.0018	0.000	0.0024	0 (!).
100	1	0.0052	0.0036	0.010	0.0028	0.0100
100	2	0.0067	0.0053	0.020	0.0031	0.0140
100	5	0.0109	0.0099	0.050	0.0037	0.0218
500	0	0.0021	0.0011	0.000	0.0015	0 (!)
500	2	0.0041	0.0032	0.004	0.0020	0.0028
500	10	0.0115	0.0108	0.020	0.0031	0.0063
500	20	0.0190	0.0185	0.040	0.0034	0.0088

Table 1: Default Probabilities: Location and Precision

Discussion of Estimators

The maximum-likelihood estimator $\widehat{\theta}$ is sensitive to small changes in the data. For n=100, the MLE ranges from 0.00-0.05 as the number of defaults ranges from 0 to 5

The posterior mean ranges in the same case from 0.0036 to 0.011

The usual estimator for the standard deviation of the maximum-likelihood estimator gives 0 when no defaults are observed.

The major differences between the posterior statistics ($\overline{\theta}$ and $\dot{\theta}$) and $\widehat{\theta}$ occur at unusual samples, and at zero.

Lesson: the posterior mean "pulls" the MLE in the direction of the prior expectation.

Remarks: Information and a suggested estimator

Information is difficult to measure. By one measure the prior information has 5 times the information in the sample of 100, almost twice the information of the 300 sample, and about the same information as the 500 sample.

(An insider comment) The confidence estimator (Pluto-Tasche) is $\theta_c = \widehat{\theta} + f(r, n)$, where $0 < f(r, n) < 1$.

Suppose we consider the probability of nondefault, $\rho = 1 - \theta$. Then $\widehat{\theta} + \widehat{\rho} = 1$ and $\overline{\theta} + \overline{\rho} = 1$, but $\theta_c + \rho_c > 1$.

The confidence estimator is *incoherent*.

A Mid-Portfolio Application

The bulk of a typical bank's portfolio are in the middle range (roughly S&P Baa or Moody's BBB)
Expert assessment: The minimum value for θ was 0.0001. Mean value 0.01
A value above 0.035 would occur with probability less than 10%, and an absolute upper bound was 0.3.
The expert did not want to rule out the possibility that the rates were much higher than anticipated (prudence?).
The median value was 0.01. Quartiles were assessed at 0.0075 and 0.0125.
The expert seemed to be thinking in terms of a normal distribution.

Prior Distribution (fit to 4-parameter Beta)

