

Hierarchical Models for Permutations: Analysis of Auto Racing Results

Todd Graves, C. Shane Reese & Mark Fitzgerald*

ABSTRACT

The popularity of the sport of auto racing is increasing rapidly, but its fans remain less interested in statistics than the fans of other sports. In this paper, we propose a new class of models for permutations which closely resembles the behavior of auto racing results. We pose the model in a Bayesian hierarchical framework. This framework permits hierarchical specification and fully hierarchical estimation of interaction terms. The methodology is demonstrated using several rich datasets which consist of repeated rankings for a collection of drivers. Our models can potentially identify individuals who are racing in “minor league” divisions who have higher potential for competitive performance at higher levels. We also present evidence that one of the sport’s more controversial figures, Jeff Gordon, is a statistically dominant figure.

KEY WORDS: Bayesian Hierarchical Models; Bradley-Terry Model; Markov Chain Monte Carlo.

*Todd Graves is in the Statistical Sciences Group at Los Alamos National Laboratory. Shane Reese is in the Department of Statistics at Brigham Young University. Mark Fitzgerald is in the Mathematics Department at the University of Colorado at Denver. The authors are grateful for the comments of the editor, the associate editor, and three referees which greatly improved the manuscript.

1. INTRODUCTION

Stock car racing is the fastest growing sport in the United States with a 91% growth in attendance in the 1990's (Martin 2000). Furthermore, the National Association for Stock Car Auto Racing (NASCAR) is second only to the National Football League in television viewership with over 180 million viewers in 1998. Despite the popularity of the sport, NASCAR has received little attention as a subject of study by statisticians. As a classroom example, Tenebein (1971) proposed a simple probability model for the number of cars on a track at any given time. Sullivan (2000) used ordered probit modeling (under which all drivers' finishing positions in the same race are treated as independent) to estimate driver abilities on each of four types of tracks, overall, in qualifying, and in avoiding accidents. Mockus, Hibino, and Graves (2000) used NASCAR data to demonstrate visualization tools.

NASCAR is big business. Racing teams are composed of pit crews, mechanics, drivers, and owners. Owners must find drivers, pit crews, and mechanics who they think will be most successful racing their cars, and successful teams generate sponsorships that equate to money. The three NASCAR racing series are the Winston Cup (WC), the Busch Grand National (BGN), and the Craftsman Truck (CT) series. The WC is considered the major league series, and BGN and CT are analogous to the minor leagues in baseball. New drivers for the WC series are generally selected from the BGN and CT series. Therefore, significant money is at stake when deciding which BGN or CT driver will be promoted to the WC series. Reliable prediction of how performances in the lower levels translate to higher levels (CT→BGN, CT→WC, and BGN→WC) of competition are of great interest.

In this paper we construct a probability model for finishing positions in a NASCAR race. A feature of this modeling is that we can address such questions as: "How likely would it be for Jason Keller (a BGN driver) to win a title if he were to race in the Winston Cup?" or "what is the probability that Dale Earnhardt would have won the championship in 2001?" Earnhardt, arguably the greatest driver in NASCAR history, was killed on the last lap of the Daytona 500, the first race of the 2001 season. He shares with Richard Petty the record of seven NASCAR championships, and his attempt to win his eighth after a second place finish at age 49 in 2000 would have been watched carefully in the 2001 season.

These types of questions have led us to develop a rich class of models for permutation data. Racing data

is well-suited to explore such models because we have repeated rankings of subsets of individuals, where the ranking mechanisms may differ in interesting ways (e.g. different tracks or different competition) and are subject to substantial noise.

While repeated rankings allow us to estimate driving abilities, there are a number of complicating factors. Exploring rates of improvement of young drivers is of great interest, so we explore abilities that change over time. Most sports have home field advantages as complicating factors, but auto racing has potentially as many home field effects as there are track–driver combinations. One measurable characteristic of a track is its predictability, that is, to what extent good drivers are more likely than average drivers to perform well there. We will also explore the driver–track interaction to measure the extent to which drivers specialize on certain (types of) tracks. This is done via a hierarchical specification of interactions. Our modeling approach is somewhat exploratory in nature: rather than striving for a single model explaining all the data, we list several factors that we expect to be important, and for each, use an appropriate subset of our data to explore that factor. This approach keeps each individual analysis small enough to be computationally feasible, and makes concessions to the fact that the data set is too small to consider all hypotheses simultaneously.

