Technical Report No. 0108, Department of Statistics, University of Toronto

Transferring Prior Information Between Models Using Imaginary Data

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
http://www.cs.utoronto.ca/~radford/
radford@stat.utoronto.ca

17 July 2001

Abstract. Bayesian modeling is limited by our ability to formulate prior distributions that adequately represent our actual prior beliefs — a task that is especially difficult for realistic models with many interacting parameters. I show here how a prior distribution formulated for a simpler, more easily understood model can be used to modify the prior distribution of a more complex model. This is done by generating imaginary data from the simpler "donor" model, which is conditioned on in the more complex "recipient" model, effectively transferring the donor model's well-specified prior information to the recipient model. Such prior information transfers are also useful when comparing two complex models for the same data. Bayesian model comparison based on the Bayes factor is very sensitive to the prior distributions for each model's parameters, with the result that the wrong model may be favoured simply because the prior for the right model was not carefully formulated. This problem can be alleviated by modifying each model's prior to potentially incorporate prior information transferred from the other model. I discuss how these techniques can be implemented by simple Monte Carlo and by Markov chain Monte Carlo with annealed importance sampling. Demonstrations on models for two-way contingency tables and on graphical models for categorical data show that prior information transfer can indeed overcome deficiencies in prior specification for complex models.

1 Introduction

One obstacle to Bayesian modeling of complex situations is that it can be difficult to formulate a prior distribution over a high-dimensional parameter space that is an adequate representation of our prior beliefs. We may sometimes find it easier to specify an appropriate prior distribution for the parameters of a simpler model, even though we believe that this model does not capture all the complexities of the real situation. In this paper, I show how the well-specified prior information incorporated into such a simple model can be transferred to the more complex model. Prior information can also be transferred between models of comparable complexity, evening out the quality of their prior specifications. This technique is especially useful when these models are to be compared using Bayes factors, since such comparisons are very sensitive to the priors on each model's parameters, and consequently can favour the wrong model merely because the right model's prior was not carefully formulated.

These transfers of prior information are accomplished by means of imaginary data that is generated using the well-specified prior distribution of the "donor" model. The "recipient" model has a less-well-specified *pro forma* prior — which might, for example, be overly-diffuse — but its real prior distribution is obtained by conditioning on observation of this imaginary data. In the "donor-weighted" variety of this procedure, the imaginary data is determined solely by the donor model's prior; in the "jointly-weighted" variety, the priors for both models are used. These varieties differ both in their statistical properties and in the contexts in which they are computationally feasible. I also consider limits of both procedures as the amount of imaginary data goes to infinity.

In interesting applications, these techniques are likely to be implementable only by Monte Carlo methods. Simple Monte Carlo based on sampling data from the donor model's prior is adequate, but it becomes extremely inefficient for higher-dimensional models. Prior information transfer for complex models will therefore usually require use of Markov chain Monte Carlo methods. When the marginal likelihood (the posterior normalizing constant) needs to be computed in order to compare models, Markov chain sampling will usually have to be combined with a technique such as annealed importance sampling (Neal 2001), which uses a series of distributions that link to a distribution with a known normalizing constant. These computational methods are discussed in Section 3 and illustrated in Section 4 on two example problems.

The first example shows how prior information transfer can be used in a model for two-way contingency tables. Here, the simple model assumes independence, and uses informative priors for the marginal distributions. The more complex model is unrestricted. Directly formulating an appropriate prior for this unrestricted model that captures our prior beliefs might be difficult. Prior information transfer will be appropriate if these beliefs can be adequately approximated by taking from the simple model both a preference for distributions exhibiting some degree of approximate independence and the simple model's prior information regarding marginal probabilities. In a simulated example where the true distribution does exhibit approximate independence, I find that predictive performance is indeed improved by using prior information transfer from the simple independence model.

As a second example, I show how prior information transfer can be used to ensure a fair comparison of two graphical models for data on the educational aspirations of high school students. These alternative models were previously considered by Heckerman, Meek, and Cooper (1999), who found that the Bayes factor strongly favoured one model over the other. I found this puzzling, since the favoured model seems much less plausible. However, when each model is allowed to receive prior information transferred from the other, the strong preference for the less plausible model disappears. Transfer of prior information has here corrected a Bayes factor that does not reflect the real merits of the models, but only the differential effects of bad prior specifications.

I conclude by outlining some areas for future research, such as how to transfer information amongst multiple models. I also discuss how prior information transfer relates to previous attempts at overcoming problems of prior specification and model comparison.

For general background on prior distributions and Bayes factors, see Bernardo and Smith (1994) and Kass and Raftery (1995).

2 Techniques for prior information transfer

I will consider only models under which the data is exchangeable — that is, conditional on values for the model parameters, the data items are independent and identically distributed. We are ultimately interested in defining a prior distribution for a "recipient" model, whose parameters will be denoted by θ. We have a *pro forma* prior distribution for this model, whose density will be written as $P(\theta)$, but we are not confident that this distribution is an adequate representation of our prior beliefs. It may sometimes be permissible for the *pro forma* prior to be improper.

We also have a "donor" model, whose parameters will be denoted by ϕ. Although we doubt that this model is an adequate representation of reality, we believe that it may have some approximate validity, and on this basis, we have formulated a prior distribution, $Q(\phi)$, for this model's parameters, expressing our beliefs about which values of ϕ are more likely to produce a good approximation to reality. We hope to transfer the prior information incorporated into this model to the recipient model, using imaginary data generated from the donor model. The prior probability of data $y_1, \ldots, y_k$ under the donor model is

$$Q(y_1, \ldots, y_k) \;=\; \int Q(y_1, \ldots, y_k \,|\, \phi)\, Q(\phi)\, d\phi \;=\; \int \Big[\prod_{i=1}^{k} Q(y_i \,|\, \phi)\Big]\, Q(\phi)\, d\phi \qquad (1)$$

Here, $Q(y_i \,|\, \phi)$ gives the probability of a single data item under the donor model, conditional on model parameters ϕ.

2.1 Donor-weighted transfer of prior information

The "donor-weighted" prior for the recipient model based on k imaginary data points, denoted by $\dot{P}_k$, is defined as follows:

$$\dot{P}_k(\theta) \;=\; \sum_{y_1,\ldots,y_k} P(\theta \,|\, y_1, \ldots, y_k)\, Q(y_1, \ldots, y_k) \qquad (2)$$

Here, $P(\theta \,|\, y_1, \ldots, y_k)$ is the posterior density for θ given imaginary data $y_1, \ldots, y_k$, based on the recipient model's *pro forma* prior. The sum above is over all possible values for these k imaginary data points. We can also write this prior in the following equivalent way:

$$\dot{P}_k(\theta) \;=\; \sum_{y_1,\ldots,y_k} \frac{P(y_1, \ldots, y_k \,|\, \theta)\, P(\theta)}{\int P(y_1, \ldots, y_k \,|\, \theta)\, P(\theta)\, d\theta}\, Q(y_1, \ldots, y_k) \qquad (3)$$

$$\;=\; \sum_{y_1,\ldots,y_k} \frac{\Big[\prod_{i=1}^{k} P(y_i \,|\, \theta)\Big]\, P(\theta)}{\int \Big[\prod_{i=1}^{k} P(y_i \,|\, \theta)\Big]\, P(\theta)\, d\theta}\, Q(y_1, \ldots, y_k) \qquad (4)$$

The number, k, of imaginary data points affects the prior in two ways. A larger value of k produces a larger modification of the *pro forma* prior, as a result of conditioning on more data points. A larger value of k also reduces the variability in the effect of the imaginary data, since the Law of Large Numbers makes most large data sets have almost equivalent effect. To separate these two influences, we can generalize the prior by introducing a parameter ω that determines the effective mass of the imaginary data:

$$\dot{P}_k^\omega(\theta) \;=\; \sum_{y_1,\ldots,y_k} \frac{\left[\prod_{i=1}^{k} P(y_i \,|\, \theta)\right]^{\omega/k} P(\theta)}{\int \left[\prod_{i=1}^{k} P(y_i \,|\, \theta)\right]^{\omega/k} P(\theta)\, d\theta}\; Q(y_1,\ldots,y_k) \tag{5}$$

When $\omega = k$, this prior is equivalent to that of equations (2) and (4). When k is large but ω is small, the *pro forma* prior is changed only to a small degree, but this change is based on many "fractional" data points, rather than on a few randomly selected data points. If this is desirable, it may be best to use the prior obtained in the limit as the amount of imaginary data goes to infinity, while ω stays constant:

$$\dot{P}^\omega(\theta) \;=\; \lim_{k\rightarrow\infty} \dot{P}_k^\omega(\theta) \tag{6}$$

It may also sometimes be sensible to let the mass of imaginary data go to infinity along with the number of imaginary data points, producing the following prior:

$$\dot{P}_\infty(\theta) \;=\; \lim_{k\rightarrow\infty} \dot{P}_k(\theta) \tag{7}$$

This will not be useful when the donor model is nested within the recipient model, since conditioning on an infinite mass of imaginary data generated from such a donor model will have the effect of forcing the recipient model to behave in exactly the same way as the donor model. However, when the two models are not nested, an infinite mass of imaginary data produces the interesting effect of transforming the prior for the donor model's parameters to the prior for the recipient model that is obtained by mapping the donor parameters to the values of the recipient model's parameters that are closest, in terms of likelihood. The recipient model's *pro forma* prior has no effect on the result (except to eliminate parameter values that are not within its support). As a trivial example, consider two models for i.i.d. pairs of Bernoulli variables, (X, Y). In the donor model, $P(X=1 \,|\, \phi) = \phi$ and $P(Y = 1 \,|\, \phi) = 1/2$. In the recipient model, $P(X = 1 \,|\, \theta) = P(Y = 1 \,|\, \theta) = \theta$. The posterior for θ given an infinite mass of imaginary data generated with a particular ϕ will be concentrated at the maximum likelihood value of $\theta = 1/4 + \phi/2$. If $Q(\phi)$ is uniform over $(0, 1)$, and $P(\theta)$ is any distribution with full support, $\dot{P}_\infty(\theta)$ will be uniform on $(1/4,\ 3/4)$.

Such a complete transfer of the donor model's prior to the recipient model may sometimes be useful, though I think that the fuzzier effect of a finite mass of imaginary data will usually be preferable. When the donor model has fewer parameters than the recipient model, $\dot{P}_\infty$ will be concentrated on a lower-dimensional sub-manifold of the recipient model's parameter space. This may also be useful at times, although it does effectively eliminate some of the recipient model's flexibility.

In general, priors based on donor-weighted transfer of information can be viewed as mixtures of distributions obtained by conditioning the *pro forma* prior on the various possible values for the imaginary data. The effect of the imaginary data can be to narrow the prior distribution, if the data the donor model is likely to produce is a subset of that which would be produced by the recipient model under its *pro forma* prior. It is also possible, however, for the imaginary data to cause a shift or a widening in the prior. This can occur if the donor model gives substantial probability to data whose probability under the *pro forma* prior is very small (though not zero). In the mixture defining $\dot{P}$, imaginary data that is unlikely under P will nevertheless be included, with weight given by its probability under Q, forcing the final prior to place an appreciable probability on values of θ that are compatible with such data.