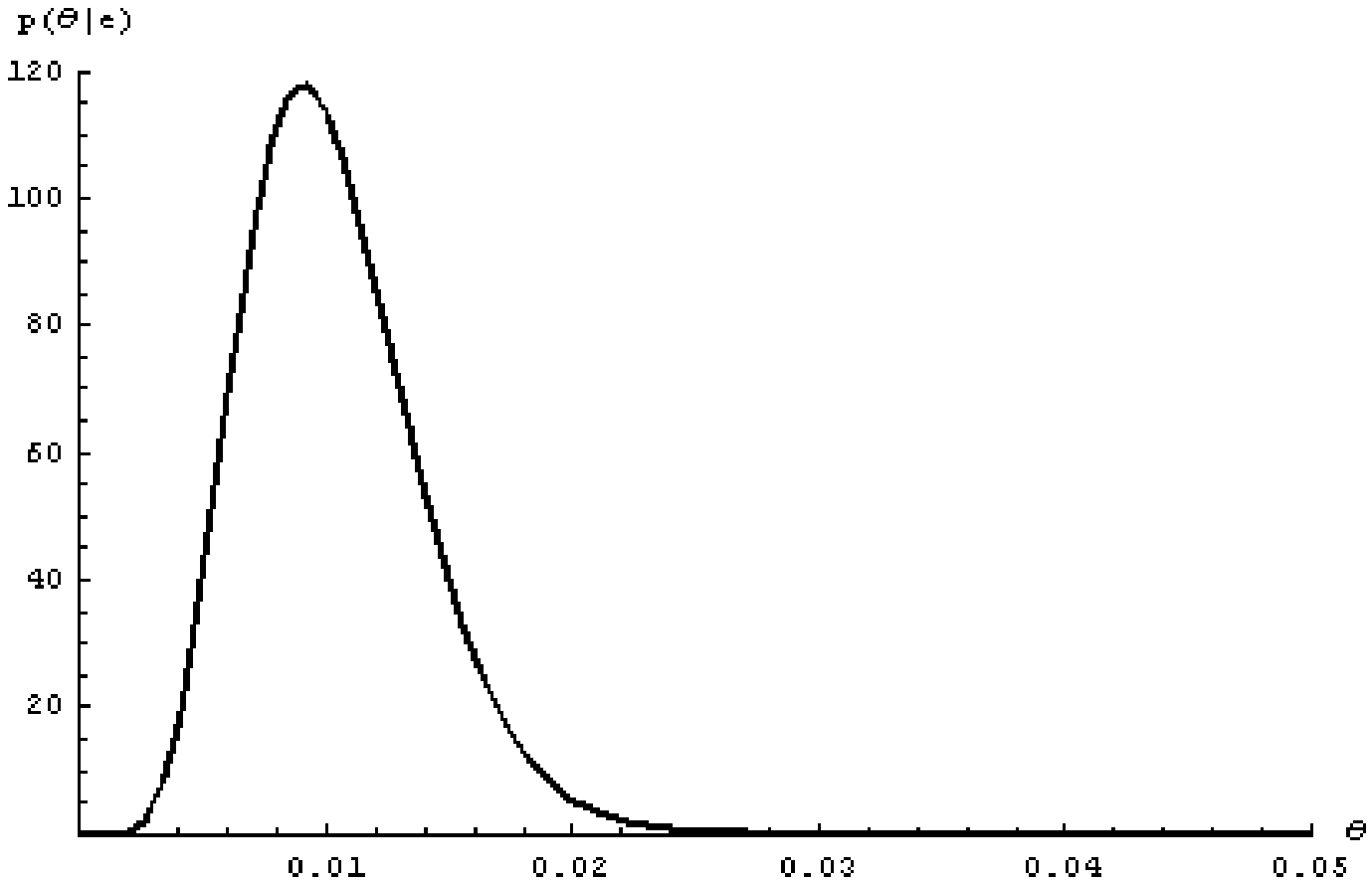


Figure 4: Expert information (closeup)