Inherent in our modeling approach is the parameter estimation associated with driving team ability. Throughout this paper we refer to driver abilities; however, the modeling employed in this paper is only able to rate all members of the racing teams together, including owners, crew chiefs, pit crews, and mechanics. Our data are insufficient to assess driver ability independent of team ability. Another concern is that drivers who drive partial seasons can choose races on tracks for which their skills are well-suited. For most comparisons between drivers, this is not an issue, in part because most of the above-average drivers compete in every race, missing races only at random due to injury. However, race selection bias can potentially be a problem when estimating the ability of road course specialists who drive only in the one or two races per year that feature right turns.

This paper is organized as follows. In Section 1.1 we review related work in both sports and non-sports applications. The data on which our analysis is based is discussed in Section 2. Our fundamental model is presented in Section 3. Section 4 details the strategy for estimation of the model we pose. Results from and extensions to that model are contained in Section 5. In Section 6 we conclude with some discussion and

proposals for future work.

1.1 Related Work

While the statistical literature contains little evidence of work on auto racing, there is a substantial literature on permutation models. Marden (1995) provides a nice review of models for permutations.

Bradley and Terry (1952) present the seminal work on modeling head-to-head competition, in which two opponents have ability parameters λ_1 and λ_2 , and the probability of a win by the first opponent is $\lambda_1/(\lambda_1 + \lambda_2)$. Luce (1959) presents a natural extension of the Bradley-Terry model by allowing more general comparisons than just paired comparisons. Plackett (1975) presents a saturated logistic model for probabilities of winning and illustrates his methodology with voting results.

Yu and Lam (1996) provide a nice analysis of duplicate bridge tournament data, using the Bradley-Terry model. Ties are quite common in bridge, which severely restricts modeling choices. Johnson, Deaner, and van Schaik (2000) propose a latent variable model for primate intelligence. The responses are multiple rankings of the skills of several genera of primates at performing different tasks, and thus this is a competing model for that proposed here. The approach is couched in the Bayesian hierarchical modeling framework. A strength of such an approach is that the Gaussian error structure makes it possible to study correlated rankings. We studied a similar approach for racing data, but Gaussian errors fit the data poorly.

Stern (1990) presents a general class of models for permutations based on gamma random variables. Stern's approach is to estimate the distribution of finishing position assuming a given first place probability for each individual. The results of the model are applied to horse racing. Our model most closely resembles Stern's model, except that we use a *last* place probability for each individual, which results in a better fit for car racing, where early exits from a race are common (due to wrecks, engine failure, etc.).

Little work has been done on hierarchical modeling of interactions. Several authors indicate a desire to fit interaction terms in a hierarchical sense (Berry, Reese, and Larkey 1999, for example), but are limited by the type of data collected. Spiegelhalter, et al. (1999) present Bayesian hierarchical models for problems posed by Breslow (1984), Breslow and Clayton (1993), and Clayton (1994) that were originally analyzed under a more classical framework. In this work we present an approach to modeling interactions in a hierarchical

framework and in the next section we describe the data that support our specification.

2. DATA

We obtained race data from 1996-2000 from NASCAR's web site, www.nascar.com, for the WC, BGN, and CT series. Typical numbers of races per year are 34 for WC, 32 for the BGN series, and 25 for CT. Data quality for the WC series is essentially flawless. Periodically the BGN data are missing certain finishing positions (in these rare cases we act as if the missing drivers did not participate at all and move the remaining drivers up one finishing position if necessary). The CT data are the most incomplete; some results from past years are missing, and archived data include fewer data fields.

Available data fields include finishing position, starting position (determined by speed in a qualifying lap), driver, car number, sponsor, car manufacturer, number of laps completed, indicator of whether driver led any laps, indicator of whether driver led more laps than any other driver, status of driver at end of race (either "Running" or a reason why the car wasn't running), money and championship points earned (points are determined by finish position and the indicators of laps led), date of race, and the track (and therefore various qualities of the track, such as its length, amount of banking in degrees, and qualifying speed). The data are available for download at the web site madison.byu.edu/racing/racing.html.

The WC series is more prestigious than the BGN series, but WC drivers (e.g. Mark Martin) frequently drive in BGN races, making it possible to compare the relative strengths of drivers in the two series. There is less cross-driving with the CT series, but there is some potential for comparison using multiple years of data and drivers who were promoted from the CT to either the BGN or WC series. These comparisons will need to allow for the possibility that some drivers improve over time.