Computationally, weighting imaginary data according to the donor model's prior requires evaluation of normalizing constants of posterior distributions under P. For example, simple Monte Carlo estimation of (4) or (5) based on random sampling of $y_1, \ldots, y_k$ from Q will require evaluation of the integrals appearing in these expressions. For the models with conjugate priors used in the examples of Section 4, these integrals can be done analytically, but this will not be true in general.

2.2 Jointly-weighted transfer of prior information

An alternative, "jointly-weighted" form of prior information transfer is interesting both because it has different statistical properties, and because it will sometimes be easier to implement from a computational standpoint.

The jointly-weighted prior using k imaginary data points, denoted by $\ddot{P}_k$, can be written as the following proportionality:

$$\ddot{P}_k(\theta) \;\propto\; \sum_{y_1,\ldots,y_k} P(\theta \,|\, y_1,\ldots,y_k)\, P(y_1,\ldots,y_k)\, Q(y_1,\ldots,y_k) \tag{8}$$

Comparing this with equation (2), we see that it differs in that the imaginary data is weighted by the product of its probability under P and its probability under Q. Also, to produce a normalized prior for θ, the expression on the right must be divided by the sum of these weights, $\sum P(y_1,\ldots,y_k)\, Q(y_1,\ldots,y_k)$. We can write this prior equivalently as follows:

$$\begin{aligned} \ddot{P}_k(\theta) \;&\propto\; P(\theta) \sum_{y_1,\ldots,y_k} P(y_1,\ldots,y_k \,|\, \theta)\, Q(y_1,\ldots,y_k) & (9) \\ &\propto\; P(\theta) \sum_{y_1,\ldots,y_k} \Big[\prod_{i=1}^{k} P(y_i \,|\, \theta)\Big]\, Q(y_1,\ldots,y_k) & (10) \end{aligned}$$

As for donor-weighted transfer, we can generalize this by introducing a parameter ω for the effective mass of the imaginary data:

$$\ddot{P}_k^{\omega}(\theta) \;\propto\; P(\theta) \sum_{y_1,\ldots,y_k} \Big[\prod_{i=1}^{k} P(y_i \,|\, \theta)\Big]^{\omega/k} Q(y_1,\ldots,y_k) \tag{11}$$

and we can consider the limit of this as the amount of imaginary data goes to infinity, with its effective mass kept constant:

$$\ddot{P}^{\omega}(\theta) \;=\; \lim_{k\to\infty} \ddot{P}_k^{\omega}(\theta) \tag{12}$$

As before, we might also consider letting the mass of imaginary data go to infinity:

$$\ddot{P}_\infty(\theta) \;=\; \lim_{k\to\infty} \ddot{P}_k(\theta) \tag{13}$$

However, this prior seems to have stranger and less desirable properties than the corresponding donor-weighted prior of equation (7). In the example of pairs of Bernoulli variables following (7), if $Q(\phi)$ and $P(\theta)$ are both uniform on $(0, 1)$, direct calculation shows that the resulting $\ddot{P}_\infty(\theta)$ assigns probability $1/2$ to each of the two extreme values of $\theta = 0$ and $\theta = 1$, with zero probability for other values. This is not useful, and not what one would intuitively expect.

With jointly-weighted transfer of prior information, we might expect that it will be less likely that the donor model will induce a prior for the recipient model that is wider than its *pro forma* prior, or that is shifted with respect to it, since weights that are products of probabilities under both models will tend to be high only for data that is of high probability under both models. However, it could be that no data has high prior probability under both models, in which case the resulting prior could put high probability on parameter values that produce the best compromise. Which method for transferring prior information produces better results will depend on the actual situation, but it does seem that the prior resulting from jointly-weighted transfer is harder to visualize than that resulting from donor-weighted transfer. When the *pro forma* prior is diffuse, however, it may be that the two methods produce similar results, in which case one might choose between them according to computational convenience.

When $P(y_1, \ldots, y_k)$ is easily computed, it seems easier to use donor-weighted transfer, since this avoids the need to compute the normalizing constant for (8). However, computing $P(y_1, \ldots, y_k)$ may sometimes be very difficult, in which case donor-weighted transfer may be infeasible, but jointly-weighted transfer may still be possible using Markov chain Monte Carlo methods to sample from the posterior distribution for θ. These issues are discussed in Section 3.

2.3 Idempotence of prior information transfer

The donor-weighted prior of equation (2) has a pleasing idempotence property — if the donor model and the recipient model are identical, the transfer of prior information has no effect. In other words, if Q and P are identical, then $\dot{P}_k$ is the same as P. This is easily seen as follows:

$$\dot{P}_k(\theta) \;=\; \sum_{y_1,\ldots,y_k} P(\theta \,|\, y_1, \ldots, y_k)\, Q(y_1, \ldots, y_k) \tag{14}$$

$$\;=\; \sum_{y_1,\ldots,y_k} \frac{P(\theta,\, y_1, \ldots, y_k)}{P(y_1, \ldots, y_k)}\, P(y_1, \ldots, y_k) \tag{15}$$

$$\;=\; \sum_{y_1,\ldots,y_k} P(y_1, \ldots, y_k \,|\, \theta)\, P(\theta) \;=\; P(\theta) \tag{16}$$

This idempotence property need not hold when the effective mass of prior information is adjusted to a value other than k, as in equation (5), including when k is infinite as in equation (6). This occurs, for example, with models for Bernoulli data when both P and Q use a uniform prior for the unknown probability — direct calculation shows that $\dot{P}_1^{1/2}$ and $\dot{P}^1$ are both non-uniform in this case. The idempotence property also need not hold for jointly-weighted priors — for the Bernoulli example, direct calculation shows that $\ddot{P}_2$ and $\ddot{P}^1$ are both non-uniform. This lack of idempotence should make us cautious about drawing conclusions about these priors on the basis of intuitive ideas about information transfer, but it does not seem to me to be reason enough to rule out use of such priors in circumstances where they appear to be the most expedient means of expressing our prior beliefs.

The idempotence of donor-weighted prior information transfer with $\omega = k$ preserves a modularity property that is useful when comparing graphical models, as in the example of Section 4.2. Suppose that each data item consists of two parts, $x_i = (x_i', x_i'')$. One of the models has two parameters, θ and ψ, and a likelihood that can be written as $P(x_i \,|\, \theta, \psi) = P(x_i' \,|\, \theta)\, P(x_i'' \,|\, x_i', \psi)$; the other model has parameters ϕ and ψ, and a likelihood that can be written as $Q(x_i \,|\, \phi, \psi) = Q(x_i' \,|\, \phi)\, Q(x_i'' \,|\, x_i', \psi)$. Suppose that the likelihood factors involving x'' are identical for the two models — ie, $P(x_i'' \,|\, x_i', \psi) = Q(x_i'' \,|\, x_i', \psi)$. Suppose further that θ and ψ are independent in the first model's prior, ϕ and ψ are independent in the second model's prior, and the priors for ψ in the two models are the same — ie, $P(\psi) = Q(\psi)$. These models can be seen as containing a common component, parameterized by ψ, which models the conditional distribution of x'' given x'. Since this component is the same in the two models, it does not affect the Bayes factor for comparing them. The marginal likelihood for each model will be a product of a factor pertaining to ψ, which is the same for both models, and a factor particular to that model, which is the same as the marginal likelihood that would have been found by considering only the first part of the data, x', and only the first parameter, θ or ϕ, ignoring x'' and ψ.

This ability to ignore x'' and ψ when comparing the models will be preserved when each model's prior is modified using donor-weighted transfer from the other model with $\omega = k$. In these modified priors, ψ will still be independent of the other parameter, and will have the same prior as before. This can be seen as follows for transfer from Q to P (using y as shorthand for $y_1, \ldots, y_k$):

$$\begin{aligned}
\dot{P}_k(\psi, \theta) &= \sum_{y', y''} P(\psi, \theta \,|\, y', y'')\, Q(y', y'') && (17)\\
&= \sum_{y', y''} P(\theta \,|\, y')\, P(\psi \,|\, y', y'')\, Q(y') \int Q(y'' \,|\, y', \tilde{\psi})\, Q(\tilde{\psi})\, d\tilde{\psi} && (18)\\
&= \sum_{y', y''} P(\theta \,|\, y')\, \frac{P(\psi)\, P(y'' \,|\, y', \psi)}{\int P(y'' \,|\, y', \tilde{\psi})\, P(\tilde{\psi})\, d\tilde{\psi}}\, Q(y') \int P(y'' \,|\, y', \tilde{\psi})\, P(\tilde{\psi})\, d\tilde{\psi} && (19)\\
&= \sum_{y', y''} P(\theta \,|\, y')\, P(\psi)\, P(y'' \,|\, y', \psi)\, Q(y') && (20)\\
&= P(\psi) \sum_{y'} P(\theta \,|\, y')\, Q(y') \sum_{y''} P(y'' \,|\, y', \psi) && (21)\\
&= P(\psi) \sum_{y'} P(\theta \,|\, y')\, Q(y') && (22)
\end{aligned}$$

The sum in the last expression has the form of a prior obtained by donor-weighted transfer using models that ignore the second component of the data. We therefore see that with this form of prior information transfer, model comparison is not affected by the presence of extra data that is treated identically by the two models.

Unfortunately, this property is not shared by donor-weighted transfer with $\omega \neq k$, or by jointly-weighted transfer. When using these methods, it seems most reasonable to reduce the two models by eliminating any parts such as x'' that they have in common, in order to avoid any influence these arguably irrelevant aspects of the data might have on the answer. This will reduce the computational cost as well.

2.4 Controlling the amount of information transferred

Since the degree to which prior information in the donor model is actually helpful will generally be hard to determine *a priori*, we should usually not fix the amount of information transferred. Perhaps the only exception to this is when we have some reason to think that an infinite mass of imaginary data would be appropriate, as in (7). When we don't know exactly how much information should be transferred, we can specify a fairly vague prior distribution for the number of imaginary data points, k, or for the mass parameter, ω, and then allow the data to determine a posterior distribution for how much prior information is transferred. I expect that the best procedure will typically be to fix k to as large a value as is feasible computationally (letting it be infinite if possible), and to then use a prior on ω to control the amount of information transferred. Conceivably, however, there could be situations in which the variability present when k is small is beneficial, in which case we might instead specify a prior distribution for k and fix ω to be equal to k, or specify some joint prior for k and ω.

3 Monte Carlo implementations of prior information transfer

Prior distributions obtained by donor-weighted or jointly-weighted transfer of prior information, such as (2) and (8), will seldom be reducible to easily computable formulas. We will instead need to resort to Monte Carlo methods. To predict future observations, or estimate the posterior expectation of some function of interest, we need some way of sampling from the posterior distribution of the model parameters, θ, based on observed data $x_1, \ldots, x_n$. If we are comparing two models, we also need to estimate the marginal likelihoods of these models based on the observed data, the ratio of which gives the Bayes factor.

For some examples in this paper, I use simple Monte Carlo estimation based on sampling from the donor model's prior, but in other cases Markov chain Monte Carlo methods are needed. Below, I discuss implementation by these two methods when the mass of imaginary data is fixed, after which I discuss the problem of letting the mass parameter, ω, be variable.

