

Lecture I

A Gentle Introduction to Markov Chain Monte Carlo (MCMC)

Ed George
University of Pennsylvania

Seminaire de Printemps
Villars-sur-Ollon, Switzerland
March 2005

1. MCMC: A “New” Approach to Simulation

- Consider the general problem of trying to calculate characteristics of a complicated multivariate probability distribution $f(x)$ on $x = (x_1, \dots, x_p)$.
- For example, suppose we want to calculate the mean of x_1 ,

$$\int \int x_1 f(x_1, x_2) dx_1 dx_2$$

where

$$f(x_1, x_2) \propto (1 + x_1^2)^{-1} x_2^{-n} \exp \left\{ -\frac{1}{2x_2^2} \sum_i (y_i - x_1)^2 - x_2 \right\}$$

(y_1, \dots, y_n are fixed constants). Bad news: This calculation is analytically intractable.

- A Monte Carlo approach: Simulate k observations $x^{(1)}, \dots, x^{(k)}$ from $f(x)$ and use this sample to estimate the characteristics of interest. (Careful: Each $x^{(j)} = (x_1^{(j)}, \dots, x_p^{(j)})$ is a multivariate observation). For example, we could estimate the mean of x_1 by

$$\bar{x}_1 = \frac{1}{k} \sum_j x_1^{(j)}.$$

- If $x^{(1)}, \dots, x^{(k)}$ were independent observations (i.e. an iid sample), we could use standard central limit theorem results to draw inference about the quality of our estimate.
- Bad news: In many problems, methods are unavailable for direct simulation of an iid sample from $f(x)$.

- Good news: In many problems, methods such as the Gibbs sampler and the Metropolis-Hastings algorithms can be used to simulate a Markov chain $x^{(1)}, \dots, x^{(k)}$ which is converging in distribution to $f(x)$, (i.e. as k increases, the distribution of $x^{(k)}$ gets closer and closer to $f(x)$).
- Recall that a Markov chain $x^{(1)}, \dots, x^{(k)}$ is a sequence such that for each $j \geq 1$, $x^{(j+1)}$ is sampled from a distribution $p(x | x^{(j)})$ which depends on $x^{(j)}$ (but not on $x^{(1)}, \dots, x^{(j-1)}$).
- The function $p(x | x^{(j)})$ is called a Markov transition kernel. If $p(x | x^{(j)})$ is time-homogeneous (i.e. $p(x | x^{(j)})$ does not depend on j) and the transition kernel satisfies

$$\int p(x | x^*) f(x^*) dx^* = f(x),$$

then the chain will converge to $f(x)$ if it converges at all.

- Simulation of a Markov chain requires a starting value $x^{(0)}$. If the chain is converging to $f(x)$, then the dependence between $x^{(j)}$ and $x^{(0)}$ diminishes as j increases. After a suitable “burn in” period of l iterations, $x^{(l)}, \dots, x^{(k)}$ behaves like a dependent sample from $f(x)$.
- Such behavior is illustrated by Figure 1.1 on page 6 of Gilks, Richardson & Spiegelhalter (1995).
- The output from such simulated chains can be used to estimate the characteristics of $f(x)$. For example, one can obtain approximate iid samples of size m by taking the final $x^{(k)}$ values from m separate chains.
- It is probably more efficient, however, to use all the simulated values. For example, $\bar{x}_1 = \frac{1}{k} \sum_j x_1^{(j)}$ will still converge to the mean of x_1 .

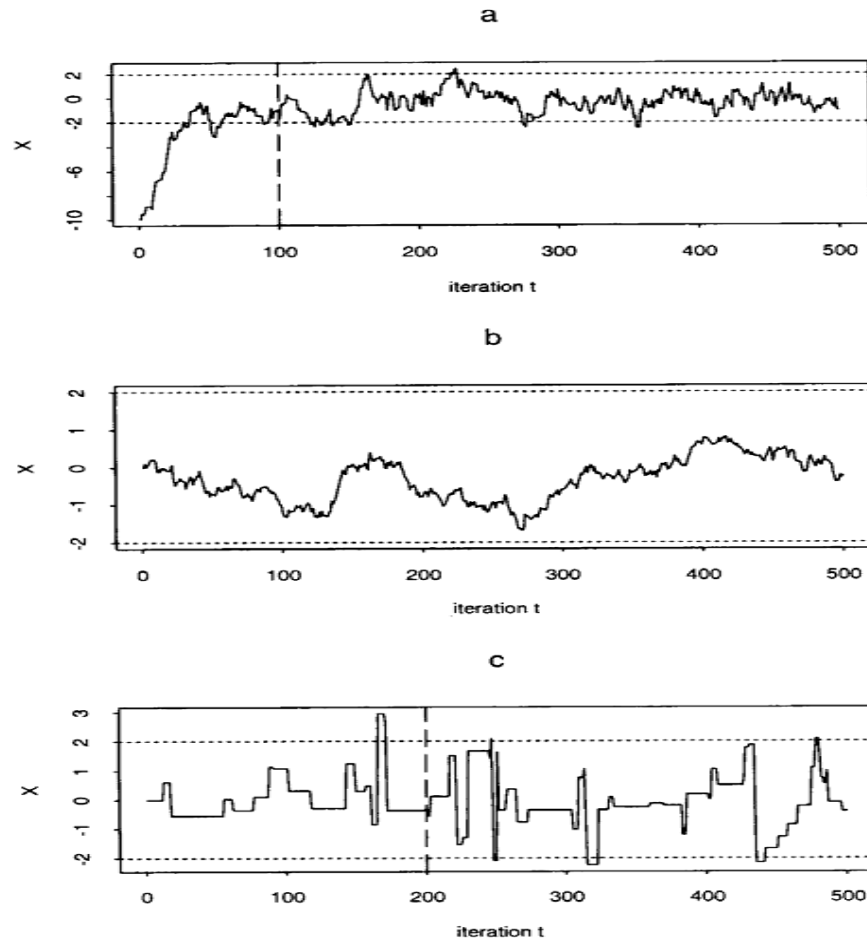


Figure 1.1 500 iterations from Metropolis algorithms with stationary distribution $N(0, 1)$ and proposal distributions (a) $q(\cdot|X) = N(X, 0.5)$; (b) $q(\cdot|X) = N(X, 0.1)$; and (c) $q(\cdot|X) = N(X, 10.0)$. The burn-in is taken to be to the left of the vertical broken line.

- MCMC is the general procedure of simulating such Markov chains and using them to draw inference about the characteristics of $f(x)$.
- Methods which have ignited MCMC are the Gibbs sampler and the more general Metropolis-Hastings algorithms. As will we now see, these are simply prescriptions for constructing a Markov transition kernel $p(x|x^*)$ which generates a Markov chain $x^{(1)}, \dots, x^{(k)}$ converging to $f(x)$.

2. The Gibbs Sampler (GS)

- The GS is an algorithm for simulating a Markov chain $x^{(1)}, \dots, x^{(k)}$ which is converging to $f(x)$, by successively sampling from the full conditional component distributions $f(x_i|x_{-i})$, $i = 1, \dots, p$, where x_{-i} denotes the components of x other than x_i .

- For simplicity, consider the case where $p = 2$. The GS generates a Markov chain

$$(x_1^{(1)}, x_2^{(1)}), (x_1^{(2)}, x_2^{(2)}), \dots, (x_1^{(k)}, x_2^{(k)})$$

converging to $f(x_1, x_2)$, by successively sampling

$$\begin{array}{lll} x_1^{(1)} & \text{from} & f(x_1 \mid x_2^{(0)}) \\ x_2^{(1)} & \text{from} & f(x_2 \mid x_1^{(1)}) \\ x_1^{(2)} & \text{from} & f(x_1 \mid x_2^{(1)}) \\ & \vdots & \\ x_1^{(k)} & \text{from} & f(x_1 \mid x_2^{(k-1)}) \\ x_2^{(k)} & \text{from} & f(x_2 \mid x_1^{(k)}) \end{array}$$

(To get started, prespecify an initial value for $x_2^{(0)}$).

- For example, suppose

$$f(x_1, x_2) \propto \binom{n}{x_1} x_2^{x_1 + \alpha - 1} (1 - x_2)^{n - x_1 + \beta - 1}$$

$$x_1 = 0, 1, \dots, n, \quad 0 \leq x_2 \leq 1.$$

The GS proceeds by successively sampling from

$$f(x_1 | x_2) = \text{Binomial}(n, x_2)$$

$$f(x_2 | x_1) = \text{Beta}(x_1 + \alpha, n - x_1 + \beta)$$

- To illustrate the GS for the above, Figure 1 of Casella & George (1992) presents a histogram of a sample of $m = 500$ final values of x_1 from separate GS runs of length $k = 10$ when $n = 16$, $\alpha = 2$ and $\beta = 4$. This is compared with an iid sample from the actual distribution $f(x_1)$, (which here can be shown to be Beta-Binomial).

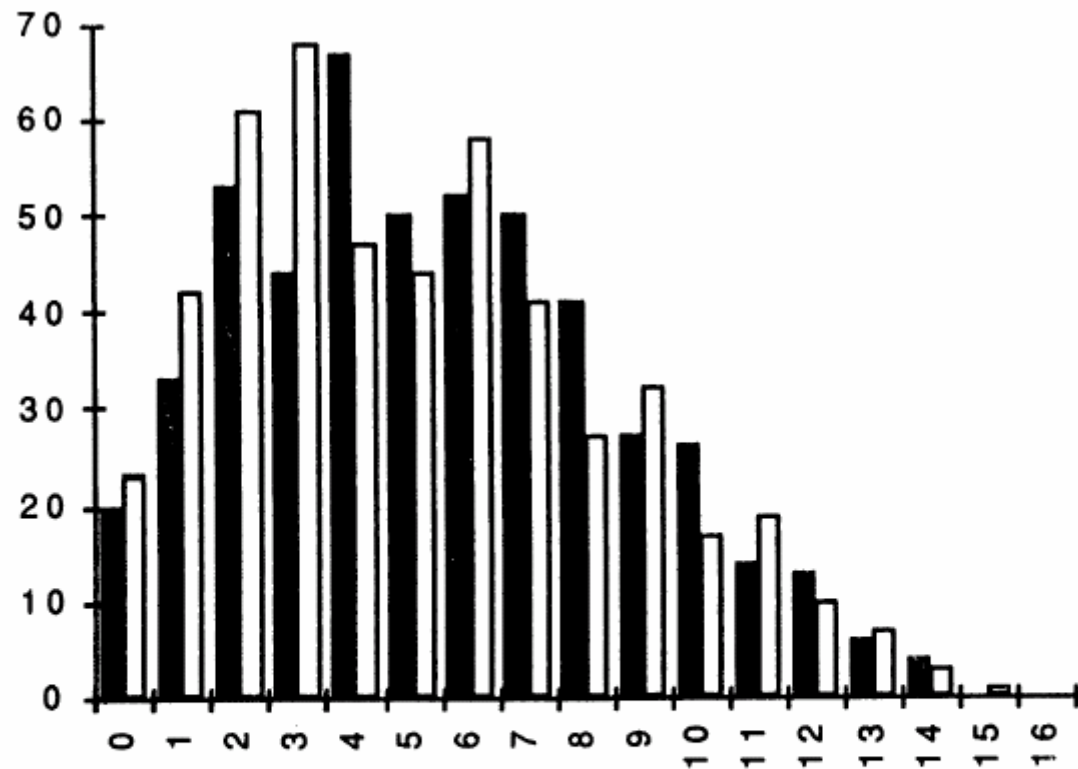


Figure 1. Comparison of Two Histograms of Samples of Size $m = 500$ From the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram sample was obtained using Gibbs sampling with $k = 10$. The white histogram sample was generated directly from the beta-binomial distribution.