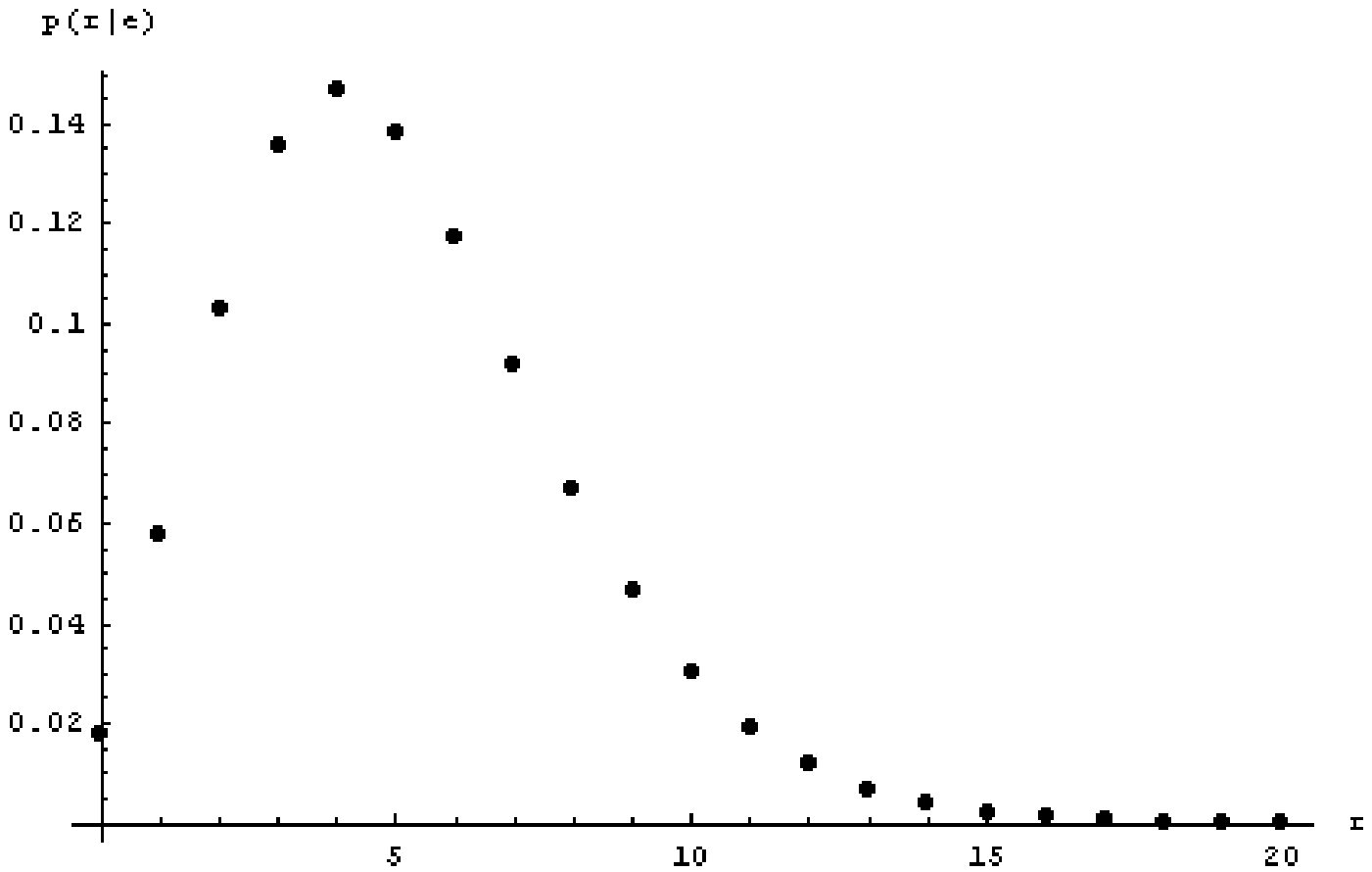


Figure 5: Predictive distribution $p(r, e)$

All Likely Datasets

Total defaults numbering 0-9 characterize 92% of expected data sets.

Analysis for these 10 data types, comprising about 2^{62} distinct datasets, covers 92% of the 2^{500} possible datasets.

Defaults are expected to be rare events.

This is the key to the ALD approach: results are applicable to 92% of the likely datasets.

n	r	$\overline{\theta}$	$\dot{\theta}$	$\widehat{\theta}$	σ_θ	$\sigma_{\widehat{\theta}}$
500	0	0.0063	0.0081	0.000	0.0022	0 (!).
500	1	0.0071	0.0092	0.002	0.0023	0.0020
500	2	0.0079	0.0103	0.004	0.0025	0.0028
500	3	0.0086	0.0114	0.006	0.0026	0.0035
500	9	0.0132	0.0180	0.018	0.0032	0.0060
500	20	0.0215	0.0296	0.040	0.0040	0.0088
500	50	0.0431	0.0425	0.100	0.0053	0.0134
500	100	0.0753	0.0749	0.200	0.0065	0.0179
500	200	0.1267	0.1266	0.400	0.0069	0.0219