3.1 Implementation using simple Monte Carlo

A simple Monte Carlo approach is possible when the donor model allows easy sampling from its prior, and the recipient model, with its *pro forma* prior, allows easy sampling from its posterior and efficient computation of marginal likelihoods. In particular, the donor model must allow easy sampling from the prior for its parameters, $Q(\phi)$, and from the distribution for data given parameters, $Q(y_i \,|\, \phi)$. The recipient model with its *pro forma* prior must allow easy sampling from posterior distributions, such as $P(\theta \,|\, y_1, \ldots, y_k)$, and efficient computation of marginal likelihoods, such as $P(y_1, \ldots, y_k) = \int P(y_1, \ldots, y_k \,|\, \theta)\, P(\theta)\, d\theta$. This will usually be possible only when the *pro forma* prior, $P(\theta)$, is conjugate.

The simple Monte Carlo estimates will be based on N sets of k imaginary data points, with the j'th such set denoted by $y_1^{(j)}, \ldots, y_k^{(j)}$. These data sets are drawn independently from the donor model's prior, typically by first drawing $\phi^{(j)}$ from $Q(\phi)$ and then drawing each $y_i^{(j)}$ for i from 1 to k from $Q(y_i \,|\, \phi^{(j)})$.

Using such a sample of imaginary data sets, we could estimate the prior of equation (2), based on donor-weighted transfer, as follows:

$$\dot{P}_k(\theta) \;\approx\; \frac{1}{N} \sum_{j=1}^{N} P(\theta \,|\, y_1^{(j)}, \ldots, y_k^{(j)}) \tag{23}$$

This is usually not of direct interest, but it forms the basis for the following estimate of the marginal likelihood for actual observations $x_1, \ldots, x_n$:

$$\begin{aligned}
\dot{P}_k(x_1, \ldots, x_n) &= \int P(x_1, \ldots, x_n \,|\, \theta)\, \dot{P}_k(\theta)\, d\theta & (24)\\
&\approx \frac{1}{N} \sum_{j=1}^{N} \int P(x_1, \ldots, x_n \,|\, \theta)\, P(\theta \,|\, y_1^{(j)}, \ldots, y_k^{(j)})\, d\theta & (25)\\
&= \frac{1}{N} \sum_{j=1}^{N} P(x_1, \ldots, x_n \,|\, y_1^{(j)}, \ldots, y_k^{(j)}) & (26)\\
&= \frac{1}{N} \sum_{j=1}^{N} \frac{P(x_1, \ldots, x_n,\, y_1^{(j)}, \ldots, y_k^{(j)})}{P(y_1^{(j)}, \ldots, y_k^{(j)})} & (27)
\end{aligned}$$

By assumption, the marginal likelihoods in this last expression are easily computable. Alternatively, when the *pro forma* prior is conjugate, conditioning on imaginary data simply alters this conjugate prior's parameters, so each term of the sum in (26) above can be found by calculating the marginal likelihood for the observed data under such an altered prior.

The posterior distribution for θ given the observed data can be estimated as follows:

$$\begin{aligned}
\dot{P}_k(\theta \,|\, x_1, \ldots, x_n) &= \frac{P(x_1, \ldots, x_n \,|\, \theta)\, \dot{P}_k(\theta)}{\dot{P}_k(x_1, \ldots, x_n)} & (28)\\
&\approx \frac{1}{N} \sum_{j=1}^{N} \frac{P(x_1, \ldots, x_n \,|\, \theta)\, P(\theta \,|\, y_1^{(j)}, \ldots, y_k^{(j)})}{\dot{P}_k(x_1, \ldots, x_n)} & (29)\\
&= \frac{1}{N} \sum_{j=1}^{N} \frac{P(x_1, \ldots, x_n \,|\, y_1^{(j)}, \ldots, y_k^{(j)})}{\dot{P}_k(x_1, \ldots, x_n)}\, P(\theta \,|\, x_1, \ldots, x_n,\, y_1^{(j)}, \ldots, y_k^{(j)}) & (30)
\end{aligned}$$

For the tractable models assumed here, an estimate of the posterior density for some value of θ can computed using this expression. The posterior expectation of a function of θ can be estimated as well, assuming that the posterior expectation of this function using the *pro forma* prior is easily computable. These estimators have the same form as arises with importance sampling and with regenerative simulation, so their accuracy can be assessed by the methods used in those contexts (Ripley 1987, Section 6.4; Geweke 1989; Neal 2001). If a value for θ drawn from the posterior distribution is desired, it can be obtained by choosing a value for j with probabilities given by $P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_k^{(j)}) \,/\, \dot{P}_k(x_1,\ldots,x_n)$, and then picking θ from the posterior distribution based on the *pro forma* prior and data $x_1,\ldots,x_n,\, y_1^{(j)},\ldots,y_k^{(j)}$.

Similar methods can be applied when using the donor-weighted prior of equation (5), in which the effective mass of the imaginary data is adjusted to be ω. When the recipient model is in the regular exponential family and the *pro forma* prior is conjugate, conditioning on imaginary data takes the form of updating the conjugate prior's parameters by adding the appropriate sufficient statistics computed from the imaginary data (Bernardo and Smith 1994, Section 5.2). To take a trivial example, for a model for Bernoulli data that uses a beta(α,β) *pro forma* prior for θ, the unknown probability of a 1, the distribution conditional on imaginary data is also beta, with parameters found by adding the number of imaginary 1's to α and the number of imaginary 0's to β. To adjust the effective mass of the imaginary data to ω, we need only multiply these imaginary data statistics by ω/k when computing the conditional probabilities in equations (23) and (26) and the two conditional probabilities in equation (30).

Letting k go to infinity, giving the prior of equation (6), is also feasible for such models. Rather than generate N imaginary data sets, we generate N values for the parameters of the donor model, $\phi^{(1)},\ldots,\phi^{(N)}$, and where we would condition on imaginary data by adding appropriate statistics to the parameters of the conjugate *pro forma* prior, we instead add the expected value of these statistics for one data point (with respect to the distribution defined by $\phi^{(j)}$), multiplied by the desired effective mass, ω. For the Bernoulli example above, with ϕ and θ both being the unknown probability of a 1, the expected number of 1's in an imaginary data set with one observation is $\phi^{(j)}$, and the expected number of 0's is $1-\phi^{(j)}$, so the conditional distribution for θ given a mass ω of imaginary data is beta$\,(\alpha{+}\omega\phi^{(j)},\, \beta{+}\omega(1-\phi^{(j)}))$.

The prior of equation (8), based on jointly-weighted transfer, can be estimated as follows:

$$\ddot{P}_k(\theta) \;\approx\; \frac{\dfrac{1}{N}\sum\limits_{j=1}^{N} P(\theta \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})}{\dfrac{1}{N}\sum\limits_{j=1}^{N} P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{31}$$

(Here and later, the two factors of $1/N$ will cancel, but they have been retained for clarity.) From this, we get an estimate of the marginal likelihood for the observed data:

$$\ddot{P}_k(x_1,\ldots,x_n) \;=\; \int P(x_1,\ldots,x_n \,|\, \theta)\, \ddot{P}_k(\theta)\, d\theta \tag{32}$$

$$\approx\; \frac{\dfrac{1}{N}\sum\limits_{j=1}^{N} \int P(x_1,\ldots,x_n \,|\, \theta)\, P(\theta \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})\, d\theta}{\dfrac{1}{N}\sum\limits_{j=1}^{N} P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{33}$$

$$= \frac{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})}{\frac{1}{N}\sum_{j=1}^{N} P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{34}$$

$$= \frac{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n,\, y_1^{(j)},\ldots,y_k^{(j)})}{\frac{1}{N}\sum_{j=1}^{N} P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{35}$$

Comparing (34) and (35) above with with (26) and (27) earlier gives some more insight into how donor-weighted and jointly-weighted prior information transfers differ. In particular, one can see that the marginal likelihoods based on the two priors will be the same if $P(x_1,\ldots,x_n \,|\, y_1,\ldots,y_k)$ and $P(y_1,\ldots,y_k)$ are independent under $Q(y_1,\ldots,y_k)$.

The posterior distribution for θ using the jointly-weighted prior can be estimated using equations (31) and (34), as follows:

$$\ddot{P}_k(\theta \,|\, x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n \,|\, \theta)\, \ddot{P}_k(\theta)}{\ddot{P}_k(x_1,\ldots,x_n)} \tag{36}$$

$$\approx \frac{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n \,|\, \theta)\, P(\theta \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})}{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{37}$$

$$= \frac{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})\, P(\theta \,|\, x_1,\ldots,x_n,\, y_1^{(j)},\ldots,y_k^{(j)})}{\frac{1}{N}\sum_{j=1}^{N} P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_k^{(j)})\, P(y_1^{(j)},\ldots,y_k^{(j)})} \tag{38}$$

To use the prior of equation (11), which is based on jointly-weighted information transfer using imaginary data that is adjusted to have effective mass ω, the probabilities conditional on $y_1^{(j)},\ldots,y_k^{(j)}$ in (31), (34), and (38) must be adjusted appropriately. As for donor-weighted transfer, this can easily be done for regular exponential family models when the *pro forma* prior is conjugate — the statistics added to the parameters of the conjugate prior are simply multiplied by ω/k. In addition, the occurrences of $P(y_1^{(j)},\ldots,y_k^{(j)})$ in these equations must be replaced by the analogue of marginal likelihood for the adjusted mass of data, which is $\int P(y_1^{(j)},\ldots,y_k^{(j)} \,|\, \theta)^{\omega/k} P(\theta)\, d\theta$. Use of the prior of equation (12), obtained by letting k go to infinity, will generally also be feasible for tractable models with conjugate priors.

Many models and priors do not permit the easy computation of marginal likelihoods and posterior distributions needed to apply these simple Monte Carlo methods. Even when the required quantities can be calculated, the efficiency of the Monte Carlo estimates will sometimes be extremely poor, because estimates such as that of equation (26) may be dominated by a very small fraction of the sample, with the result that the estimate will have a high variance. This will be the case when a few of the imaginary data sets, $y_1^{(j)},\ldots,y_k^{(j)}$, lead to much higher probabilities of the actual data, $x_1, \ldots, x_n$, than do the others, which will often be the case when the amount of imaginary data is large. Markov chain Monte Carlo methods will sometimes provide a way around these problems.

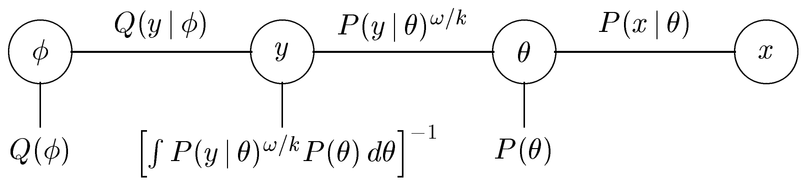
3.2 Implementation using Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods can be used to make posterior inferences about the recipient model's parameters, and to predict future observations. With somewhat more difficulty, the marginal likelihoods needed for model comparison can also be found. See Gilks, *et al* (1996) for general background on MCMC methods.