- Note that $f(x_1) = \int f(x_1, x_2) dx_2 = \int f(x_1 | x_2) f(x_2) dx_2$. This expression suggests that an improved estimate of $f(x_1)$ in this example can be obtained by inserting the m values of $x_2^{(k)}$ into

$$\hat{f}(x_1) = \frac{1}{m} \sum_{i=1}^m f(x_1 | x_2^{(i)}).$$

Figure 3 of Casella & George (1992) illustrates the improvement obtained by this estimate.

- Note that the conditional distributions for the above setup, the Binomial and the Beta, can be simulated by routine methods. This is not always the case. For example, $f(x_1 | x_2)$ from page 2 is not of standard form. Fortunately, such distributions can be simulated using envelope methods such as rejection sampling, the ratio-of-uniforms method or adaptive rejection resampling. As we'll see, Metropolis-Hastings algorithms can also be used for this purpose.

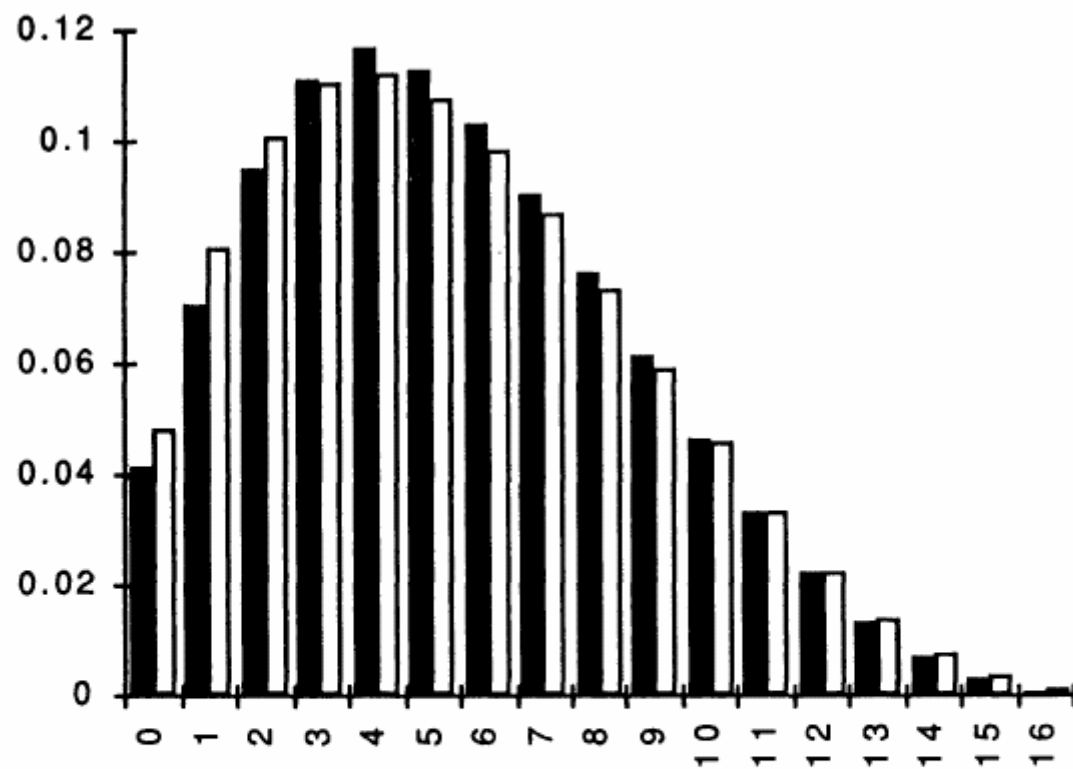


Figure 3. Comparison of Two Probability Histograms of the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram represents estimates of the marginal distribution of X using Equation (2.11), based on a sample of Size $m = 500$ from the pair of conditional distributions in (2.6). The Gibbs sequence had length $k = 10$. The white histogram represents the exact beta-binomial probabilities.

3. Metropolis-Hastings Algorithms (MH)

- MH algorithms generate Markov chains which converge to $f(x)$, by successively sampling from an (essentially) arbitrary proposal distribution $q(x|x^*)$ (i.e. a Markov transition kernel) and imposing a random rejection step at each transition.
- An MH algorithm for a candidate proposal distribution $q(x | x^*)$, entails simulating $x^{(1)}, \dots, x^{(k)}$ as follows:
 - Simulate a transition candidate x^C from $q(x | x^{(j)})$
 - Set $x^{(j+1)} = x^C$ with probability

$$\alpha(x^{(j)}, x^C) = \min \left\{ 1, \frac{q(x^{(j)} | x^C) f(x^C)}{q(x^C | x^{(j)}) f(x^{(j)})} \right\}$$

Otherwise set $x^{(j+1)} = x^{(j)}$.

- The original Metropolis algorithm was based on symmetric q , (i.e. $q(x | x^*) = q(x^* | x)$), for which α is of the simple form

$$\alpha(x^{(j)}, x^C) = \min \left\{ 1, \frac{f(x^C)}{f(x^{(j)})} \right\}.$$

- If $q(x | x^*)$ is chosen such that the Markov chain satisfies modest conditions (e.g. irreducibility and aperiodicity), then convergence to $f(x)$ is guaranteed. However, the rate of convergence will depend on the relationship between $q(x | x^*)$ and $f(x)$.
- When x is continuous, a popular choice for $q(x | x^*)$ is $x = x^* + z$ where $z \sim N_p(0, \Sigma)$. The resulting chain is called a random walk chain. Note that the choice of scale Σ can critically affect the mixing (i.e. movement) of the chain. Figure 1.1 on page 6 of Gilks, Richardson & Spiegelhalter (1995) illustrates this when $p = 1$. Other distributions for z can also be used.

- Another useful choice, called an independence sampler, is obtained when the proposal $q(x | x^*) = q(x)$ does not depend on x^* . The resulting α is of the form

$$\alpha(x^{(j)}, x^C) = \min \left\{ 1, \frac{q(x^{(j)})}{q(x^C)} \frac{f(x^C)}{f(x^{(j)})} \right\}.$$

Such samplers work well when $q(x)$ is a good heavy-tailed approximation to $f(x)$.

- It may be preferable to use an MH algorithm which updates the components $x_i^{(j)}$ of x one at a time. It can be shown that the Gibbs sampler is just a special case of such a single-component MH algorithm where q is chosen so that $\alpha \equiv 1$.

- Finally, to see why MH algorithms work, it is not too hard to show that the implied transition kernel $p(x | x^*)$ of any MH algorithm satisfies

$$p(x | x^*)f(x^*) = p(x^* | x)f(x),$$

a condition called detailed balance or reversibility. Integrating both sides of this identity with respect to x^* yields

$$\int p(x | x^*)f(x^*)dx^* = f(x),$$

showing that $f(x)$ is the limiting distribution when the chain converges.

4. The Model Liberation Movement

- Advances in computing technology have unleashed the power of Monte Carlo methods, which in turn, are now unleashing the potential of statistical modeling.

- Our new ability to simulate from complicated multivariate probability distributions via MCMC is having impact in many areas of Statistics, but most profoundly for Bayesian approaches to statistical modeling.
- The Bayesian paradigm uses probability to characterize **ALL** uncertainty as follows:
 - *Data* is a realization from a model $p(\text{Data} | \Theta)$, where Θ is an unknown (possibly multivariate) parameter.
 - Θ is treated as a realization from a prior distribution $p(\Theta)$.
 - Post-data inference about Θ is based on the posterior distribution

$$p(\Theta | \text{Data}) = \frac{p(\text{Data} | \Theta)p(\Theta)}{\int p(\text{Data} | \Theta)p(\Theta)d\Theta}$$

- In the past, analytical intractability of the expression for $p(\Theta | Data)$ severely stymied realistic practical Bayesian methods. Unrealistic, oversimplified models were too often used to facilitate calculations. MCMC has changed this, and opened up vast new realms of modeling possibilities.
- My initial example

$$f(x_1, x_2) \propto (1 + x_1^2)^{-1} x_2^{-n} \exp \left\{ -\frac{2}{x_2^2} \sum_i (y_i - x_1)^2 - x_2 \right\}$$

was a just a disguised posterior distribution for the Bayesian setup

$$y_1, \dots, y_n \text{ iid } \sim N(\mu, \sigma^2)$$

$$\mu \sim \text{Cauchy}(0, 1) \quad \sigma \sim \text{Exponential}(1).$$

The posterior of the parameters μ and σ is

$$p(\mu, \sigma | Data) \propto (1 + \mu^2)^{-1} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 - \sigma \right\}.$$

- In the above example, $f(x)$ can only be specified up to a norming constant. This is typical of Bayesian formulations. A huge attraction of GS and MH algorithms is that these norming constants are not needed.
- The previous example is just a toy problem. MCMC is in fact enabling posterior calculation for extremely complicated models with hundreds and even thousands of parameters.
- Going even further, the Bayesian approach can be used to obtain posterior distributions over model spaces. Under such formulations, MCMC algorithms are leading to new search engines which automatically identify promising models.

References For Getting Started

- Casella, G. & George, E.I. (1992) Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.
- Chib, S. & Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.
- Gilks, W. R., Richardson, S. & D.J. Spiegelhalter (1995) *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.
- Robert, C.P. & Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd Edition, Springer, New York.

Lecture II

Bayesian Approaches for Model Uncertainty

Ed George
University of Pennsylvania

Seminaire de Printemps
Villars-sur-Ollon, Switzerland
March 2005

1. A Probabilistic Setup for Model Uncertainty

- Suppose a set of K models $\{M_1, \dots, M_K\}$ are under consideration for data Y .
- Under M_k , Y has density $p(Y | \theta_k, M_k)$ where θ_k is a vector of unknown parameters that indexes the members of M_k . (More precisely, M_k is a model class).
- The Bayesian approach proceeds by assigning a prior probability distribution $p(\theta_k | M_k)$ to the parameters of each model, and a prior probability $p(M_k)$ to each model.
- Intuitively, this complete specification can be understood as a three stage hierarchical mixture model for generating the data Y ; first the model M_k is generated from $p(M_1), \dots, p(M_K)$, second the parameter vector θ_k is generated from $p(\theta_k | M_k)$, and third the data Y is generated from $p(Y | \theta_k, M_k)$.

- Letting Y_f be a future unknown observation, this formulation induces a joint distribution

$$p(Y_f, Y, \theta_k, M_k) = p(Y_f, Y \mid \theta_k, M_k)p(\theta_k \mid M_k)p(M_k).$$

- Conditioning on Y , all remaining uncertainty is captured by the joint posterior distribution $p(Y_f, \theta_k, M_k \mid Y)$. Through conditioning and marginalization, this can be used for a variety Bayesian inferences and decisions.
- For example, for prediction one would margin out both θ_k and M_k and use the predictive distribution $p(Y_f \mid Y)$ which in effect averages over all the unknown models.

- Of particular interest are the posterior model probabilities

$$p(M_k | Y) = \frac{p(Y | M_k)p(M_k)}{\sum_j p(Y | M_j)p(M_j)}$$

where

$$p(Y | M_k) = \int p(Y | \theta_k, M_k)p(\theta_k | M_k)d\theta_k$$

is the marginal or integrated likelihood of M_k .

- In terms of the three stage hierarchical mixture formulation, $p(M_k|Y)$ is the probability that M_k generated the data, i.e. that M_k was generated from $p(M_1), \dots, p(M_K)$ in the first step.
- The model posterior distribution $p(M_1|Y), \dots, p(M_K|Y)$ provides a complete post-data representation of model uncertainty and is the fundamental object of interest for model selection and model averaging.

- A natural and simple strategy for model selection is to choose the most probable M_k , the one for which $p(M_k | Y)$ largest. However, for the purpose of prediction with a single model, it may be better to use the median posterior model. Alternatively one might prefer to report a set of high posterior models along with their probabilities to convey the model uncertainty.
- Based on these posterior probabilities, pairwise comparison of models is summarized by the posterior odds

$$\frac{p(M_1 | Y)}{p(M_2 | Y)} = \frac{p(Y | M_1)}{p(Y | M_2)} \times \frac{p(M_1)}{p(M_2)}.$$

Note how the data, through the Bayes factor

$$\frac{p(Y | M_1)}{p(Y | M_2)},$$

updates the prior odds to yield the posterior odds.