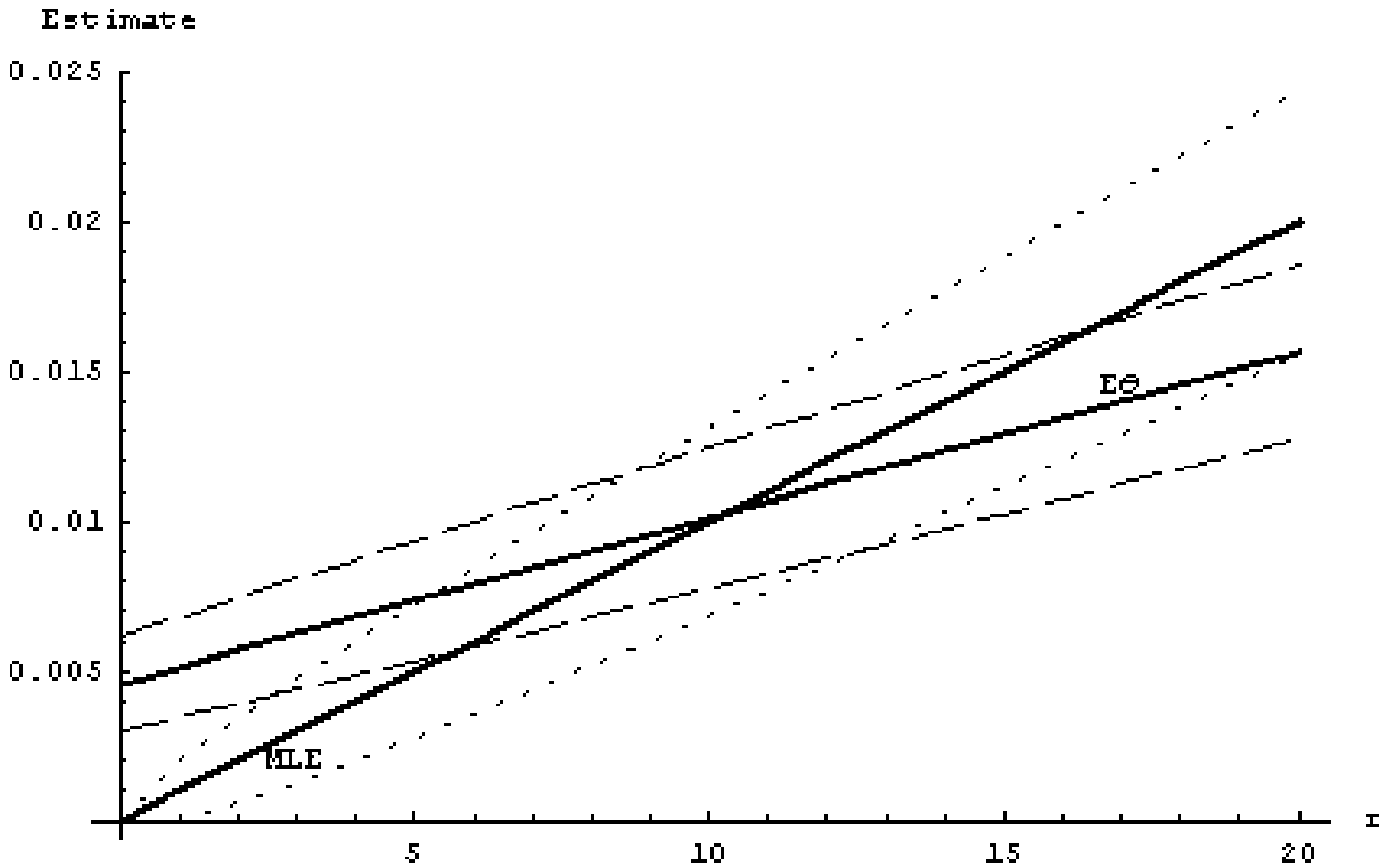


Figure 7: Eθ and MLE for n=1000

Robustness: The Cautious Bayesian

$$p(\theta|e,\epsilon) = (1-\epsilon)p(\theta|\alpha,\beta,a,b)I(\theta \in [a,b]) + \epsilon$$

n	r	$\overline{\theta}$; $\epsilon = .01$	$\overline{\theta}$; $\epsilon = .1$	$\overline{\theta}$; $\epsilon = .2$	$\overline{\theta}$; $\epsilon = .3$	$\overline{\theta}$; $\epsilon = .4$
500	0	0.0063	0.0063	0.0062	0.0061	0.0061
500	1	0.0071	0.0071	0.0071	0.0071	0.0070
500	2	0.0079	0.079	0.0079	0.0079	0.0078
500	3	0.0086	0.0086	0.0086	0.0086	0.0086
500	20	0.0358	0.0358	0.0386	0.0398	0.0405
500	50	0.1016	0.1016	0.1016	0.1016	0.1016
500	100	0.2012	0.2012	0.2012	0.2012	0.2012
500	200	0.4004	0.4004	0.4004	0.4004	0.4004