Consider first the problem of sampling from the posterior distribution given data $x_1, \ldots, x_n$, based on the donor-weighted prior of equation (5), in which the total mass of the imaginary data is adjusted to a fixed value, ω. One possible approach is to explicitly represent the donor model's parameters, ϕ, and the recipient model's parameters, θ, as well as the imaginary data, $y_1, \ldots, y_k$. Letting x denote all the observed data, and y all the imaginary data, the posterior distribution for ϕ, y, and θ is given by the following proportionality:

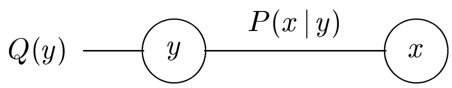
$$\dot{P}_k^\omega(\phi,\, y,\, \theta) \quad \propto \quad Q(\phi) \cdot Q(y \,|\, \phi) \cdot \frac{P(y \,|\, \theta)^{\omega/k} \cdot P(\theta)}{\int P(y \,|\, \theta)^{\omega/k} P(\theta)\, d\theta} \cdot P(x \,|\, \theta) \tag{39}$$

This expression can be diagramed as below:



(40)

Here, each factor is written adjacent to a line that connects to the variables involved in that factor. Markov chain sampling for this distribution can be done by updating each of these variables in turn (apart from x, which is fixed to the observed data), using some method that leaves the distribution invariant, such as Gibbs sampling or a Metropolis-Hastings update. When updating a variable, only the factors associated with the lines connecting to that variable need be considered. Unfortunately, computing $\int P(y \,|\, \theta)^{\omega/k} P(\theta)\, d\theta$ will usually be infeasible when the *pro forma* prior is not conjugate, in which case this Markov chain sampling scheme will be inapplicable. When we do have conjugate priors for the recipient model or for the donor model, it may be desirable to marginalize away the corresponding model parameters. When both model's parameters are eliminated, the Markov chain scheme diagramed below can be used, when $\omega = k$:



(41)

When $\omega \neq k$, the factor of $P(x \,|\, y)$ is replaced by a corresponding factor in which the conditioning on y is adjusted to an equivalent mass of ω.

When the jointly-weighted prior of equation (11) is used, the troublesome integral factor in (39) is eliminated, leaving only factors that are likely to be more easily computable:

$$\ddot{P}_k^\omega(\phi,\, y,\, \theta) \quad \propto \quad Q(\phi) \cdot Q(y \,|\, \phi) \cdot P(y \,|\, \theta)^{\omega/k} \cdot P(\theta) \cdot P(x \,|\, \theta) \tag{42}$$

We can diagram this expression as follows:

$$
\begin{tikzcd}
\phi \arrow[r, dash, "Q(y\,|\,\phi)"] \arrow[d, dash] & y \arrow[r, dash, "P(y\,|\,\theta)^{\omega/k}"] & \theta \arrow[r, dash, "P(x\,|\,\theta)"] \arrow[d, dash] & x \\
Q(\phi) & & P(\theta) &
\end{tikzcd}
\tag{43}
$$

Markov chain sampling will be feasible for a much wider class of models when such a jointly-weighted prior is used, which is one motive for considering these priors.

When the recipient model is in the regular exponential family, these methods can be adapted for use when k is infinite. For such models, the effect of y on the recipient model is determined by its sufficient statistics, which here are assumed to be expressed as sample averages (not totals), so that they stay finite as k goes to infinity. The Law of Large Numbers ensures that when k is infinite, these statistics will equal their expected values, as determined by ϕ. Markov chain updates for θ can therefore be done much as for finite k, using the expected value of the statistics computed from the current ϕ. Markov chain updates for ϕ will have to account for how these expected statistics change when ϕ changes. Looked at another way, one must update ϕ and y simultaneously, with the infinite-dimensional y being represented by its finite-dimensional sufficient statistics, s, which are functions of ϕ. If θ has been marginalized away, as in (41), the state of the Markov chain will consist of ϕ alone, and the distribution to sample from can be diagrammed as below:

$$
\begin{tikzcd}
Q(\phi) \arrow[r, dash] & \phi \arrow[r, dash, "P(x\,|\,\omega s(\phi))"] & x
\end{tikzcd}
\tag{44}
$$

Since the interaction of ϕ and x will generally be complicated, using Gibbs sampling to update ϕ will usually be impossible, but other techniques such as Metropolis-Hastings updates can be used.

The sample of values for θ obtained by Markov chain sampling can be used to estimate posterior expectations, but this sample does not directly provide any feasible method for estimating the marginal likelihood. The marginal likelihood for the recipient model with the donor-weighted prior is the normalizing constant of (39). This can be estimated using methods such as path sampling (Gelman and Meng 1998) or annealed importance sampling (Neal 2001), which require that we embed the distribution of interest in a family of distributions. Such a family can be obtained by replacing the factor $P(x\,|\,\theta)$ with $P(x\,|\,\theta)^{\alpha}$. The normalizing constant when $\alpha = 1$ is the desired marginal likelihood. It can be estimated using samples from a sequence of distributions at varying values of α, starting with $\alpha = 0$, for which the normalizing constant is known to be one, and ending at $\alpha = 1$.

When the jointly-weighted prior is used, the marginal likelihood is the normalizing constant of (42) divided by the normalizing constant of the expression for the prior given by (11). This ratio of normalizing constants can be estimated using a sequence of values for α in the same way as described above for the donor-weighted prior. However, randomly drawing a starting state for the annealing run from the distribution with $\alpha = 0$ and ω set to the desired mass of imaginary data may be difficult. This problem can be handled by by varying the ω parameter as well, from a starting value of zero (where drawing a starting state easy) up to the desired mass, while holding α fixed at zero. The final value for ω is then kept fixed while α is increased to one.

3.3 Letting the mass parameter be variable

As discussed in Section 2.4, we should usually not fix the amount of information transferred from the donor model, but should instead specify a prior distribution for ω that represents our uncertainty regarding the appropriate amount.

If the simple Monte Carlo methods of Section 3.1 are used, the easiest way to implement this is to apply the methods described there for values of ω at equally-spaced quantiles of its prior distribution. The overall marginal likelihood can then be estimated by the average of the marginal likelihoods for these values of ω. Predictions for future observations and posterior expectations for θ can be estimated by weighted averages of estimates obtained at each ω, with the weights being proportional to the marginal likelihood for that ω.

This method could also be used in conjunction with a Markov chain Monte Carlo implementation, but we might instead try to include ω as a hyperparameter that is updated as part of the Markov chain simulation. This should be feasible when a donor-weighted prior is used, since the effects of changing ω are captured by the factors that are already explicitly present in (39). However, updating ω as part of the Markov chain looks difficult when a jointly-weighted prior is used, since there seems to be no easy way to account for the fact that changing ω may change the normalizing constant for (42).

4 Examples of prior information transfer

In this section, I present two examples illustrating the effects of prior information transfer and the computational methods needed to implement it. The first example is fairly simple, with the recipient model being just complex enough that directly formulating an appropriate prior might sometimes be difficult. The second example revisits a more complex modeling problem from the literature, and shows how prior information transfer can avoid problems with prior specifications that might otherwise go unnoticed.

The programs used for these examples were written in R, and are available from my web page. (Note, however, that they are intended only for reproducing the examples in this paper; they are not suitable for general use.) The examples were run on an 866 MHz Pentium III processor. The computation times reported could be reduced by at least a factor of ten by rewriting the programs in a compiled language such as C.

4.1 Example 1: Models for two-way contingency tables

As a first illustration of prior information transfer, I will consider models for two-way contingency tables that summarize data concerning a pair of variables, (X_1, X_2), where X_1 is an integer from 1 to h_1 and X_2 is an integer from 1 to h_2. I will assume that n pairs, $(x_{1,j}, x_{2,j})$, have been observed, and that these were drawn independently from some joint distribution with unknown probabilities $P(X_1 = a, X_2 = b) = \theta_{ab}$.

Often, we may believe that any valid joint probabilities are possible, but we may also believe that X_1 and X_2 are likely to be approximately, though not exactly, independent. We may be able to formulate appropriate priors for the marginal distributions of X_1 and of X_2, but find it difficult to properly formulate a higher-dimensional prior distribution for the

True probabilities of pairs							*Counts of observed pairs*						
.0000	.0000	.0010	.0053	.0174	.0042	.0113	0	0	1	7	22	6	14
.0014	.0159	.0068	.0116	.0121	.0421	.0183	4	14	9	11	14	48	18
.0157	.0184	.0025	.0063	.0089	.0363	.0189	12	18	7	6	7	41	18
.0018	.0065	.0091	.0265	.0064	.0132	.0151	2	8	10	31	6	15	14
.0080	.0027	.0104	.0093	.0123	.0460	.0564	8	4	12	8	9	40	63
.0091	.0118	.0107	.0469	.0388	.1023	.0431	5	16	10	51	35	79	50
.0159	.0133	.0149	.0420	.0259	.0679	.0792	18	10	12	42	23	67	75

Figure 1: The true probabilities for each possible pair, (X_1, X_2), used in the simulation, and the contingency table showing how many times each pair occurred in the 1000 simulated observations. Rows correspond to different values for X_1, columns to different values for X_2.

joint probabilities, θ_{ab}. We would like the prior for these joint probabilities to be compatible with our prior beliefs about the marginal distributions, and for it to also express the idea that the joint distribution is likely, but not certain, to exhibit some degree of approximate independence. This is a situation in which transferring prior information from a simple model that assumes independence may be useful.

I have tested the use of prior information transfer in a simulated context of this sort, in which $h_1 = h_2 = 7$. I used a donor model that assumes independence of X_1 and X_2, and whose parameters are the two vectors of marginal probabilities, ϕ_1 and ϕ_2. The priors for these marginal probability vectors where Dirichlet:

$$\phi_1 \sim \text{Dirichlet}(\alpha_1), \qquad \phi_2 \sim \text{Dirichlet}(\alpha_2)$$

with the parameters of these Dirichlet priors fixed as follows:

$$\alpha_1 = [12\ 18\ 18\ 18\ 36\ 42\ 48], \qquad \alpha_2 = [12\ 12\ 12\ 24\ 24\ 54\ 54]$$

The recipient model allowed for any joint distribution. Its parameters, θ_{ab}, were given a *pro forma* prior that was uniform over the simplex of valid probability distributions, which is equivalent to a Dirichlet distribution with all parameters equal to one.

The actual joint probabilities — ie, the true values for the h_1h_2 parameters θ_{ab} — were simulated from the Dirichlet distribution with parameters $\alpha_{ab} = \alpha_{1,a}\alpha_{2,b}/\sum \alpha_{1,a}$. These true parameters (which were rounded to four decimal places before use) are shown on the left in Figure 1. One thousand observations of pairs were then generated independently with these probabilities. The contingency table of counts from these observations is shown on the right in Figure 1.

The Dirichlet distribution from which the true simulation probabilities were drawn is such that the consequent distributions for the marginal probabilities of X_1 and of X_2 are the Dirichlet distributions that are used as priors in the donor model. This simulates a situation in which we have informative priors for these marginal probabilities, which we have been able to express mathematically. In this simulated situation, the ideal prior for the unrestricted recipient model would of course be the Dirichlet distribution from which the actual probabilities were drawn. In a real situation, however, we might have difficulty in extending our prior for the marginal probabilities to a prior for the whole joint distribution, which involves many more parameters. (There are 48 free parameters in the joint distribution, versus 12 for the two marginal distributions.) Note in this respect that the technique of defining a Dirichlet distribution for the joint probabilities that was used in this simulation — setting $\alpha_{ab} = \alpha_{1,a}\alpha_{2,b}/\sum \alpha_{1,a}$ — produces a prior that is consistent with the priors for the marginal probabilities only when, as here, $\sum \alpha_{1,a} = \sum \alpha_{2,b}$, so that our priors for the two marginal distributions are equally informative. Even when our priors for marginal distributions are Dirichlet, there is no reason in general for them to satisfy this constraint, so this technique is not generally applicable. Of course, a 7-by-7 contingency table is not extraordinarily complex, so with a bit of thought, one can devise various ways of adequately expressing the available prior information. The aim here is to see whether the generally-applicable methods of prior information transfer will also work well.