2. Examples

- As a first example, consider the problem of choosing between two nonnested models, M_1 and M_2 for discrete count data $Y = (y_1, \dots, y_n)$ where

$$p(Y | \theta_1, M_1) = \pi^n (1 - \pi)^s, \quad s = \sum y_i,$$

a geometric distribution where $\theta_1 = \pi$, and

$$p(Y | \theta_2, M_2) = \frac{e^{n\lambda} \lambda^s}{\prod x_i!}$$

a Poisson distribution where $\theta_2 = \lambda$.

- Suppose further that uncertainty about $\theta_1 = \pi$ is described by a uniform prior

$$p(\pi | M_1) = 1 \text{ for } \pi \in [0, 1]$$

and uncertainty about $\theta_2 = \lambda$ is described by an exponential prior

$$p(\lambda | M_2) = e^{-\lambda} \text{ for } \lambda \in [0, \infty).$$

- Under these priors, the marginal distributions are

$$p(Y | M_1) = \int_0^1 \pi^n (1 - \pi)^s d\pi = \frac{n!s!}{(n + s + 1)!}$$

$$p(Y | M_2) = \int_0^\infty \frac{e^{-(n+1)\lambda} \lambda^s}{\prod x_i!} d\lambda = \frac{s!}{(n + 1)^{s+1} \prod x_i!}$$

- The Bayes Factor for M_1 vs M_2 is then

$$\frac{p(Y | M_1)}{p(Y | M_2)} = \frac{n!(n + 1)^{s+1} \prod x_i!}{(n + s + 1)!}.$$

When $p(M_1) = p(M_2) = 1/2$, this equals the posterior odds.

- Note that in contrast to the likelihood ratio statistic which compares maximized likelihoods, the Bayes factor compares averaged likelihoods.
- Caution - the choice of priors here can be very influential.

- As our second example, consider the problem of testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ when y_1, \dots, y_n iid $\sim N(\mu, 1)$.
- This can be treated as a Bayesian model selection problem by letting

$$p(Y | \theta_1, M_1) = p(Y | \theta_2, M_2) = (2\pi)^{-n/2} \exp \left\{ -\frac{\sum (y_i - \mu)^2}{2} \right\}$$

and assigning different priors to $\theta_1 = \theta_2 = \mu$, namely

$$Pr(\mu = 0 | M_1) = 1, \text{ i.e. a point mass at } 0$$

$$p(\mu | M_2) = (2\pi\tau^2)^{-1/2} \exp \left\{ -\frac{\mu^2}{2\tau^2} \right\}$$

- These priors yield marginal distributions $p(Y | M_1)$ and $p(Y | M_2)$ that result in a Bayes factor of the form

$$\frac{p(Y | M_1)}{p(Y | M_2)} = (1 + n\tau^2)^{1/2} \exp \left\{ -\frac{n^2\tau^2\bar{y}^2}{2(1 + n\tau^2)} \right\}$$

3. General Considerations for Prior Selection

- For a given set of models \mathcal{M} , the effectiveness of the Bayesian approach rests firmly on the specification of the parameter priors $p(\theta_k | M_k)$ and the model space prior $p(M_1), \dots, p(M_K)$.
- The most common and practical approach to prior specification in model uncertainty problems, especially large ones, is to try and construct noninformative, semi-automatic formulations, using subjective and empirical Bayes considerations where needed.
- A simple and popular choice for the model space prior is

$$p(M_k) \equiv 1/K$$

which is noninformative in the sense of favoring all models equally. However, this can be deceptive because it may not be uniform over other characteristics such as model size.

- Turning to the choice of parameter priors $p(\theta_k | M_k)$, the use of improper noninformative priors must be ruled out because their arbitrary norming constants are problematic for posterior odds comparisons.
- Proper priors guarantee the internal coherence of the Bayesian formulation and allow for meaningful hyperparameter specifications.
- An important consideration for prior specification is the analytical or numerical tractability for obtaining marginals $p(Y | M_k)$.
- For nested model formulations, centering priors is often straightforward. The crucial challenge is setting the prior dispersion. It should be large enough to avoid too much prior influence, but small enough to avoid overly diffuse specifications. Note that in our previous normal example, the Bayes factor goes to ∞ as $\tau \rightarrow \infty$, the Bartlett-Lindley paradox.

4. Extracting Information from the Posterior

- When exact calculation of the posterior is not feasible, MCMC methods can often be used to simulate an approximate sample from the posterior. This can be used to estimate posterior characteristics or to search for high probability models.
- For a model characteristic η , MCMC methods such as the such as the GS and MH algorithms entail simulation of a Markov chain, say $\eta^{(1)}, \eta^{(2)}, \dots$, that is converging to its posterior distribution $p(\eta | Y)$.
- When $p(Y | M_k)$ can be obtained analytically, the GS and MH algorithms can be applied to directly simulate a model index from

$$p(M_k | Y) \propto p(Y | M_k)p(M_k).$$

Otherwise, one must simulate from $p(\theta_k, M_k | Y)$.

- Conjugate priors are often used because of the computational advantages of having closed form expressions for $p(Y | M_k)$.
- Alternatively, it is sometimes useful to use a computable approximation for $p(Y | M_k)$ such as a Laplace approximation

$$p(Y | M_k) \approx (2\pi)^{d_k/2} |H(\tilde{\theta}_k)|^{1/2} p(Y | \tilde{\theta}_k, M_k) p(\tilde{\theta}_k | M_k)$$

where d_k is the dimension of θ_k , $\tilde{\theta}_k$ is the maximum of $h(\theta_k) \equiv \log p(Y | \theta_k, M_k) p(\theta_k | M_k)$, and $H(\tilde{\theta}_k)$ is minus the inverse Hessian of $h(\theta_k)$ evaluated at $\tilde{\theta}_k$.

- This is obtained by substituting the Taylor series approximation $h(\theta_k) \approx h(\tilde{\theta}_k) - \frac{1}{2}(\theta_k - \tilde{\theta}_k)' H(\tilde{\theta}_k)(\theta_k - \tilde{\theta}_k)$ for $h(\theta_k)$ in $p(M_k | Y) = \int \exp\{h(\theta_k)\} d\theta_k$.
- Going further people sometimes use the BIC approximation

$$\log p(Y | M) \approx \log p(Y | \hat{\theta}_k, M_k) - (d_k/2) \log n$$

obtained by using the MLE $\hat{\theta}_k$ and ignoring the terms that are constant in large samples.

References For Getting Started

Chipman, H., George, E.I. and McCulloch, R.E. (2001). The Practical Implementation of Bayesian Model Selection (with discussion). In *Model Selection* (P. Lahiri, ed.) IMS Lecture Notes – Monograph Series, Volume 38, 65-134.

Clyde, M. & George, E.I. (2004). Model Uncertainty, *Statistical Science*, 19 1 81-94.

George, E.I. (1999). Bayesian Model Selection. In *Encyclopedia of Statistical Sciences, Update Volume 3*, (eds. S. Kotz, C. Read and D. Banks), pp 39-46, Wiley, N.Y.

Lecture III

Bayesian Variable Selection

Ed George
University of Pennsylvania

Seminaire de Printemps
Villars-sur-Ollon, Switzerland
March 2005

1. The Variable Selection Problem

- Suppose one wants to model the relationship between Y a variable of interest, and a subset of x_1, \dots, x_p a set of potential explanatory variables or predictors, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and x_1, \dots, x_p is thought to contain many redundant or irrelevant variables.
- This problem has received the most attention under the normal linear model

$$Y = \beta_1 x_1 + \dots + \beta_p x_p + \epsilon \text{ where } \epsilon \sim N_n(0, \sigma^2 I)$$

when some unknown subset of regression coefficients are so small that it would be preferable to ignore them.

- This normal linear model setup is important not only because of its analytical tractability, but also because it is a canonical version of other important problems such as modern nonparametric regression.

- It will be convenient here to index each of the 2^p possible subset choices by

$$\gamma = (\gamma_1, \dots, \gamma_p)',$$

where $\gamma_i = 0$ or 1 according to whether β_i is small or large, respectively. The size of the γ th subset is denoted $q_\gamma \equiv \gamma'1$. We refer to γ as a model since it plays the same role as M_k described in Lecture II.

2. Model Space Priors for Variable Selection

- For the specification of the model space prior, most Bayesian variable selection implementations have used independence priors of the form

$$p(\gamma) = \prod w_i^{\gamma_i} (1 - w_i)^{1 - \gamma_i}.$$

- Under this prior, each x_i enters the model independently with probability $p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = w_i$.

- A useful simplification of this yields

$$p(\gamma) = w^{q_\gamma} (1 - w)^{p - q_\gamma},$$

where w is the expected proportion of x'_i 's in the model. A special case being the popular uniform prior

$$p(\gamma) \equiv 1/2^p.$$

Note that both of these priors are informative about the size of the model.

- Related priors that might also be considered are

$$p(\gamma) = \frac{B(\alpha + q_\gamma, \beta + p - q_\gamma)}{B(\alpha, \beta)}$$

obtained putting a Beta prior on w , and more generally

$$p(\gamma) = \binom{p}{q_\gamma}^{-1} h(q_\gamma)$$

obtained by putting a prior $h(q_\gamma)$ on the model size.

3. Parameter Priors for Selection of Nonzero β_i

- When the goal is to ignore only those x_i for which $\beta_i = 0$, the problem then becomes that of selecting a submodel of the form

$$Y = X_\gamma \beta_\gamma + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I)$$

where X_γ is the $n \times q_\gamma$ matrix whose columns correspond to the γ th subset of x_1, \dots, x_p and β_γ is a $q_\gamma \times 1$ vector of unknown regression coefficients. Here, (β_γ, σ^2) plays the role of θ_k described in Lecture II.

- Perhaps the most commonly applied parameter prior form for this setup is the conjugate normal-inverse-gamma prior

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(0, \sigma^2 \Sigma_\gamma),$$

$$p(\sigma^2 \mid \gamma) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2).$$

($p(\sigma^2)$ here is equivalent to $\nu\lambda/\sigma^2 \sim \chi_\nu^2$).

- A valuable feature of this prior is its analytical tractability; β_γ and σ^2 can be eliminated by routine integration to yield

$$p(Y | \gamma) \propto |X'_\gamma X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (\nu\lambda + S_\gamma^2)^{-(n+\nu)/2}$$

where

$$S_\gamma^2 = Y'Y - Y'X_\gamma(X'_\gamma X_\gamma + \Sigma_\gamma^{-1})^{-1} X'_\gamma Y.$$

The use of these closed form expressions can substantially speed up posterior evaluation and MCMC exploration, as we will see.

- In choosing values for the hyperparameters that control $p(\sigma^2)$, λ may be thought of as a prior estimate of σ^2 , and ν may be thought of as the prior sample size associated with this estimate.
- Let σ_{FULL}^2 and σ_Y^2 denote the traditional estimates of σ^2 based on the saturated and null models respectively. Treating σ_{FULL}^2 and σ_Y^2 as rough under- and over-estimates of σ^2 , one might choose λ and ν so that $p(\sigma^2)$ assigns substantial probability to the interval $(\sigma_{FULL}^2, \sigma_Y^2)$. This should at least avoid gross misspecification.