3. FUNDAMENTAL MODEL

We use the model of Luce (1959), a natural extension of the Bradley-Terry paired comparison model (Bradley and Terry, 1952), and a modification of the models discussed in Stern (1990) as the basis of our analyses.

In each series we observe R_{ij} , the finishing position of the i th driver in the j th race. From these finishing positions we want to estimate the ability of each driver for a given race. Letting θ_{ij} represent the i th driver's

ability in the j th race, our simplest model is

$$\theta_{ij} \equiv \theta_i \text{ and } \theta_i \sim N(\mu, \sigma_\theta^2), \quad (1)$$

in which each driver has a single ability parameter that is constant across races. σ_θ^2 is given a prior distribution, and for our models, μ can be taken to be zero without loss of generality. In more sophisticated models, we will put structure on the θ_{ij} 's and analyze the parameters inherent in this structure.

To create finishing order from the θ_{ij} coefficients, we choose a driver to finish in last place with probability proportional to $\lambda_{ij} = \exp(-\theta_{ij})$. We then choose the second-to-last place finisher from the remaining drivers, again with probability proportional to λ_{ij} , and repeat this process until drivers i_1 and i_2 remain, and the second place finisher is chosen to be driver i_1 with probability $\lambda_{i_1j}/(\lambda_{i_1j} + \lambda_{i_2j})$. This model is appealing because it resembles what actually happens in a race, as drivers often drop out one by one; we refer to it as the ‘‘attrition model.’’ (Slower cars can also be viewed as ‘‘dropping out,’’ though perhaps not at well-defined times.) Let π_j be defined so that $\pi_j(k) = i$ means that driver i finished in k th place in race j . If the data set consists of J races and I_j drivers took part in race j , this model gives rise to the likelihood function

$$P(\pi|\lambda) = \prod_{j=1}^J \prod_{i=1}^{I_j} \lambda_{\pi_j(i),j} \left(\sum_{k=1}^i \lambda_{\pi_j(k),j} \right)^{-1}.$$

For example, consider a race j with three drivers, with $\lambda_{1j} = 1$, $\lambda_{2j} = 2$, and $\lambda_{3j} = 3$. Then the probability that driver 3 finishes third and driver 2 finishes second, $P(123)$, is $3/(1+2+3) \times 2/(1+2) = 1/3$. Similarly, $P(132) = 1/4$, $P(213) = 1/6$, $P(231) = 1/10$, $P(312) = 1/12$, and $P(321) = 1/15$.

The model has a further intuitive interpretation: one mechanism that would lead to this probability model for permutations is if each driver remains in the race for an exponential length of time with rate (inverse mean) parameter λ_{ij} . The finishing positions are then determined by ranking the ‘‘failure times.’’ (This follows from the facts that if Z_1 and Z_2 are independent exponentials with rates λ_1 and λ_2 , $P\{Z_1 < Z_2\} = \lambda_1/(\lambda_1 + \lambda_2)$, the minimum of several independent exponentials is exponential with rate equal to the sum of rates, and the memoryless property of the exponential distribution.) A consequence of this formulation is that the probability that driver i_1 finishes ahead of driver i_2 in race j is equal to $\lambda_{i_2j}/(\lambda_{i_1j} + \lambda_{i_2j})$, independently

of which other drivers are in the race. The model for the ranks also results from a more general model for failure times in which the various drivers have failure intensities given by $\lambda_{ij}(t) = \lambda_{ij}g_j(t)$, where the λ_{ij} 's are positive constants and the g_j 's are positive functions not varying across drivers, so we are not in effect assuming failure rates which are constant in time.

This formulation differs from the one in Stern (1990) in a subtle way: Stern's model also in effect determines the finishing order based on exponential random variables, but Stern's winner is the racer with the smallest exponential variable, while our winner has the largest. Stern's exponential variables are to be interpreted as lengths of time each competitor requires to complete the course, rather than as failure times. Stern's models also allow the ranking times to take on non-exponential distributions; in particular, a gamma time with integer shape parameter r can be interpreted as the time it takes to score r (independent exponentially distributed) points. Stern was motivated by horse racing, in which there is the appealing interpretation that each of these points represents a fraction of the race length. In auto racing, in which many cars that could have won the race see their chances end due to mechanical problems or crashes, the failure time interpretation is more appropriate.