Prior information transfer was implemented by the simple Monte Carlo methods of Section 3.1. This is possible because the marginal likelihood using a Dirichlet prior can be found analytically. If in m observations, pair (a, b) was observed m_{ab} times, the marginal likelihood for the unrestricted model using a Dirichlet prior with parameters α_{ab} is

$$\frac{\prod\limits_{a,b} \Gamma(m_{ab} + \alpha_{ab}) \;/\; \Gamma(\alpha_{ab})}{\Gamma(m + \alpha) \;/\; \Gamma(\alpha)}$$

where $\alpha = \sum \alpha_{ab}$. By using this formula, one can compute both $P(y_1^{(j)}, \ldots, y_k^{(j)})$ and $P(x_1, \ldots, x_n \,|\, y_1^{(j)}, \ldots, y_k^{(j)})$, as needed for the donor-weighted methods of equations (26) and (30) and the jointly-weighted methods of equations (34) and (38). When computing $P(y_1^{(j)}, \ldots, y_k^{(j)})$, the parameters α_{ab} are those of the *pro forma* prior (all one in the present case), and the counts are for the imaginary data (so $m = k$). When computing $P(x_1, \ldots, x_n \,|\, y_1^{(j)}, \ldots, y_k^{(j)})$, the parameters α_{ab} are the counts for the imaginary data plus the parameters of the *pro forma* prior, and the counts are for the actual data (so $m = n$). The effective mass of the imaginary data can be adjusted to be ω by multiplying the counts for the imaginary data by ω/k. To produce the effect of k being infinite, we generate parameter values ϕ_1 and ϕ_2 from the donor model's prior, but we do not attempt to generate the infinite-size imaginary data vector y. Instead, we proceed as above using $\omega\phi_{1,a}\phi_{2,b}$ in place of the counts for the imaginary data.

The success of the various models can be measured by looking at the marginal likelihood of each for increasing portions of the data, since the marginal likelihood for the first so-many data points measures the model's predictive success on these data points, with respect to the loss given by the sum of minus the log of the predictive probability of each data point in succession up to this point. Such marginal likelihood profiles also show us which model would have been preferred if we had had to choose a model based on each amount of data, on the basis of the Bayes factor.

The plot on the left in Figure 2 shows the log of the marginal likelihood based on increasing portions of the data, for the donor model, the recipient model with its *pro forma* prior, the recipient model with the simulation prior used to choose the actual probabilities, and finally for the model whose parameters are fixed to these true probabilities. Since these are not readily distinguishable on this scale, adjusted marginal likelihood profiles are shown on the

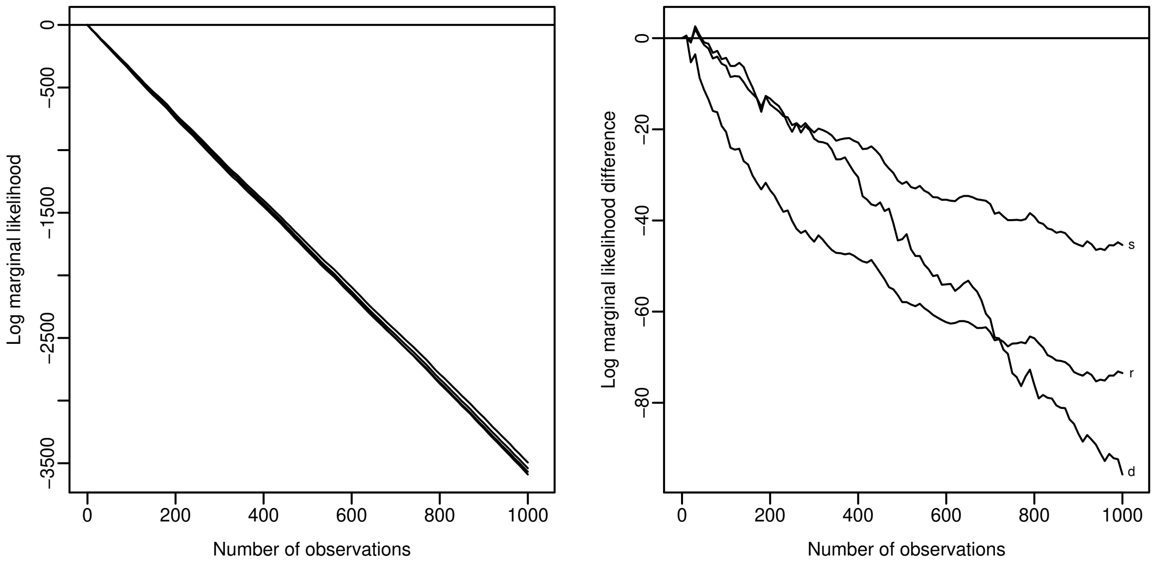


Figure 2: Marginal likelihood profiles for basic models. On the left, the log of the marginal likelihood is shown for increasing amounts of data. On the right, adjusted profiles are shown, in which the log marginal likelihood for the model whose parameters are fixed at the true probabilities has been subtracted. The plots show marginal likelihoods after every ten data points, up to the full one thousand. The models shown (in addition to the model with the true parameters) are the donor model (d), the unrestricted recipient model with its *pro forma* prior (r), and the recipient model with the simulation prior from which the actual probabilities were drawn (s).

right, in which the log marginal likelihood for the model with parameters fixed at their true values has been subtracted, making the differences more easily discernible.

From these plots, we can see that the donor model, which assumes independence and has informative priors, appears better than the unrestricted recipient model until 714 out of the 1000 observations have been seen (ignoring situations after very small numbers of observations, when by chance the recipient model is sometimes favoured). Until this point, the Bayes factor would lead us to chose the model assuming independence, even though the actual probabilities do not satisfy this assumption. Partly, this is due to the donor model's informative priors for the marginal probabilities, which correspond well with the actual marginal probabilities, and partly it is due to the fact that the actual distribution does have some degree of approximate independence. Due to this good prior, for about the first 300 observations, the donor model's marginal likelihood is virtually indistinguishable from that of the model from which the true parameter values were actually drawn, which we could not hope to do better than (except by chance).

Figure 3 shows marginal likelihood profiles for the unrestricted recipient model with various forms of donor-weighted transfer of prior information from the donor model, which assumes independence and has informative priors for marginal probabilities. The plot in the upper left shows the marginal likelihood profiles using the prior of equation (2), with the number of imaginary observations (each observation being a pair in this case) varying from 50 to 800. All these values for k lead to substantially better results than are obtained

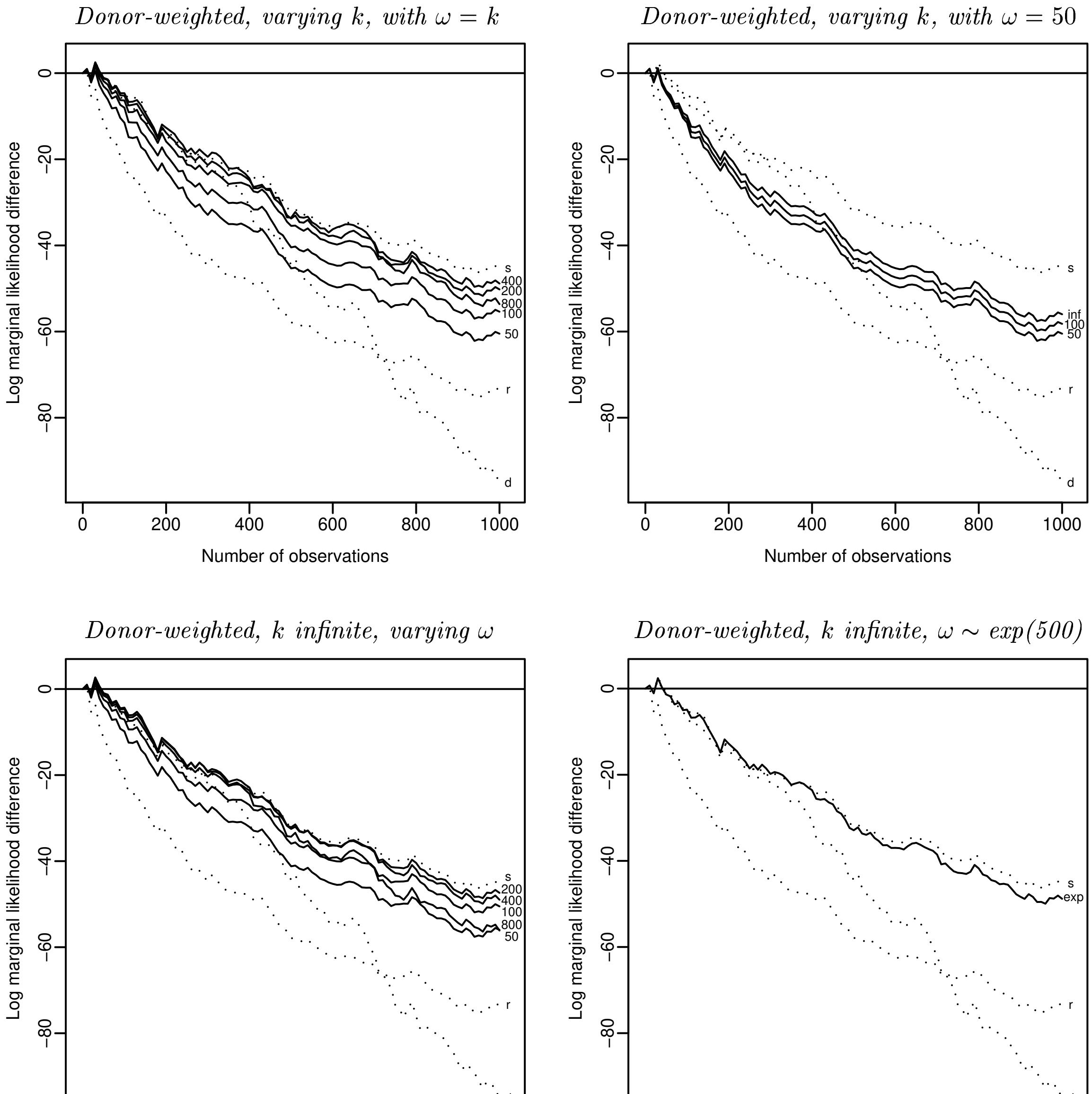


Figure 3: Results using donor-weighted prior information transfer. The plot in the upper left shows marginal likelihood profiles for recipient models receiving prior information from varying numbers of imaginary data points, each with full mass. The plot in the upper right shows the effect of keeping the effective mass constant, while varying the number of data points, up to infinity. The plot in the lower left shows the effect of varying the effective mass of an infinite amount of imaginary data. Finally, the plot in the lower right shows the marginal likelihood profile for a model in which the mass of imaginary data is given an exponential prior with mean 500. For comparison, dotted lines show the marginal likelihood profiles for the basic models of Figure 2.

using the recipient model's *pro forma* uniform prior. The best results are obtained when $k=400$.