- Alternatively, the explicit choice of λ and ν can be avoided by using $p(\sigma^2) \propto 1/\sigma^2$, the limit of the inverse-gamma prior as $\nu \rightarrow 0$.
- For choosing the prior covariance matrix Σ_γ that controls $p(\beta_\gamma | \sigma^2, \gamma)$, specification is substantially simplified by setting $\Sigma_\gamma = c V_\gamma$, where c is a scalar and V_γ is a preset form such as $V_\gamma = (X'_\gamma X_\gamma)^{-1}$ or $V_\gamma = I_{q_\gamma}$, the $q_\gamma \times q_\gamma$ identity matrix.
- Having fixed V_γ , the goal is then to choose c large enough so that $p(\beta_\gamma | \sigma^2, \gamma)$ is relatively flat over the region of plausible values of β_γ , thereby reducing prior influence. At the same time it is important to avoid excessively large values of c because the Bayes factors will eventually put increasing weight on the null model as $c \rightarrow \infty$, the Bartlett-Lindley paradox. For practical purposes, a rough guide is to choose c so that $p(\beta_\gamma | \sigma^2, \gamma)$ assigns substantial probability to the range of all plausible values for β_γ . Choices of c between 10 and 10,000 seem to yield good results.

4. Posterior Calculation and Exploration

- The previous conjugate prior formulations allow for analytical margining out of β and σ^2 from $p(Y, \beta, \sigma^2 | \gamma)$ to yield a computable, closed form expression

$$g(\gamma) \propto p(Y | \gamma)p(\gamma) \propto p(\gamma | Y)$$

that can greatly facilitate posterior calculation and exploration.

- For example, when $\Sigma_\gamma = c(X'_\gamma X_\gamma)^{-1}$, we can obtain

$$g(\gamma) = (1 + c)^{-q_\gamma/2} (\nu\lambda + Y'Y - (1 + 1/c)^{-1}W'W)^{-(n+\nu)/2} p(\gamma)$$

where $W = T'^{-1}X'_\gamma Y$ for upper triangular T such that $T'T = X'_\gamma X_\gamma$ (obtainable by the Cholesky decomposition). This representation allows for fast updating of T , and hence W and $g(\gamma)$, when γ is changed one component at a time, requiring $O(q_\gamma^2)$ operations per update, where γ is the changed value.

- The availability of $g(\gamma) \propto p(\gamma | Y)$ allows for the flexible construction of MCMC algorithms that simulate a Markov chain

$$\gamma^{(1)}, \gamma^{(2)}, \gamma^{(3)}, \dots$$

converging (in distribution) to $p(\gamma | Y)$.

- A variety of such MCMC algorithms can be conveniently obtained by applying the GS with $g(\gamma)$. For example, by generating each γ component from the full conditionals

$$p(\gamma_i | \gamma_{(i)}, Y)$$

($\gamma_{(i)} = \{\gamma_j : j \neq i\}$) where the γ_i may be drawn in any fixed or random order.

- The generation of such components can be obtained rapidly as a sequence of Bernoulli draws using simple functions of the ratio

$$\frac{p(\gamma_i = 1, \gamma_{(i)} | Y)}{p(\gamma_i = 0, \gamma_{(i)} | Y)} = \frac{g(\gamma_i = 1, \gamma_{(i)})}{g(\gamma_i = 0, \gamma_{(i)})}.$$

- Such $g(\gamma)$ also facilitates the use of MH algorithms. Because $g(\gamma)/g(\gamma') = p(\gamma | Y)/p(\gamma' | Y)$, these are of the form:
 1. Simulate a candidate γ^* from a transition kernel $q(\gamma^* | \gamma^{(j)})$.
 2. Set $\gamma^{(j+1)} = \gamma^*$ with probability

$$\alpha(\gamma^* | \gamma^{(j)}) = \min \left\{ \frac{q(\gamma^{(j)} | \gamma^*)}{q(\gamma^* | \gamma^{(j)})} \frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1 \right\}. \quad (1)$$

Otherwise, $\gamma^{(j+1)} = \gamma^{(j)}$.

- A useful class of MH algorithms, the Metropolis algorithms, are obtained from the class of symmetric transition kernels of the form

$$q(\gamma^1 | \gamma^0) = q_d \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = d. \quad (2)$$

which simulate a candidate γ^* by randomly changing d components of $\gamma^{(j)}$ with probability q_d .

- When available, fast updating schemes for $g(\gamma)$ can be exploited in all these MCMC algorithms.

5. Extracting Information from the Output

- The simulated Markov chain sample $\gamma^{(1)}, \dots, \gamma^{(K)}$ contains valuable information about the posterior $p(\gamma | Y)$.
- Empirical frequencies provide consistent estimates of individual model probabilities or characteristics such as $p(\beta_i \neq 0 | Y)$.
- When closed form $g(\gamma)$ is available, we can do better. For example, the exact relative probability of any two values γ^0 and γ^1 is obtained as $g(\gamma^0) / g(\gamma^1)$ in the sequence of simulated values.

- Such $g(\gamma)$ also facilitates estimation of the normalizing constant $p(\gamma|Y) = C g(\gamma)$. Let A be a preselected subset of γ values and let $g(A) = \sum_{\gamma \in A} g(\gamma)$ so that $p(A|Y) = C g(A)$. Then, a consistent estimate of C is

$$\hat{C} = \frac{1}{g(A)K} \sum_{k=1}^K I_A(\gamma^{(k)})$$

where $I_A(\cdot)$ is the indicator of the set A .

- This yields improved estimates of the probability of individual γ values

$$\hat{p}(\gamma|Y) = \hat{C} g(\gamma),$$

as well as an estimate of the total visited probability

$$\hat{p}(B|Y) = \hat{C} g(B),$$

where B is the set of visited γ values.

- The simulated $\gamma^{(1)}, \dots, \gamma^{(K)}$ can also play an important role in model averaging. For example, suppose one wanted to predict a quantity of interest Δ by the posterior mean

$$E(\Delta | Y) = \sum_{\text{all } \gamma} E(\Delta | \gamma, Y) p(\gamma | Y).$$

When p is too large for exhaustive enumeration and $p(\gamma | Y)$ cannot be computed, $E(\Delta | Y)$ is unavailable and is typically approximated by something of the form

$$\hat{E}(\Delta | Y) = \sum_{\gamma \in S} E(\Delta | \gamma, Y) \hat{p}(\gamma | Y, S)$$

where S is a manageable subset of models and $\hat{p}(\gamma | Y, S)$ is a probability distribution over S . (In some cases, $E(\Delta | \gamma, Y)$ will also need to be approximated).

- Letting S be the sampled values, a natural and consistent choice for $\hat{E}(\Delta | Y)$ is

$$\hat{E}_f(\Delta | Y) = \sum_{\gamma \in S} E(\Delta | \gamma, Y) \hat{p}_f(\gamma | Y, S)$$

where $\hat{p}_f(\gamma | Y, S)$ is the relative frequency of γ in S . However, it appears that when $g(\gamma)$ is available, one can do better by using

$$\hat{E}_g(\Delta | Y) = \sum_{\gamma \in S} E(\Delta | \gamma, Y) \hat{p}_g(\gamma | Y, S)$$

where $\hat{p}_g(\gamma | Y, S) = g(\gamma)/g(S)$ is the renormalized value of $g(\gamma)$.

- For example, when S is an iid sample from $p(\gamma | Y)$, $\hat{E}_g(\Delta | Y)$ approximates the best unbiased estimator of $E(\Delta | Y)$ as the sample size increases. To see this, note that when S is an iid sample, $\hat{E}_f(\Delta | Y)$ is unbiased for $E(\Delta | Y)$. Since S (together with g) is sufficient, the Rao-Blackwellized estimator $E(\hat{E}_f(\Delta | Y) | S)$ is best unbiased. But as the sample size increases, $E(\hat{E}_f(\Delta | Y) | S) \rightarrow \hat{E}_g(\Delta | Y)$.

6. Calibration and Empirical Bayes Variable Selection

- Let us now focus on the special case when the conjugate normal-inverse-gamma prior,

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(0, c\sigma^2(X'_\gamma X_\gamma)^{-1}),$$

is combined with

$$p(\gamma) = w^{q_\gamma}(1 - w)^{p - q_\gamma}$$

the simple independence prior; for the moment, let's assume σ^2 is known.

- The hyperparameter c controls the expected size of the nonzero coefficients of $\beta = (\beta_1, \dots, \beta_p)'$. The hyperparameter w controls the expected proportion of such nonzero components.

- Surprise! We will see that this prior setup is related to the canonical penalized sum-of-squares criterion

$$C_F(\gamma) \equiv SS_\gamma / \sigma^2 - F q_\gamma$$

where $SS_\gamma = \hat{\beta}'_\gamma X'_\gamma X_\gamma \hat{\beta}_\gamma$, $\hat{\beta}_\gamma \equiv (X'_\gamma X_\gamma)^{-1} X'_\gamma Y$ and F is a fixed penalty value for adding a variable.

- Popular model selection criteria simply entail maximizing $C_F(\gamma)$ with particular choices of F and $\sigma^2 = \hat{\sigma}^2$.
- For orthogonal variables, x_i added $\Leftrightarrow t_i^2 > F$.
- Some choices for F
 - $F = 0$: Select full model
 - $F = 2$: Cp and AIC
 - $F = \log n$: BIC
 - $F = 2 \log p$: RIC

- The relationship with $C_F(\gamma)$ is obtained by reexpressing the model posterior under the prior setup as

$$p(\gamma | Y) \propto \exp \left[\frac{c}{2(1+c)} \{SS_\gamma / \sigma^2 - F(c, w) q_\gamma\} \right],$$

where

$$F(c, w) = \frac{1+c}{c} \left\{ 2 \log \frac{1-w}{w} + \log(1+c) \right\}.$$

- As a function of γ for fixed Y , $p(\gamma | Y)$ is increasing in $C_F(\gamma)$ when $F = F(c, w)$. Thus, Bayesian model selection based on $p(\gamma | Y)$ is equivalent to model selection based on the criterion $C_{F(c, w)}(\gamma)$. For example, by appropriate choice of c, w , the mode of $p(\gamma | Y)$ can be made to correspond to the best C_p , AIC, BIC or RIC models.
- Since c and w control the expected size and proportion of the nonzero components of β , the dependence of $F(c, w)$ on c and w provides an implicit connection between the penalty F and the profile of models for which its value may be appropriate.

- The awful truth: c and w are unknown
- Empirical Bayes Idea: Use \hat{c} and \hat{w} which maximize the marginal likelihood

$$\begin{aligned}
 L(c, w | Y, \sigma) &\propto \sum_{\gamma} p(\gamma | w) p(Y | \sigma, \gamma, c) \\
 &\propto \sum_{\gamma} w^{q_{\gamma}} (1 - w)^{p - q_{\gamma}} (1 + c)^{-q_{\gamma}/2} \exp \left\{ \frac{c SS_{\gamma}}{2\sigma^2(1 + c)} \right\}.
 \end{aligned}$$

- For orthogonal x 's (and σ known), this simplifies to

$$L(c, w | Y, \sigma) \propto \prod_{i=1}^p [(1 - w)e^{-t_i^2/2} + w(1 + c)^{-1/2} e^{-t_i^2/2(1+c)}]$$

where $t_i = b_i v_i / \sigma$ is the t-statistic associated with x_i

- At least in the orthogonal case, \hat{c} and \hat{w} can be found numerically using Gauss-Seidel, EM algorithm, etc.

- The best marginal maximum likelihood model is then the one which maximizes the “posterior” $p(\gamma | Y, \hat{c}, \hat{w}, \sigma)$ or equivalently

$$C_{\text{MML}} \equiv C_{F(\hat{c}, \hat{w})}$$

- In contrast to criteria of the form $C_F(\gamma)$ with prespecified fixed F , C_{MML} uses an adaptive penalty $F(\hat{c}, \hat{w})$ that is implicitly based on the estimated distribution of the regression coefficients.
- Estimating β_γ after selecting $\hat{\gamma}_{\text{MML}}$ might then proceed using

$$E(\beta_\gamma | Y, \hat{c}, \hat{w}, \sigma, \hat{\gamma}_{\text{MML}}) = \frac{\hat{c}}{1 + \hat{c}} \hat{\beta}_{\hat{\gamma}_{\text{MML}}}$$

- A computable conditional maximum likelihood approximation C_{CML} for the nonorthogonal case is available.