Entropy

The entropy of a distribution $p(x)$ or of the random variable X is a measure of the information value of an observation.
Entropy is

$$\begin{aligned} H(p) &= H(X) = -E\log(X) \\ &= -\sum p(x_i)\log(x_i) \\ &= -\int \log(x)p(x)dx \\ &= -\int \log(x)dP \end{aligned}$$

Entropy

The definition is most intuitive for a discrete random variable and extends to continuous or mixed variables by direct definition or by taking discrete approximations and limits.
Changing the base of the logarithm is irrelevant.
Entropy using the base 2 log can be interpreted as the expected number of binary questions ("is $x < a$") necessary to find the value of the realization. The base 2 log is extremely useful for coding results. This interpretation is not as compelling in the continuous case of "differential entropy," which can be negative. For continuous distributions it is often useful to use natural logs.

Entropy and Prior Distributions

Here we take the approach of fitting the distribution that meets the expert specification and otherwise imposes as little additional information as possible

Thus, we maximize the entropy in the distribution subject to the constraints given by the assessments.

Since we are looking for continuous distributions, we use the natural log. The general framework is to solve for the distribution p

$$\max_p \{-\int p \ln(x) dx\}$$

$$s.t. \int p(x) c_k(x) dx = 0 \text{ for } k = 1, ..., K$$

$$\text{and} \int p(x) dx = 1$$

Entropy and Prior Distributions 2

Form the Lagrangian with multipliers λ_k and μ.and differentiate with respect to $p(x)$for each x, obtaining the FOC

$$-\ln(p(x)) - 1 + \sum_k \lambda_k c_k(x) + \mu = 0$$

or

$$p(x) = \exp\{-1 + \sum_k \lambda_k c_k(x) + \mu\}$$

The multipliers are chosen so that the constraints are satisfied.

Entropy and Prior Distributions 3

The constraints are written in terms of indicator functions for the quartiles, for example the median constraint corresponds to $c(x) = I(x < .01) - 0.5$. Thus the functional form for the prior density on θ is

$$p(\theta) = k \exp\{\sum_k \lambda_k (I(\theta < q_k) - \alpha_k) + \beta\theta\}$$

Without the mean constraint the linear term in the exponent drops.

Default Probabilities - Location and Precision, n=500

r	$\overline{\theta}$	$\dot{\theta}$	$\widehat{\theta}$	σ_θ	$\sigma_{\widehat{\theta}}$	$\overline{\theta}(ent)$	$\sigma_\theta(ent)$
0	0.0063	0.0081	0.000	0.0022	0 (!).	0.0024	0.0024
1	0.0071	0.0092	0.002	0.0023	0.0020	0.0048	0.0032
2	0.0079	0.0103	0.004	0.0025	0.0028	0.0070	0.0033
3	0.0086	0.0114	0.006	0.0026	0.0035	0.0085	0.0031
4	0.0094	0.0125	0.008	0.0027	0.0040	0.0096	0.0030
5	0.0102	0.0136	0.010	0.0028	0.0044	0.0106	0.0030
6	0.0109	0.0147	0.012	0.0029	0.0049	0.0115	0.0031
7	0.0117	0.0158	0.014	0.0030	0.0053	0.0114	0.0033
8	0.0125	0.0169	0.016	0.0031	0.0056	0.0133	0.0033
9	0.0132	0.0180	0.018	0.0032	0.0060	0.0144	0.0028

Issues

The remainder of the paper discusses:
The 2004 Basel document and Newsletter 6; also relevant is the BBA, LIBA and ISDA report.
Technical issues: assessing expert information, simultaneous inference, etc
Supervisory issues: Official and formal attitude toward subjectivity.
Guidelines for using the probability approach.

Conclusion

Future defaults are unknown, so this uncertainty is modelled with a probability distribution.
The default probability is unknown. But experts do know something about it.
Uncertainty about the default probability should be modeled with a probability distribution.
A realistic example demonstrated the feasibility of the probability approach.