The upper-right plot in Figure 3 shows that, for this problem at least, a given mass of imaginary data is best obtained by using a larger number of imaginary data points, whose mass is then adjusted, as in equation (5). For a mass of $\omega=50$, we see an improvement in the plot as we go from obtaining this mass using 50 full imaginary data points, to obtaining it using 100 data points whose mass has been reduced by one half, and finally to obtaining a mass of 50 using an infinite number of imaginary data points.

It therefore makes sense to fix k to be infinite, and investigate which value for ω is best. The plot in the lower right of Figure 3 shows marginal likelihood profiles for values of ω from 50 to 800. We see that when k is infinite, a mass of 200 is slightly better than 400, the reverse of what was seen in the upper left, where $k = \omega$.

In an actual application, we would seldom know *a priori* what value for ω was best. The plot in the lower right shows the marginal likelihood profile when ω is given a prior distribution that is exponential with mean 500. This marginal likelihood profile is almost as good as that for the model whose prior is the one from which the true probabilities were actually simulated. For up to about 250 observations, the profiles for these two models are close to the marginal likelihood profile for the donor model, indicating that this amount of data is needed in order for the lack of complete independence to become apparent.

The results in Figure 3 were obtained using the simple Monte Carlo estimator of equation (26). The Monte Carlo sample size was $N=2000$ for the models with infinite k, which required about six minutes of computation time. The standard error for the estimated log marginal likelihood for all 1000 observations increased roughly linearly with the mass of imaginary data, from 0.03 for $\omega=50$ to 0.29 for $\omega=800$. When k was equal to ω, obtaining precise estimates required a larger Monte Carlo sample size, which is not surprising, since letting k be infinite eliminates one source of variability in $P(x_1,\ldots,x_n \,|\, y_1^{(j)},\ldots,y_n^{(j)})$. Accordingly, the Monte Carlo sample size was increased to $N = 10000$ for the estimates with finite k, which required about thirty minutes of computation time. The resulting standard errors for the log marginal likelihood increased with k, from 0.13 when $k=\omega=50$ to 0.50 when $k = \omega = 800$. The marginal likelihood with $\omega \sim \exp(500)$ was estimated by averaging simple Monte Carlo estimates for the marginal likelihood (*not* its log) with ω set to quantiles 0.00, 0.02, …, 0.98 of the exp(500) distribution, with the sample size for each of these computations being $N=200$. The total computation time for this was about thirty minutes. The standard error for the resulting estimate of the log marginal likelihood was 0.14.

Marginal likelihoods using jointly-weighted prior information transfer were also estimated by simple Monte Carlo, using equation (34). Because the effective sample size in (34) is reduced in comparison to that in (26) due to the variability in $P(y_1^{(j)},\ldots,y_n^{(j)})$, a much larger sample size of $N = 50000$ was used for these estimates, taking about three hours of computation time. As was the case for donor-weighted transfer, accuracy was lower for $k = \omega$ than for infinite k, and it was lower in both cases when ω was larger. For given values of k and ω, the marginal likelihood for jointly-weighted transfer was slightly higher than for donor-weighted transfer, at least for small values of ω, which were the only ones for which results accurate enough for a clear comparison were obtained. For infinite k, the log marginal likelihood with $\omega = 10$ was -3562.035 ± 0.010 for donor-weighted transfer, versus -3561.973 ± 0.002 for jointly-weighted transfer; with $\omega = 50$, the corresponding values were -3550.50 ± 0.03 for donor-weighted transfer and -3549.01 ± 0.05 for jointly-weighted transfer. With $k = \omega = 10$, the log marginal likelihood was -3563.51 ± 0.04 for donor-weighted transfer and -3563.30 ± 0.04 for jointly-weighted transfer. (All estimates are given with $\pm$ the estimated standard error.)

Marginal likelihoods were also estimated using annealed importance sampling, using the scheme diagramed in (40) for donor-weighted transfer and that of (43) for jointly-weighted transfer. Gibbs sampling was used to update the parameters and imaginary data, which was possible for donor-weighted transfer because the marginal likelihood using the *pro forma* prior can be computed analytically. For a donor-weighted prior with $k=\omega=50$, a standard error of 0.08 for the log marginal likelihood was achieved using 50 annealing runs, in each of which α was varied from 0 to 1 in steps of 0.0005, with a single Gibbs sampling scan updating θ, y, and ϕ at each value of α. This required 72 minutes of computation time in total, as compared with 32 minutes for the simple Monte Carlo estimate, which had a somewhat larger standard error of 0.13. For this problem, simple Monte Carlo and annealed importance sampling were about equally efficient.

When using a jointly-weighted prior with $k=\omega=10$, a standard error of 0.13 was achieved using 50 annealing runs in which ω was first varied from 0 to 10 in steps of 0.005 and α was then varied from 0 to 1 in steps of 0.0005, with one Gibbs sampling update at each stage. This required 40 minutes of computation (less per stage than for donor-weighted transfer), compared to 162 minutes for finding a simple Monte Carlo estimate whose standard error was 0.04. Simple Monte Carlo is a bit more efficient in this case, but when $k = \omega = 50$, no useful estimate was obtained using simple Monte Carlo with $N = 50000$ (taking 178 minutes), whereas annealed importance sampling produced an estimate with a standard error of 0.11 using 215 minutes of computation time. This estimate was obtained using 50 annealing runs in which ω was varied from 0 to 50 in steps of 0.005 and α was then varied from 0 to 1 in steps on 0.0001. The estimate found for the log marginal likelihood of -3567.41 ± 0.11 was surprisingly low, in comparison with -3554.80 ± 0.08 for donor-weighted transfer with $k = \omega = 50$, and -3563.30 ± 0.04 for jointly-weighted transfer with $k = \omega = 10$. It is not much different from the log marginal likelihood of -3568.00 for the recipient model with its *pro forma* prior. Jointly-weighted transfer with $k=\omega=50$ seems to modify the prior in an unexpected, and in this case, undesirable, way.

Annealed importance sampling was also tried for jointly-weighted priors with k infinite. For these runs, θ was updated by Gibbs sampling, as for finite k, but a Metropolis-Hastings update was done for ϕ, using $Q(\phi)$ as the proposal distribution (the infinite-sized y was in effect updated simultaneously with ϕ). With $\omega = 50$, the log marginal likelihood was estimated with a standard error of 0.10 using 50 annealing runs in which ω was varied from 0 to 50 in steps of 0.05, and then α was varied from 0 to 1 in steps of 0.0001. This took 16 minutes, compared with 158 minutes for a simple Monte Carlo estimate whose standard error was 0.05. With $\omega=100$, annealed importance sampling took 140 minutes to obtain an estimate with standard error of 0.18, using 50 annealing runs with ω varied in steps of 0.0025 and α in steps of 0.000025. This is better than simple Monte Carlo, which required

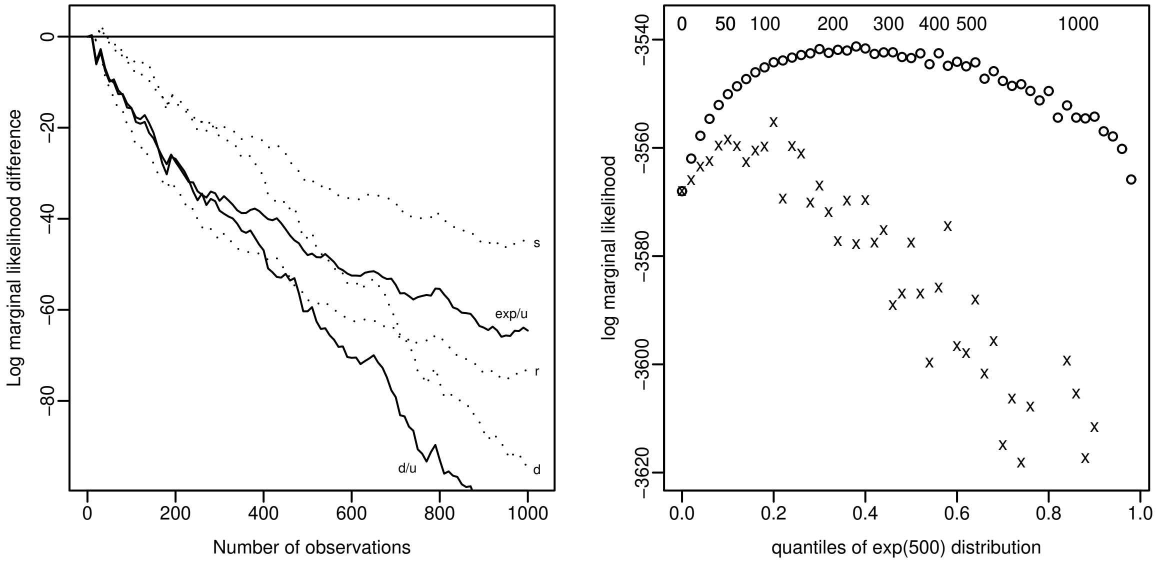


Figure 4: Comparison of results using donor models with informative and with uniform priors. The left plot shows marginal likelihood profiles for a model that assumes independence, with uniform priors for the marginal probablities (lower solid line), and for an unrestricted model with donor-weighted prior information transfer from this model (upper solid line). The transfer used infinite k, and $\omega \sim \exp(500)$. For comparison, the dotted lines show the marginal likelihood profiles from Figure 2. The plot on the right shows how the log marginal likelihood varies with ω when using the donor model with uniform prior (x, some low points omitted) and when using the donor model with the informative prior (o). The estimates shown are those used to find the estimate with $\omega \sim \exp(500)$, with ω set to quantiles of this distribution. They are based on samples sizes of $N = 200$ for the informative prior and $N = 2000$ for the uniform prior.

172 minutes to obtain an estimate with a standard error of 0.33. The estimated log marginal likelihood using jointly-weighted transfer with k infinite and $\omega = 100$ was -3543.82 ± 0.18, which is slightly better than the value of -3545.04 ± 0.06 for donor-weighted transfer. It appears that the peculiarly bad behaviour seen above for jointly-weighted transfer with $\omega = k = 50$ does not occur when k is infinite.

Figure 4 shows that prior information transfer is still beneficial when the donor model uses a uniform prior for marginal probabilities, rather than the informative prior used for the previous tests. The benefit in this case comes from transferring to the recipient model the knowledge that the distribution is likely to show to some degree the independence assumed by the donor model. The benefit of information transfer from this donor model (measured in terms of log marginal likelihood) is about one-third of that from using the donor model with the informative prior. As seen in the right of Figure 4, the optimal value of ω for this donor model (about 100) is also less than the optimal ω for the donor model with the informative prior (about 250). The sample size needed to get good estimates using simple Monte Carlo is much larger when the donor model's prior is not informative — even though the sample used was ten times larger, the marginal likelihood estimates using the donor model with uniform prior on the right in the figure are substantially noisier than those using the donor model with the informative prior.

In this example, when the donor model with the informative prior was used, donor-weighted prior information transfer with k infinite and with ω given a fairly vague prior achieved nearly as good a result as could possibly be expected, even if the actual mechanism used to choose the true parameters were assumed known. Jointly-weighted transfer with k infinite also worked well, but when k was finite, the jointly-weighted prior behaved strangely. Simple Monte Carlo computation worked adequately well for donor-weighted transfer with this model, but was not so successful for models using jointly-weighted transfer, or when the prior was not informative. With either method of prior information transfer, it is likely that larger-scale problems will required use of Markov chain methods, in conjunction with a method such as annealed importance sampling if marginal likelihoods are needed.