- Consider the simple model with $X = I$,

$$Y = \beta + \epsilon \text{ where } \epsilon \sim N_n(0, I)$$

where $\beta = (\beta_1, \dots, \beta_p)'$ is such that

$$\beta_1, \dots, \beta_q \text{ iid } \sim N(0, c)$$

$$\beta_{q+1}, \dots, \beta_p \equiv 0$$

- For $p = n = 1000$, and fixed values of c and q , simulated Y from the above model
- Evaluate $\hat{\gamma}$ by estimating

$$R(\beta, \hat{\gamma}) \equiv E_{c,q} \sum_i (Y_i I[x_i \in \hat{\gamma}] - \beta_i)^2$$

- Figures 1ab and 2 illustrate the adaptive advantages of the empirical Bayes selection criteria.

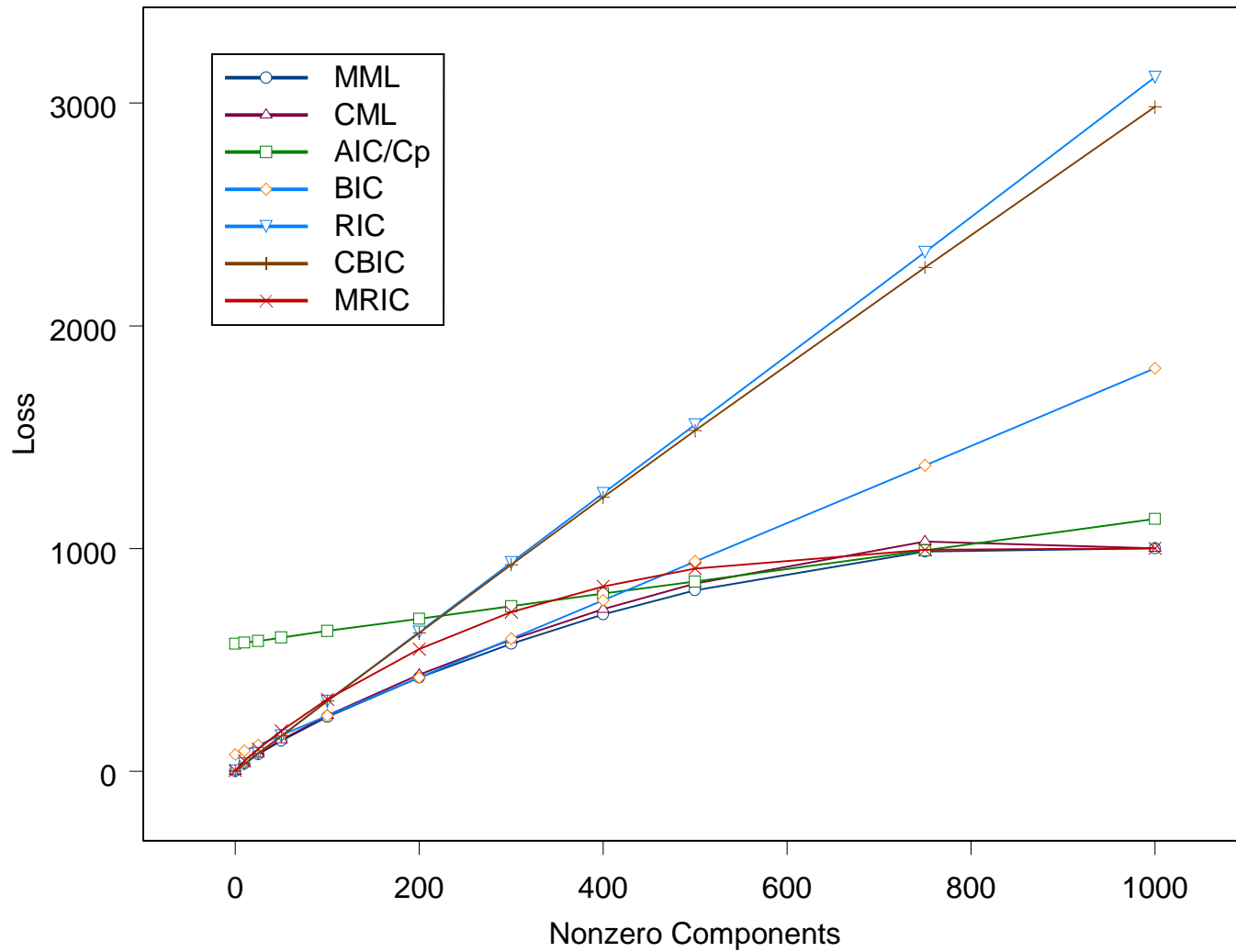


Figure 1(a). The average loss of the selection procedures when $c = 25$ and the number of nonzero components $q = 0, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000$. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

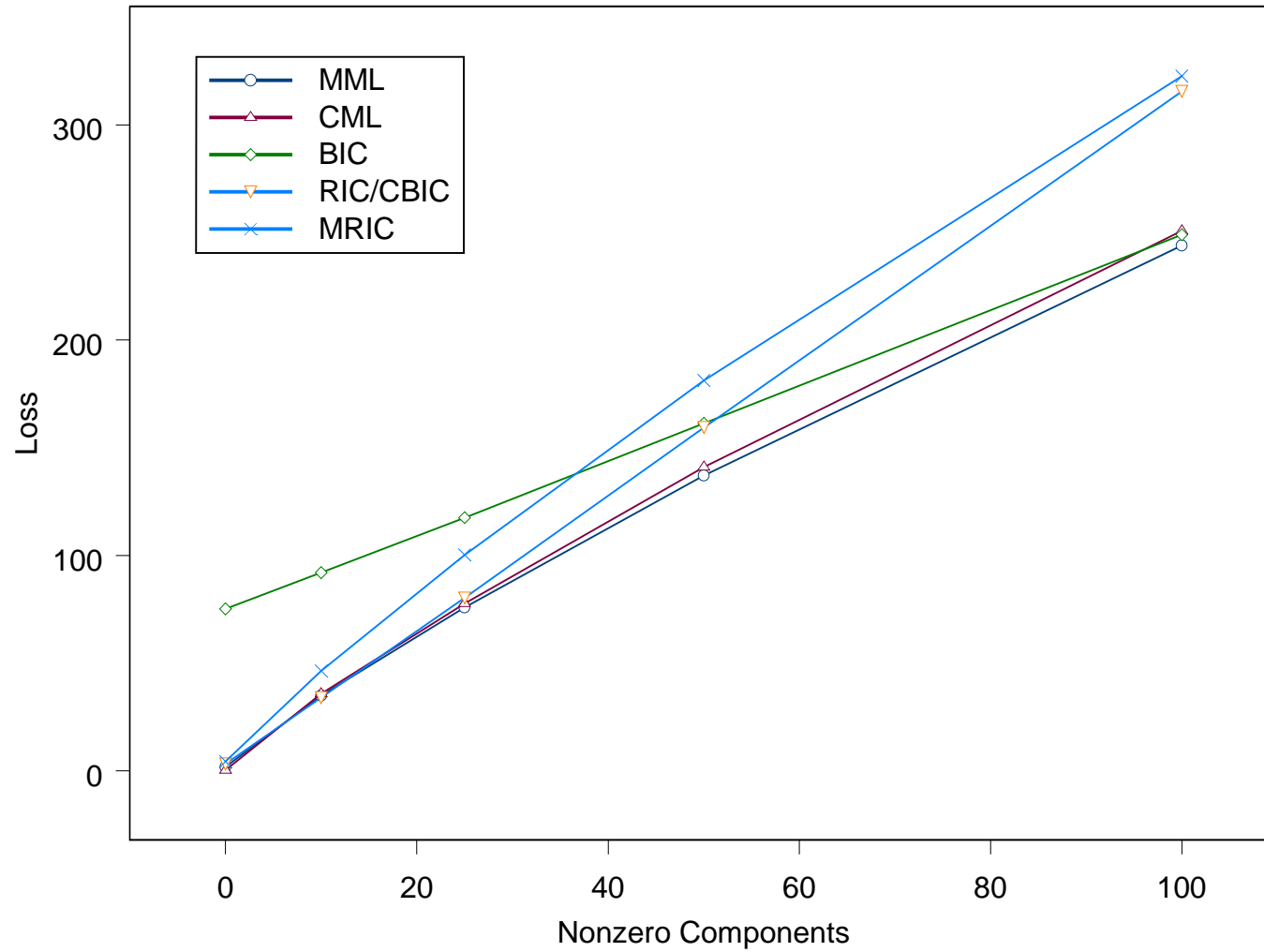


Figure 1(b). The average loss of the selection procedures when $c = 25$ and the number of nonzero components $q = 0, 10, 25, 50, 100$. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC. RIC and CBIC are virtually identical here and so have been plotted together.

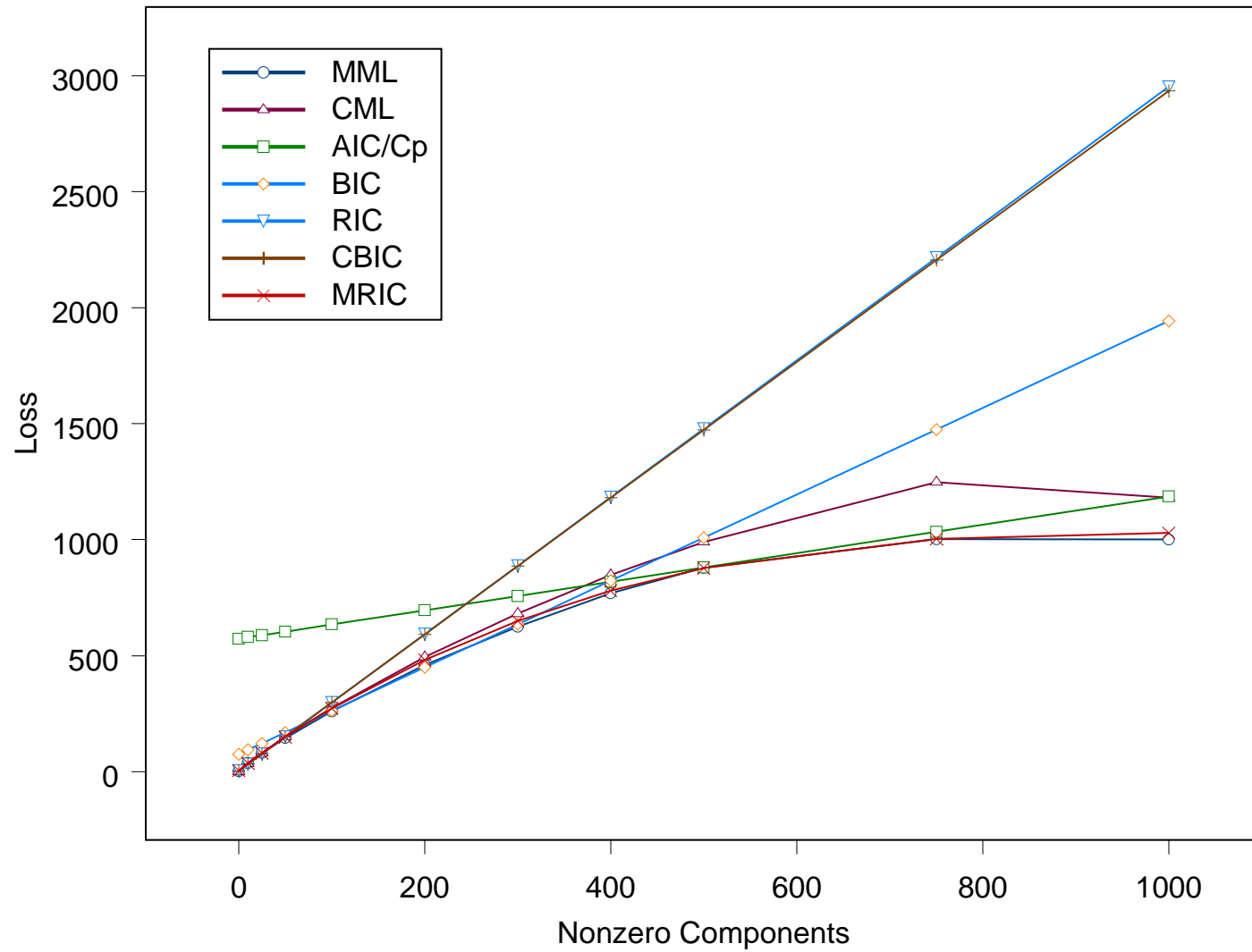


Figure 1(c). The average loss of the selection procedures when $c = 5$ and the number of nonzero components $q = 0, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000$. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

References For Getting Started

Chipman, H., George, E.I. and McCulloch, R.E. (2001). The Practical Implementation of Bayesian Model Selection (with discussion). In *Model Selection* (P. Lahiri, ed.) IMS Lecture Notes – Monograph Series, Volume 38, 65-134.

George, E.I. and Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-748.

Lecture IV

High Dimensional Predictive Estimation

Ed George
University of Pennsylvania

Seminaire de Printemps
Villars-sur-Ollon, Switzerland
March 2005

1. Estimating a Normal Mean: A Brief History

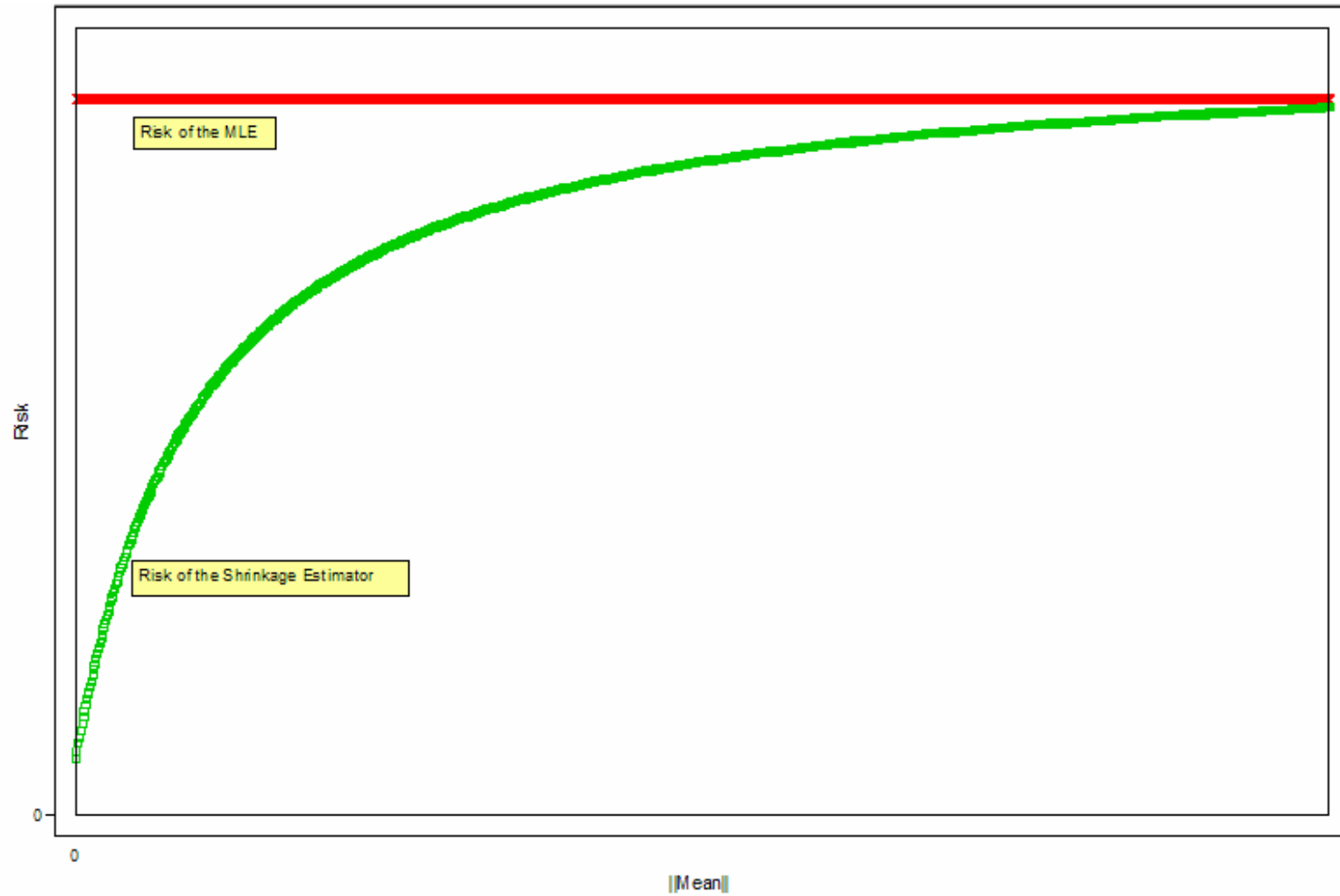
- Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2$$

- $\hat{\mu}_{MLE}(X) = X$ is the MLE, best invariant and minimax with constant risk
- Shocking Fact: $\hat{\mu}_{MLE}$ is inadmissible when $p \geq 3$. (Stein 1956)
- Bayes rules are a good place to look for improvements
- For a prior $\pi(\mu)$, the Bayes rule $\hat{\mu}_\pi(X) = E_\pi(\mu \mid X)$ minimizes $E_\pi R_Q(\mu, \hat{\mu})$
- Remark: The (formal) Bayes rule under $\pi_U(\mu) \equiv 1$ is

$$\hat{\mu}_U(X) \equiv \hat{\mu}_{MLE}(X) = X$$

The Risk Functions of Two Minimax Estimators



- $\hat{\mu}_H(X)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $\hat{\mu}_U$ when $p \geq 3$. (Stein 1974)

- $\hat{\mu}_a(X)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu \mid s \sim N_p(0, sI), \quad s \sim (1 + s)^{a-2}$$

dominates $\hat{\mu}_U$ and is proper Bayes when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Strawderman 1971)

- A Unifying Phenomenon: These domination results can be attributed to properties of the marginal distribution of X under π_H and π_a .

- The Bayes rule under $\pi(\mu)$ can be expressed as

$$\hat{\mu}_\pi(X) = E_\pi(\mu | X) = X + \nabla \log m_\pi(X)$$

where

$$m_\pi(X) \propto \int e^{-(X-\mu)^2/2} \pi(\mu) d\mu$$

is the marginal of X under $\pi(\mu)$. ($\nabla = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_p})'$)

(Brown 1971)

- The risk improvement of $\hat{\mu}_\pi(X)$ over $\hat{\mu}_U(X)$ can be expressed as

$$\begin{aligned} R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) &= E_\mu \left[(\nabla \log m_\pi(X))^2 - 2 \frac{\nabla^2 m_\pi(X)}{m_\pi(X)} \right] \\ &= E_\mu \left[-4 \frac{\nabla^2 \sqrt{m_\pi(X)}}{\sqrt{m_\pi(X)}} \right] \end{aligned}$$

($\nabla^2 = \sum_i \frac{\partial^2}{\partial x_i^2}$) (Stein 1974, 1981)

- That $\hat{\mu}_H(X)$ dominates $\hat{\mu}_U$ when $p \geq 3$, follows from the fact that the marginal $m_\pi(X)$ under π_H is superharmonic, i.e.

$$\nabla^2 m_\pi(X) \leq 0$$

- That $\hat{\mu}_a(X)$ dominates $\hat{\mu}_U$ when $p \geq 5$ (and conditions on a), follows from the fact that the sqrt of the marginal under π_a is superharmonic, i.e.

$$\nabla^2 \sqrt{m_\pi(X)} \leq 0$$

(Fourdrinier, Strawderman and Wells 1998)

2. The Prediction Problem

- Observe $X | \mu \sim N_p(\mu, v_x I)$ and predict $Y | \mu \sim N_p(\mu, v_y I)$
 - Conditionally on μ , Y is independent of X
 - v_x and v_y are known (for now)
- The Problem: To estimate $p(y | \mu)$ by $q(y | x)$.
- Measure closeness by Kullback-Leibler loss,

$$L(\mu, q(y | x)) = \int p(y | \mu) \log \frac{p(y | \mu)}{q(y | x)} dy$$

- Risk function

$$R_{KL}(\mu, \hat{p}) = \int L(\mu, q(y | x)) p(x | \mu) dx = E_\mu[L(\mu, q(y | X))]$$

3. Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule

$$p_\pi(y | x) = \int p(y | \mu)\pi(\mu | x)d\mu = E_\pi[p(y | \mu)|X]$$

minimizes $\int R_{KL}(\mu, \hat{p})\pi(\mu)d\mu$ (Aitchison 1975)

- Let $p_U(y | x)$ denote the Bayes rule under $\pi_U(\mu) \equiv 1$
- $p_U(y | x)$ dominates $p(y | \hat{\mu} = x)$, the naive “plug-in” predictive distribution (Aitchison 1975)
- $p_U(y | x)$ is best invariant and minimax with constant risk (Murray 1977, Ng 1980, Barron and Liang 2003)
- Shocking Fact: $p_U(y | x)$ is inadmissible when $p \geq 3$

- $p_H(y | x)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $p_U(y | x)$ when $p \geq 3$. (Komaki 2001).

- $p_a(y | x)$, the Bayes rule under $\pi_a(\mu)$ where

$$\mu | s \sim N_p(0, s v_0 I), \quad s \sim (1 + s)^{a-2},$$

dominates $p_U(y | x)$ and is proper Bayes when $v_x \leq v_0$ and when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Liang 2002)

- Main Question: Are these domination results attributable to the properties of m_π ?

4. A Key Representation for $p_\pi(y | x)$

- Let $m_\pi(x; v_x)$ denote the marginal of $X | \mu \sim N_p(\mu, v_x I)$ under $\pi(\mu)$.
- **Lemma:** The Bayes rule $p_\pi(y | x)$ can be expressed as

$$p_\pi(y | x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} p_U(y | x)$$

where

$$W = \frac{v_y X + v_x Y}{v_x + v_y} \sim N_p(\mu, v_w I)$$

- Using this, the risk improvement can be expressed as

$$\begin{aligned} R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) &= \int \int p_{v_x}(x | \mu) p_{v_y}(y | \mu) \log \frac{p_\pi(y | x)}{p_U(y | x)} dx dy \\ &= E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x) \end{aligned}$$

5. An Analogue of Stein's Unbiased Estimate of Risk

- **Theorem:**

$$\begin{aligned} \frac{\partial}{\partial v} E_{\mu, v} \log m_{\pi}(Z; v) &= E_{\mu, v} \left(\frac{\nabla^2 m_{\pi}(Z; v)}{m_{\pi}(Z; v)} - \frac{1}{2} \|\nabla \log m_{\pi}(Z; v)\|^2 \right) \\ &= E_{\mu, v} \left[2 \nabla^2 \sqrt{m_{\pi}(Z; v)} / \sqrt{m_{\pi}(Z; v)} \right] \end{aligned}$$

- Proof relies on using the heat equation

$$\frac{\partial}{\partial v} m_{\pi}(z; v) = \frac{1}{2} \nabla^2 m_{\pi}(z; v)$$

- Remark: This shows that the risk improvement in the quadratic risk estimation problem can be expressed in terms of $\log m_{\pi}$ as

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_{\pi}) = -2 \left[\frac{\partial}{\partial v} E_{\mu, v} \log m_{\pi}(Z; v) \right]_{v=1}$$

6. General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.
- **Theorem:** If $m_\pi(z; v)$ is finite for all z , then $p_\pi(y \mid x)$ will be minimax if either of the following hold:
 - (i) $\sqrt{m_\pi(z; v)}$ is superharmonic
 - (ii) $m_\pi(z; v)$ is superharmonic
- **Corollary:** If $m_\pi(z; v)$ is finite for all z , then $p_\pi(y \mid x)$ will be minimax if $\pi(\mu)$ is superharmonic
- $p_\pi(y \mid x)$ will dominate $p_U(y \mid x)$ in the above results if the superharmonicity is strict on some interval.