4.2 Example 2: Comparing graphical models for categorical data

As a second example, I look at the comparison of two directed graphical models for the data on factors influencing college plans that was examined by Sewell and Shah (1968), and used as an example by Heckerman, Meek, and Cooper (1999).

The data concerns 10,318 high school seniors in Wisconsin, randomly sampled from all such students in 1957. For each student, the following variables were recorded:

SEX	male or female
SES	socioeconomic status: low, lower middle, upper middle, or high
IQ	intelligence quotient: low, lower middle, upper middle, or high
PE	parental encouragement to attend college: low or high
CP	whether student plans to attend college: yes or no

The objective was to investigate which factors influenced whether the student planned to attend college, and if so how. For example, it was of interest to know whether or not SES influenced CP only through its effect on PE. Alternatives of this sort can be expressed using directed graphical models (see, for example, Cowell, *et al* 1999). These have a probabilistic interpretation as a description of the joint distribution for all variables in terms of the conditional distributions for each variable given values for its "parent" variables, from which arrows to it are drawn. One may also wish to give a causal interpretation to the arrows in such models, though whether this is justified will depend on subtleties that I will not discuss here.

For this example, I will consider only two graphical models for this data, which were the two most probable models found by Heckerman, *et al* in an exhaustive search of all models without latent variables. These models are shown in Figure 5. Model A is similar to one used by Sewell and Shah, except that they did not specify a direction for the link between SES and IQ, and they analysed data for males and females separately, which is equivalent to making SEX a parent of all the other variables. This model seems plausible on the basis of common-sense prior knowledge. Model B is the same as Model A except that the arrow from IQ to PE has been reversed. If the model is interpreted causally, this is a change from PE being determined by IQ, SEX, and SES, to it being determined only by SEX and SES, while IQ is now influenced by PE as well as SES.

From the point of view of a causal interpretation, Model B is much less plausible than Model A, based on common-sense prior knowledge. By the time parents are encouraging or

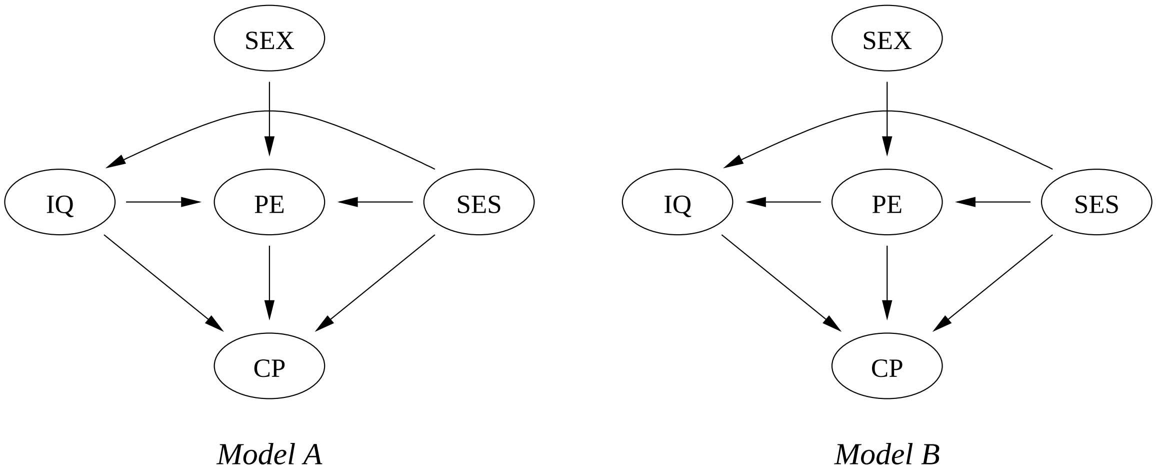


Figure 5: Two graphical models for the data on college plans.

not encouraging their children to go to college, the child's IQ must surely have stabilized, and be uninfluenced by whether such encouragement is received. It is somewhat more plausible that PE could be a surrogate for general parental encouragement of academic pursuits, including encouragement at an age early enough to influence IQ, but this is still not as plausible as the reverse causation of Model A, in which the child's IQ influences whether their parents encourage them to go to college.

From a probabilistic viewpoint, Model A is also more plausible. It implies that SEX and IQ are marginally independent. This is quite likely, since the average IQ for males and females is generally thought to be the same, and any small sex difference in the variance of IQ is likely to be concealed when its value is reduced to the four categories above. Model B lacks this marginal independence, but implies that SEX and IQ are conditionally independent given PE and SES. This would exclude, for example, situations in which parents of a certain socioeconomic status encourage all male children to go to college, but only encourage female children if they have a high IQ. Common sense prior knowledge says that such scenarios are quite possible, and indeed must certainly be present to at least a small degree.

It is therefore rather surprising that Heckerman, *et al* find that the Bayes factor favours Model B over Model A by a factor of 8.4×10^{19}. This result is of course influenced by the prior distributions for the parameters of the two models, which are the conditional probabilities for the various possible values of each variable, given each possible combination of values for its parent variables. Heckerman, *et al* use independent Dirichlet priors for these conditional distributions, with Dirichlet parameters based on the assumption that some number, η, of hypothetical previous observations occurred, spread equally over the $2 \times 4 \times 4 \times 2 \times 2 = 128$ possible combinations of values for the variables. For example, in Model B, the prior for the probabilities of the four possible values of IQ given that PE=low and SES=high is Dirichlet$(\eta/32, \eta/32, \eta/32, \eta/32)$, since $\eta/8$ of the hypothetical observations will have PE=low and SES=high, and each value of IQ will occur in a quarter of these. (Note that these counts of hypothetical observations may well be fractional.)

Heckerman, *et al* use this prior with $\eta = 5$, but they report that Model B is still favoured for any value of η from 3 to 40. Although the preference for Model B is robust with respect to choice of η, we may wonder whether use of other prior distributions might lead to a different conclusion. For this problem, we have considerable prior knowledge that is not captured by the symmetric Dirichlet priors used by Heckerman, *et al.* For instance, we know that the relationships of PE to SES and of PE to IQ are likely to be monotonic with respect to the four levels of SES and IQ, and that higher IQ and higher SES are both likely to be associated with a greater likelihood of parental encouragement to attend college. Ideally, we would specify prior distributions for each model that embody such prior knowledge, since it is the Bayes factor found when using these priors that we would wish to use when deciding which model is a better description of reality.

Even for these fairly simple models, specification of such a good prior would not be easy, however. As a substitute, we can try using independent Dirichlet distributions as described above as *pro forma* priors, and then transfer prior information from each model to the other in an attempt to even out the bad effects of these unrealistic priors. The hope is that the resulting Bayes factor will be a better approximation to the one we would have obtained if we had made the effort to properly specify the prior distribution for each model.

For this test, I used donor-weighted transfer with k infinite, as in equation (6), with various masses of imaginary data, ω. The CP variable was omitted, since its treatment is the same in Models A and B. As discussed in Section 2.3, this omission would have no effect if donor-weighted transfer with $\omega = k$ were used. In these tests, $\omega \neq k$, so this omission could have an effect, but it seems best in any case to ignore this irrelevant aspect of the models, and doing so also saves computer time. I used *pro forma* Dirichlet priors defined using the same scheme as Heckerman, *et al*, but with $\eta = 120$. This value of η gives larger marginal likelihoods for both models (with CP omitted) than the value of $\eta = 5$ used by Heckerman, *et al.* With $\eta = 5$, the log marginal likelihood is -41257.1 for Model A and -41211.2 for Model B; with $\eta = 120$, the values are -41176.0 for Model A and -41162.2 for Model B, which are each close to the maximum for any η. With this better prior, the difference in log marginal likelihoods is reduced from 45.9 to 13.8, corresponding in a reduction in the Bayes factor in favour of Model B from 8.4×10^{19} to 9.4×10^{5}. This illustrates how sensitive the Bayes factor is to the prior, but even the smaller Bayes factor still strongly favours what seems like the less plausible model.

Prior information transfer was implemented using annealed importance sampling, with Markov chain sampling based on the scheme of diagram (44). Metropolis-Hastings updates were used, in which proposals were made to change a part of ϕ corresponding to a single conditional distribution (each proposal therefore changed at most three of the parameters). The new probabilities for this conditional distribution were drawn from the prior, $Q(\phi)$, and were then accepted or rejected based on the resulting change in the probability of the observed data, x. The annealing schedules used varied α from $\exp(-8)$ to 1, in approximately geometric steps. The number of levels for α varied from 4000 for $\omega = 50$ to 56000 for $\omega = 6400$. Ten annealing runs were done to obtain each estimate (except that only five were done at $\omega = 6400$). The computation time required for a given ω, and a given direction of transfer, varied from about three hours for $\omega = 50$ to about thirty hours for $\omega = 3200$. Even with this amount of computation, the accuracy of the estimates obtained is only barely adequate. Due to the small number of annealing runs, the estimated standard errors should not be trusted too much.

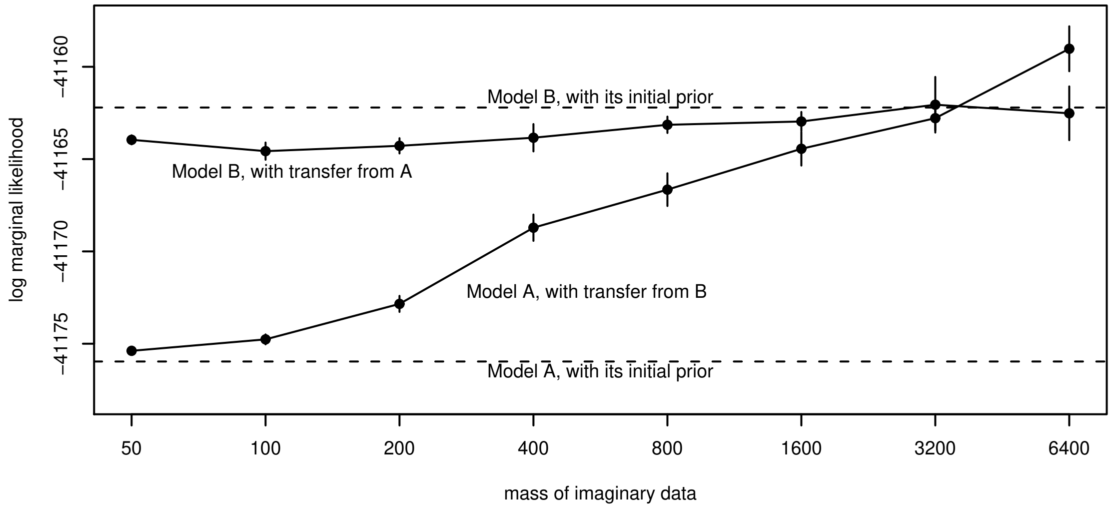


Figure 6: Marginal likelihoods for Models A and B with and without donor-weighted prior information transfer from the other model. The logarithmic horizontal axis gives the value of ω (with k infinite); the vertical axis gives the log marginal likelihood. The vertical lines through each dot extend to plus and minus twice the standard error of the Monte Carlo estimate of the log marginal likelihood. The dotted lines show the log marginal likelihoods of the initial models, with no prior information transfer. The models do not include the CP variable, since its treatment does not differ between models.