7. Sufficient Conditions for Admissibility

- **Theorem** (Blyth's Method): If there is a sequence of finite non-negative measures satisfying $\pi_n(\{\mu : \|\mu\| \leq 1\}) \geq 1$ such that

$$E_{\pi_n}[R_{KL}(\mu, q)] - E_{\pi_n}[R_{KL}(\mu, p_{\pi_n})] \rightarrow 0$$

then $q(y | x)$ is admissible.

- **Theorem:** For any two Bayes rules p_π and p_{π_n}

$$E_{\pi_n}[R_{KL}(\mu, p_\pi)] - E_{\pi_n}[R_{KL}(\mu, p_{\pi_n})] = \frac{1}{2} \int_{v_w}^{v_x} \int \frac{\|\nabla h_n(z; v)\|^2}{h_n(z; v)} m_\pi(z; v) dz dv$$

where $h_n(z; v) = m_{\pi_n}(z; v)/m_\pi(z; v)$.

- Using the explicit construction of $\pi_n(\mu)$ from Brown and Hwang (1984), we obtain tail behavior conditions that prove admissibility of $p_U(y | x)$ when $p \leq 2$, and admissibility of $p_H(y | x)$ when $p \geq 3$.

8. Minimax Shrinkage Towards 0

- Because π_H and $\sqrt{m_a}$ are superharmonic under suitable conditions, the result that $p_H(y | x)$ and $p_a(y | x)$ dominate $p_U(y | x)$ and are minimax follows immediately from the Theorem.
- By the Theorem, any of the improper superharmonic t-priors of Faith (1978) or any of the proper generalized t-priors of Fourdrinier, Strawderman and Wells (1998) yield Bayes rules that dominate $p_U(y | x)$ and are minimax.
- The risk functions $R_{KL}(\mu, p_H)$ and $R_{KL}(\mu, p_a)$ take on their minima at $\mu = 0$, and then asymptote up to $R_{KL}(\mu, p_U)$ as $\|\mu\| \rightarrow \infty$.

- Figure 1a displays the difference between the risk functions

$$[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_H)]$$

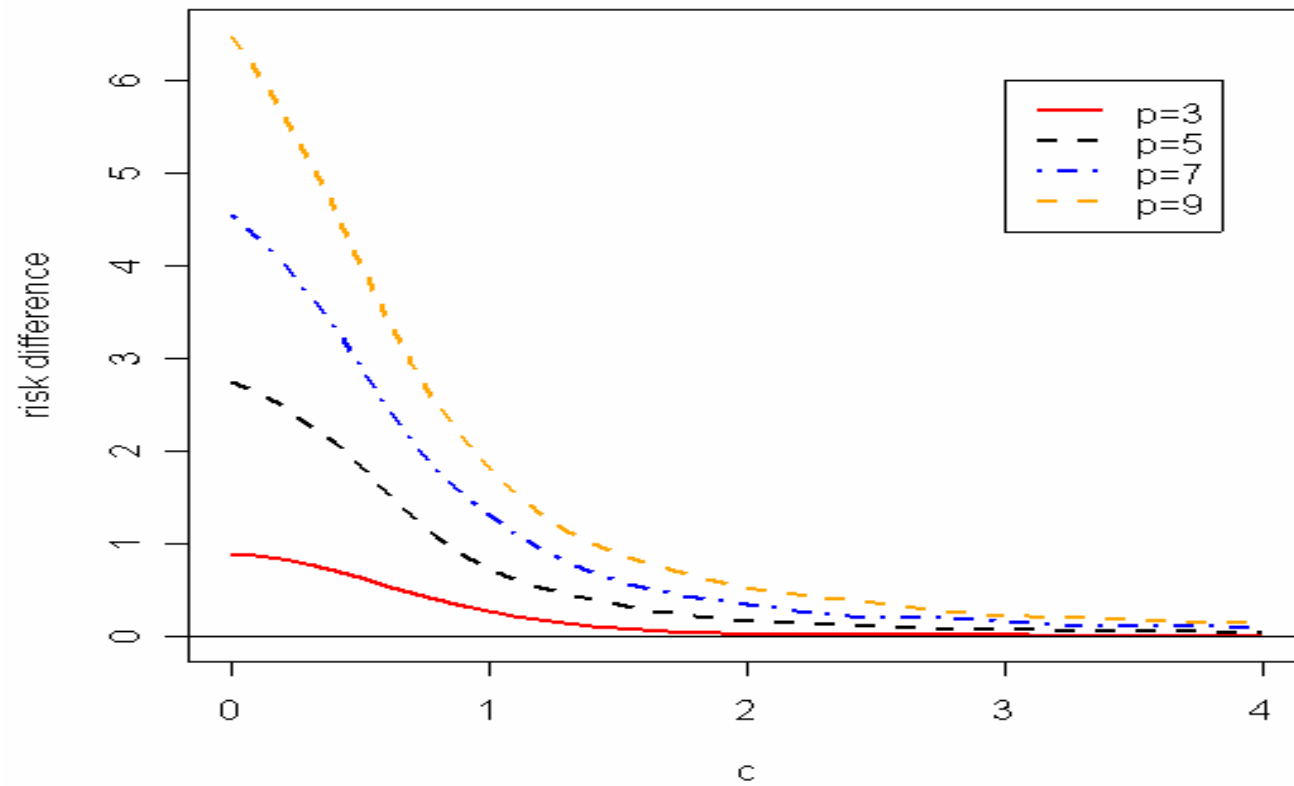
at $\mu = (c, \dots, c)'$, $0 \leq c \leq 4$ when $v_x = 1$ and $v_y = 0.2$ for dimensions $p = 3, 5, 7, 9$.

- Figure 1b displays the difference between the risk functions

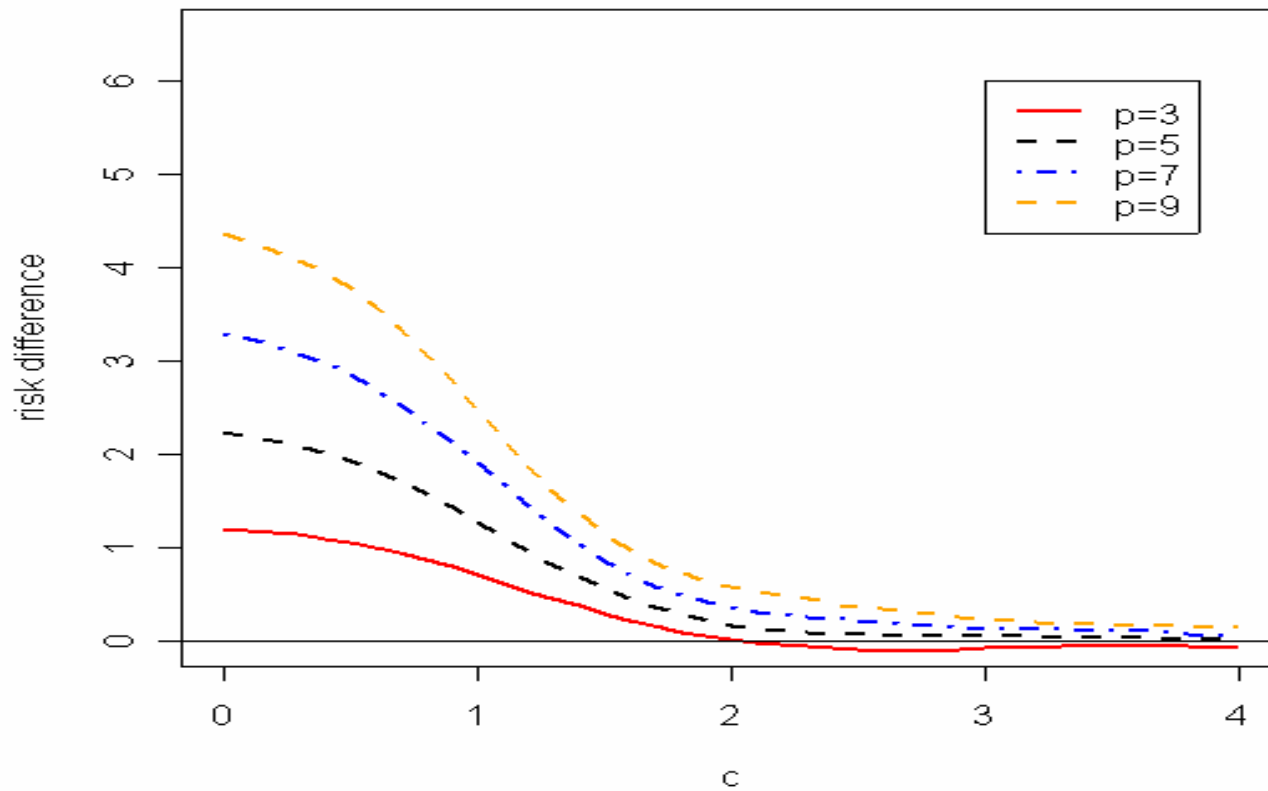
$$[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_a)]$$

at $\mu = (c, \dots, c)'$, $0 \leq c \leq 4$ when $a = 0.5$, $v_x = 1$ and $v_y = 0.2$ for dimensions $p = 3, 5, 7, 9$.

**Figure 1a. The risk difference between p_U and p_H : $R(\mu, p_U) - R(\mu, p_H)$.
Here $\theta = (c, \dots, c)$, $v_x = 1$, $v_y = 0.2$**



**Figure 1b. The risk difference between p_U and p_a with $a = 0.5$: $R(\mu, p_U) - R(\mu, p_a)$.
Here $\theta = (c, \dots, c)$, $v_x = 1$, $v_y = 0.2$**



- Our Lemma representation

$$p_H(y | x) = \frac{m_H(w; v_w)}{m_H(x; v_x)} p_U(y | x)$$

shows how $p_H(y | x)$ “shrinks $p_U(y | x)$ towards 0” by an adaptive multiplicative factor of the form

$$b_H(x, y) = \frac{m_H(w; v_w)}{m_H(x; v_x)}$$

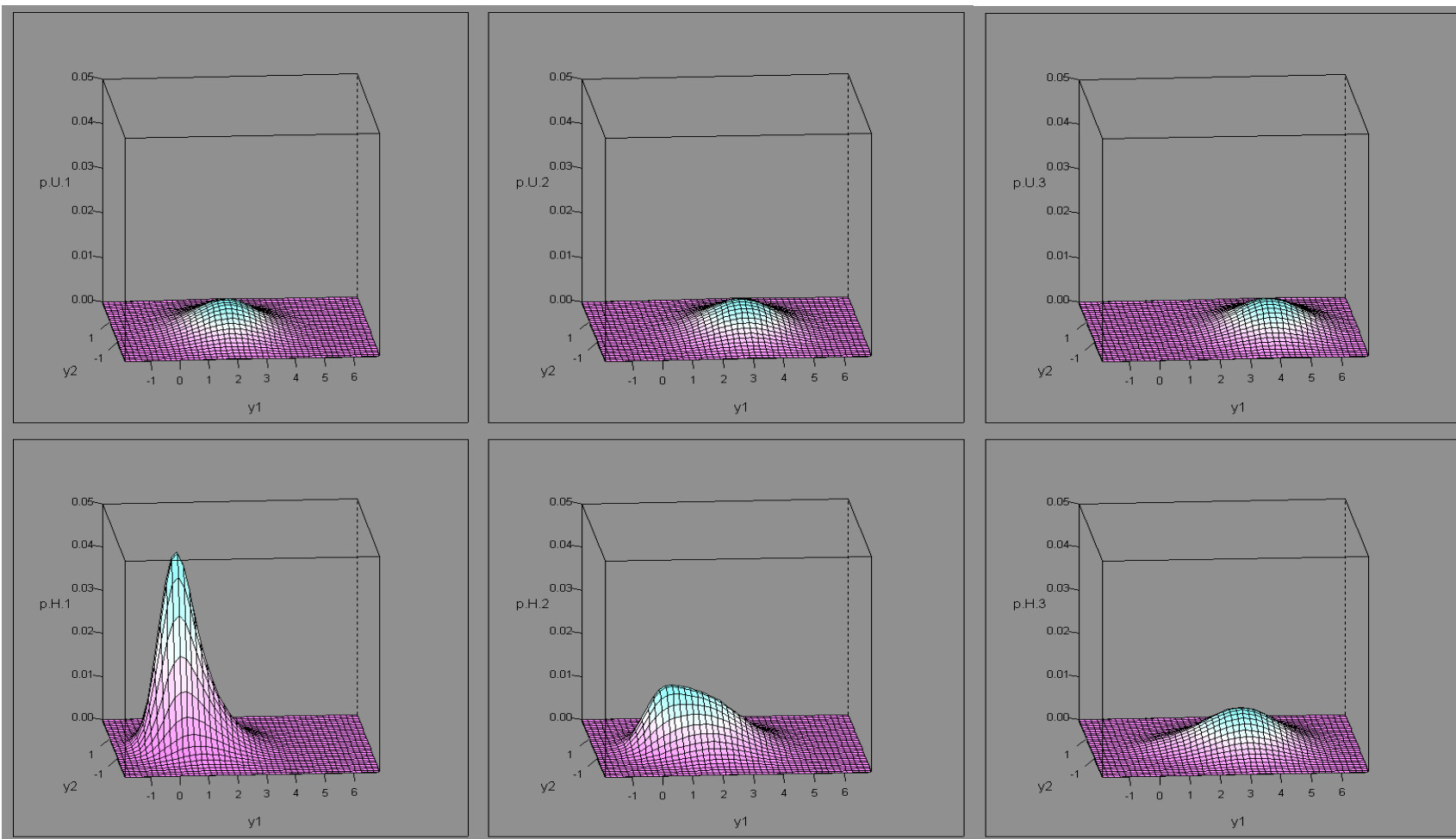
- Figure 2 illustrates how this shrinkage occurs for various values of x when $p = 5$.

Figure 2. Shrinkage of $p_U(y|x)$ to obtain $p_H(y|x)$ when $p = 5$. Here $y = (y_1, y_2, 0, 0, 0)$

$x = (2, 0, 0, 0, 0)$

$x = (3, 0, 0, 0, 0)$

$x = (4, 0, 0, 0, 0)$



9. Shrinkage Towards Points or Subspaces

- We can trivially modify the previous priors and predictive distributions to shrink towards an arbitrary point $b \in R^p$.
- Consider the recentered prior

$$\pi^b(\mu) = \pi(\mu - b)$$

and corresponding recentered marginal

$$m_\pi^b(z; v) = m_\pi(z - b; v).$$

- This yields a predictive distribution

$$p_\pi^b(y | x) = \frac{m_\pi^b(w; v_w)}{m_\pi^b(x; v_x)} p_U(y | x)$$

that now shrinks $p_U(y | x)$ towards b rather than 0.

- More generally, we can shrink $p_U(y | x)$ towards any subspace B of R^p whenever π , and hence m_π , is spherically symmetric.
- Letting $P_B z$ be the projection of z onto B , shrinkage towards B is obtained by using the recentered prior

$$\pi^B(\mu) = \pi(\mu - P_B \mu)$$

which yields the recentered marginal

$$m_\pi^B(z; v) := m_\pi(z - P_B z; v).$$

- This modification yields a predictive distribution

$$p_\pi^B(y | x) = \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)} p_U(y | x)$$

that now shrinks $p_U(y | x)$ towards B .

- If $m_\pi^B(z; v)$ satisfies any of the conditions of the Theorem, then $p_\pi^B(y | x)$ will dominate $p_U(y | x)$ and be minimax.

10. Minimax Multiple Shrinkage Prediction

- For any spherically symmetric prior, a set of subspaces B_1, \dots, B_N , and corresponding probabilities w_1, \dots, w_N , consider the recentered mixture prior

$$\pi_*(\mu) = \sum_{i=1}^N w_i \pi^{B_i}(\mu),$$

and corresponding recentered mixture marginal

$$m_*(z; v) = \sum_{i=1}^N w_i m_{\pi}^{B_i}(z; v).$$

- Applying the $\hat{\mu}_{\pi}(X) = X + \nabla \log m_{\pi}(X)$ construction with $m_*(X; v)$ yields minimax multiple shrinkage estimators of μ . (George 1986)

- Applying the predictive construction with $m_*(z; v)$ yields

$$p_*(y | x) = \sum_{i=1}^N p(B_i | x) p_{\pi}^{B_i}(y | x)$$

where $p_{\pi}^{B_i}(y | x)$ is a single target predictive distribution and

$$p(B_i | x) = \frac{w_i m_{\pi}^{B_i}(x; v_x)}{\sum_{i=1}^N w_i m_{\pi}^{B_i}(x; v_x)}$$

is the posterior weight on the i th prior component.

- **Theorem:** If each $m_{\pi}^{B_i}(z; v)$ is superharmonic, then $p_*(y | x)$ will dominate $p_U(y | x)$ and will be minimax.

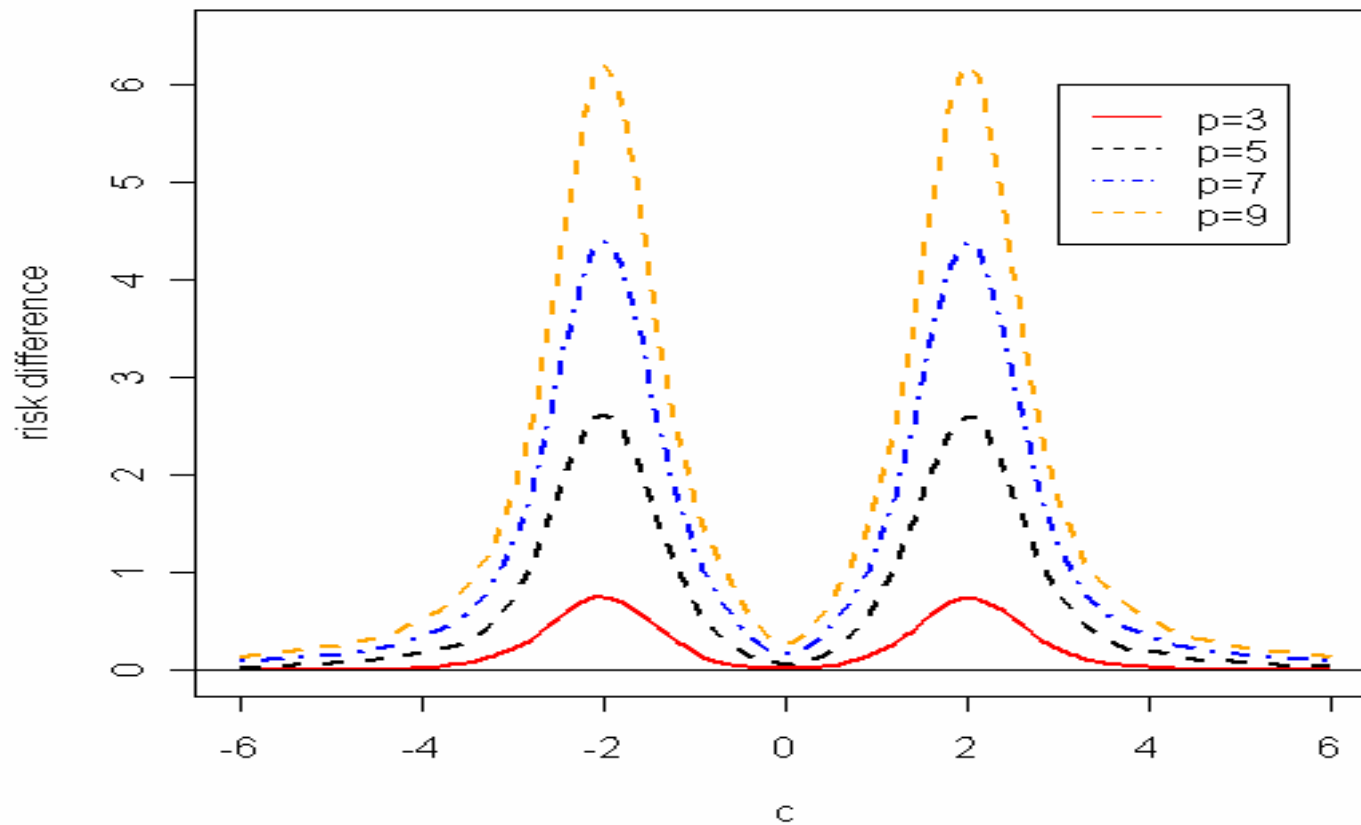
- Figure 3 illustrates the risk reduction

$$[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_{H^*})]$$

for $\mu = (c, \dots, c)'$ obtained by p_{H^*} which adaptively shrinks $p_U(y | x)$ towards the closer of the two points $b_1 = (2, \dots, 2)$ and $b_2 = (-2, \dots, -2)$ using equal weights $w_1 = w_2 = 0.5$

Figure 3. The risk difference between p_U and multiple shrinkage p_{H^*} : $R(\mu, p_U) - R(\mu, p_{H^*})$.

Here $\theta = (c, \dots, c)$, $v_x = 1$, $v_y = 0.2$, $a_1 = 2$, $a_2 = -2$, $w_1 = w_2 = 0.5$.



11. The Case of Unknown Variance

- If v_x and v_y are unknown, suppose there exists an available independent estimate of v_x of the form s/k where

$$S \sim v_x \chi_k^2.$$

Also assume that $v_y = r v_x$, for a known constant r .

- Substitute the estimates $\hat{v}_x = s/k$, $\hat{v}_y = rs/k$ and $\hat{v}_w = \frac{r}{r+1} s/k$ for v_x , v_y and v_w respectively.
- The predictor

$$p_\pi^*(y | x) = \frac{m_\pi(w; \hat{v}_w)}{m_\pi(x; \hat{v}_x)} p_U^*(y | x)$$

will still dominate $p_U^*(y | x)$ if any of the conditions of the Theorem are satisfied.

- Note however, $p_U^*(y | x)$ is no longer best invariant or minimax.

12. A Complete Class Theorem

- **Theorem:** In the KL risk problem, the class of all generalized Bayes procedures is a complete class.
- A (possibly randomized) decision procedure is a probability distribution $G(\cdot | x)$ for each x over the action space, namely the set of all densities $g(\cdot | x) : R^p \rightarrow R$ of Y . The Bayes rule under a prior π can then be denoted $G_\pi(\cdot | x) = \int p(y | \mu) p_\pi(\mu | x) d\mu$, which is a nonrandomized rule.
- The complete class result is proved by showing;
 - (i) If G is an admissible procedure, then it is non-randomized.
 - (ii) There exists a sequence of priors $\{\pi_i\}$ such that $G_{\pi_i}(\cdot | x) \rightarrow G(\cdot | x)$ weak* for a.e. x .
 - (iii) We can find a subsequence $\{\pi_{i'}\}$ of $\{\pi_i\}$ and a limiting prior π , which satisfy $\pi_{i'} \rightarrow \pi$ weak* and $G_{\pi_{i'}}(\cdot | x) \rightarrow G_\pi(\cdot | x)$ weak* for a.e. x . Therefore, $G(\cdot | x) = G_\pi(\cdot | x)$ for a.e. x , so that G is a generalized Bayes rule.

References For Getting Started

Brown, L.D., George, E.I. and Xu, X. (2005). Admissible Predictive Estimation. Working paper.

George, E.I., Liang, F. and Xu, X. (2005). Improved Minimax Predictive Densities under Kullback-Leibler Loss. *Annals of Statistics*, to appear.