Figure 6 shows the results. The marginal likelihood of Model A is greatly improved by prior information transfer from Model B, with the improvement increasing with the mass of imaginary data, up to the largest mass tested of $\omega = 6400$, for which the estimated marginal likelihood is greater than that of Model B with its initial prior. In contrast, the marginal likelihood for Model B is somewhat lower when information is transferred from Model A.

It is perhaps surprising that such a large mass of imaginary data is beneficial for transferring information from Model B to Model A. One might wonder whether an infinite mass of imaginary data would be even better. To approximate this, I tried using a mass of 10^6, and obtained an estimate for the log marginal likelihood of -41163, but the variability of the annealing runs was so large that this estimate cannot be trusted (nor can its standard error be reliably estimated). It does seem, however, that an infinite mass of imaginary data would give sensible results, if enough time were available to do the computations.

What can we now conclude regarding Models A and B? We can consider at least three priors for each model — the initial prior, the prior with transfer from the other model using the best value of ω (including the initial prior, corresponding to $\omega = 0$), and the prior obtained by giving ω a prior distribution that is uniform over the values tried (including $\omega = 0$). Here are the estimated log marginal likelihoods with these three priors:

	Initial prior	With best ω	With ω unknown
Model A	-41176.0	-41159.0	-41161.2
Model B	-41162.2	-41162.1	-41162.9

The estimates for the prior with ω unknown were obtained by simply averaging the estimated marginal likelihoods (not their logs) for each value of ω. The prior with ω regarded as unknown might seem to be the "fairest", but it might also seem "unfair" to evaluate Model B in terms of its marginal likelihood using transfer of prior information from Model A with any value of ω, since Model B does not benefit from such transfer. It may be wise to compute two Bayes factors — one in which the marginal likelihood used for Model A is the largest of the three above, while that used for Model B is for ω unknown, the other in which the reverse is done. If these two Bayes factors lead to substantially different conclusions, one might regard the results as inconclusive, or one might decide to go to the effort of formulating better priors, in order to resolve the issue.

For the present example, the first of these Bayes factors (giving Model A the benefit of the doubt) is $\exp(41162.9 - 41159.0) = 49$, in favour of Model A. The second Bayes factor (biased toward Model B) is $\exp(41162.1 - 41161.2) = 2.5$, again in Model A's favour. This corresponds to something between mild and quite strong evidence in favour of Model A. From the trend in Figure 6, one might well guess that transfer using a mass of imaginary data greater than 6400 would reveal yet stronger evidence in favour of Model A.

This is a dramatic reversal from the Bayes factor of 9.4×10^5 in favour of Model B that the initial priors produced. One indication of the cause of this is that in Model A, the conditional distributions for PE involve 32 parameters, and those for IQ involve 12 parameters, for a total of 44, whereas in Model B, the conditional distributions for PE involve 8 parameters, and those for IQ involve 24 parameters, for a total of 32. The larger number of parameters in Model A would not be a problem if the prior for these parameters were well specified, but bad priors might be more harmful to a model with many parameters than to a model with fewer. Even so, however, many bad aspects of the priors used — such as the lack of attention to the known ordering of values for IQ and SES — are just as bad in Model B as in Model A, so one might wonder how transfer from B to A can help. The benefit may come from suppression of interactions that are ruled out by the structure of Model B, but not by that of Model A, such as the possibility that parents might encourage high IQ males to attend college, but discourage high IQ females from attending. Of course, we actually believe that at least the directionality of the IQ effect on PE is likely to be the same for both sexes.

5 Discussion

As seen in the examples of Section 4, prior information transfer can sometimes correct poorly-specified priors, thereby improving the predictive performance of a model, and allowing for more meaningful comparison of alternative models. This may allow us to reduce the effort we expend in specifying a prior distribution, which for a complex model might otherwise be quite daunting. Unfortunately, we must instead spend time writing a Monte Carlo program to implement prior information transfer for the model, and this program may take a considerable amount of computer time to run. Eventually, however, prior information transfer could be included in software packages that automatically perform Bayesian computations. With further research, and further progress in computer technology, we may hope that these computations will be fast enough for routine use.

The examples in this paper were mostly done using donor-weighted transfer, which worked well for these examples, and which appears to be easier to understand than jointly-weighted transfer, which sometimes behaves strangely. However, donor-weighted transfer will be infeasible (at least using current methods) when the recipient model and its *pro forma* prior are not conjugate. Since jointly-weighted transfer will still be feasible for many such models, a better understanding of the priors that result from using jointly-weighted transfer would be of interest.

In this paper, I have considered only comparison of two models, each of which receives prior information transferred from the other. Often, we would have more than two models to compare. Prior information transfer could be extended in various ways to allow information to be transferred from more than one model, which can be seen as different ways of cascading the method — for instance, one could transfer information from Model A to Model B, and then transfer information from this modified Model B to Model C, or alternatively, one could transfer information from Model A to the result of transferring information from Model B to Model C. The best way may depend on whether these models form a natural hierarchy, or are better seen as being all on one level.

When comparing nested models, prior information transfer from the simpler to the more complex model may make it difficult for the evidence to decisively favour the simpler model, since the more complex model behaves much like the simpler model when its prior incorporates a large mass of imaginary data generated from the simpler model's prior. However, in such situations, we often do not believe that the simpler model can really be exactly true anyway, even if we think it may be approximately true. We may, for instance, think that *exact* independence could not possibly hold for some two-way contingency table, even though approximate independence is plausible. In these situations, it makes more sense to regard the simpler model merely as a device for specifying the prior for the complex model, as in the example of Section 4.1. The posterior distribution for the mass of imaginary data (assuming this is not fixed) may then be of interest as a measure of how close the simpler model is to reality.

One should keep in mind that prior information transfer is merely a way of improving prior specifications, and that it will therefore not fix problems that are inherent in the whole idea of comparing models using Bayes factors. In particular, use of Bayes factors is justified on the basis that we believe that one of the models being considered is a true description of reality — or at least close enough that any flaws are not visible given the precision of the available data. Bernardo and Smith (1994) call this the "$\mathcal{M}$-closed" scenario. When we do not believe that any of the models is more than a rough approximation to reality — the "$\mathcal{M}$-open" scenario — selecting a model based on Bayes factors cannot be justified. Indeed, which is the best of these incorrect models will depend on our purpose, so any well-justified model selection procedure must take this purpose into account. (In contrast, the true model is best for all purposes, so we needn't consider this in an $\mathcal{M}$-closed scenario.)

Imaginary data has been used before as a technique for specifying prior distributions. Conjugate priors are often viewed as expressing prior information that is equivalent to the observation of certain imaginary data points. Often, the conjugate priors used in practice are based on minimal amounts of imaginary data, chosen to be minimally informative. However, Madigan, Gavrin, and Raftery (1994) describe how they carefully elicited a substantial amount of imaginary data from an expert, and report that conditioning on this imaginary data improved predictions. The prior information transfer methods discussed here can be seen as generalizing this technique, by not fixing the imaginary data, but instead giving it a prior distribution derived from the donor model, whose prior incorporates the expert's knowledge. This generalization avoids the problem that the expert may consider two imaginary cases to each be plausible, but think that they are unlikely to both occur — ie, they are representative of mutually exclusive possibilities. This knowledge can't be expressed by a single imaginary data set, but can be expressed in a donor model's prior. On the other hand, eliciting a single imaginary data set may sometimes be more feasible than eliciting the structure and prior for a donor model.

Spiegelhalter and Smith (1982) use imaginary data in another way, to resolve the indeterminacy that results when models compared using Bayes factors are given improper priors. They advocate fixing the arbitrary ratio of constants in the Bayes factor at a value that makes the Bayes factor be one for an imaginary data set of the smallest size that permits model comparison, and which is such as to give maximum support to the simpler model. This procedure appears rather arbitrary, however.

Several procedures based not on imaginary data, but on portions of the real data, have also been proposed for resolving this "problem" that Bayes factors cannot be used to compare models with improper priors. The various "intrinsic Bayes factors" of Berger and Pericchi (1996) use priors that have been made proper by conditioning on a portion of the data, with the Bayes factor then found using the remaining portion. Since the result will vary depending on which portion of the data is selected to condition on, they explore various ways of averaging the results obtained from different selections. The "fractional Bayes factors" of O'Hagan (1995) are similar, but use a portion of the whole data set (found by raising the likelihood to a fractional power), rather than a selection of particular data points. Both schemes suffer from a dependence on arbitrary choices, such as the amount of data to condition on.

Even if a scheme of this sort could be found that had no arbitrary aspects, however, it would still lack any clear justification. Ideally, we should compare models using priors that capture our well-considered prior beliefs. If we cannot do this in practice, we should at least try to approximate this ideal — otherwise, why would we place any credence in the resulting answer? From this perspective, any sensible model comparison procedure *must* somehow incorporate real prior information, not try to avoid doing so. The challenge is to do this for high-dimensional models where direct specification is too difficult. Prior information transfer from a simpler model, or from the model being compared against, is a general technique that can help in some such situations.

Acknowledgements

I thank David Heckerman for introducing me to the Sewell and Shah data. This research was supported by the Natural Sciences and Engineering Research Council of Canada and by the Institute for Robotics and Intelligent Systems.

References

Berger, J. O. and Pericchi, L. R. (1996) "The intrinsic Bayes factor for model selection and prediction", *Journal of the American Statistical Association*, vol. 91, pp. 109-122.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*, Chichester: John Wiley.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*, New York: Springer-Verlag.

Gelman, A. and Meng, X.-L. (1998) "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling", *Statistical Science*, vol. 13, pp. 163-185.

Geweke, J. (1989) "Bayesian inference in econometric models using Monte Carlo integration", *Econometrica*, vol. 57, pp. 1317-1339.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (editors) (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Heckerman, D., Meek, C., and Cooper, G. F. (1999) "A Bayesian approach to causal discovery", in C. Glymour and G. F. Cooper (editors), *Computation, Causation, and Discovery*, AAAI Press.

Kass, R. E. and Raftery, A. E. (1995) "Bayes factors", *Journal of the American Statistical Association*, vol. 90, pp. 773-795.

Madigan, D., Gavrin, J., and Raftery, A. E. (1994) "Eliciting prior information to enhance the predictive performance of Bayesian graphical models", Technical Report 270, University of Washington, Dept. of Statistics.

Neal, R. M. (2001) "Annealed importance sampling", *Statistics and Computing*, vol. 11, pp. 125-139.

O'Hagan, A. (1995) "Fractional Bayes factors for model comparison" (with discussion), *Journal of the Royal Statistical Society B*, vol. 57, pp. 99-138.

Ripley, B. D. (1987) *Stochastic Simulation*, New York: John Wiley.

Sewell, W. H. and Shah, V. P. (1968) "Social class, parental encouragement, and educational aspirations", *American Journal of Sociology*, vol. 73, pp. 559-572.

Spiegelhalter, D. J. and Smith, A. F. M. (1982) "Bayes factors for linear and log-linear models with vague prior information", *Journal of the Royal Statistical Society B*, vol. 44, pp. 377-387.