An appealing characteristic of the attrition model is that it captures the fact that a strong finish in a race is a much more accurate measure of a driver's ability than a poor finish. While any driver can suffer an early mechanical failure or drop out due to an early accident, it is nearly impossible for a subpar driver to finish strongly. In the attrition model, the k th place finisher is in essence involved in $K - k + 1$ comparisons with other drivers. Consider a single race with K drivers and observed finishing positions $\pi_1(\cdot)$. The diagonal elements of the observed information matrix are then:

$$\mathcal{I}_{ii} = -\frac{\partial}{\partial \theta_i^2} \ln p(\theta | \pi_1(\cdot)) = \sum_{k=i}^K a_k + \frac{1}{\sigma_\theta^2} \quad (2)$$

where

$$a_k = e^{-\theta_i} \left(\sum_{m=1}^k e^{-\theta_{\pi_1(m)}} - e^{-\theta_i} \right) \left(\sum_{m=1}^k e^{-\theta_{\pi_1(m)}} \right)^{-2}$$

Since a_k is non-negative, the sum in Equation 2 will be larger for smaller values of i . That is, the better the finish, the greater the information about the driver's parameter, except that the information is the same for

the first and second place drivers, since $a_1 = 0$.

4. ANALYSIS STRATEGIES

In this section we include some discussion about how we analyze these data. We discuss our choice of prior distributions, our system for estimating Bayesian models, and which convergence diagnostics we use.

4.1 Prior selection

Our goal was to uncover differences between drivers and tracks that were driven by the data sets we used, so we chose priors under which the drivers and tracks were exchangeable. (If instead one were interested in integrating other information to attain better estimates of differences in driver abilities, one would use nonexchangeable priors, presumably based on results of races prior to 1996 or in other racing series.)

We did not, however, make any effort to use priors which were fully non informative, such as using a normal prior for θ_i with a large standard deviation. Instead we placed hyperprior distributions on these standard deviations and chose the hyperprior parameters to give reasonable quantiles for quantities like the probability of the 25th percentile driver finishing ahead of the 75th percentile driver. For example, consider the model with a single ability parameter for each driver $\theta_{ij} \equiv \theta_i$, where $(\theta_i|\sigma_\theta) \sim N(0, \sigma_\theta^2)$, and where σ_θ is exponential with mean b . To select the hyperparameter b , suppose that σ_θ is equal to its p th quantile given b (i.e. $\sigma_\theta = -b \log(1-p)$). In this case the probability that the π th percentile driver (whose θ is equal to $\sigma_\theta z_\pi$, where z_π is the π th normal quantile) beats the median driver (whose θ is zero) is $(1 + (1-p)^{bz_\pi})^{-1}$. If, for instance, we choose $b = 1$, the probability that the 90th percentile driver beats the median driver is 0.71 if σ_θ is equal to its prior mean, while this probability ranges from 0.53 to 0.95 if we allow σ_θ to range from its 10th to 90th percentiles. Our most common choice in what follows is $b = 1$; we can see from this discussion that this corresponds to a relatively non informative prior since this is a large range of probabilities.

The exchangeable prior strategy does have a weakness in that it does not take into account the fact that if drivers participate in only some of the races, they are likely to participate in the ones in which they would perform best. The effect is potentially to lead to overestimates of the abilities of these drivers. Several drivers specialize in road course racing and enter only those races; some drivers preferentially enter races at

their home tracks; and drivers who would be unlikely to do well can fail to qualify.

4.2 An Object-Oriented System for Bayesian Data Analysis

We performed the analyses presented in this paper using a software system written in Java being developed at Los Alamos (see Graves, 2001). The goals for this system include making it as easy as possible to implement MCMC analyses of new models. It updates parameters exclusively through Metropolis steps, so that conjugate priors and complete conditional distributions are not needed.

However, the software also allows parameters or groups of parameters to be updated in arbitrary ways, in principle including Gibbs steps if they are implemented by the analysts, but so far we have found it more useful to implement Metropolis steps where several (highly correlated) parameters are perturbed simultaneously to generate candidate moves. We used several types of complex Metropolis steps in the analyses presented here. A simple example would be adding a common random constant to the entire θ_i vector to prevent the average value of the θ_i 's from drifting slowly away from zero. We continue to study the circumstances under which complex update steps are useful in MCMC, and this software system provides an ideal framework for such studies, since it computes Metropolis(-Hastings) acceptance probabilities for arbitrary changes to the parameters. A typical MCMC algorithm generated by this system, then, loops over all the parameters, updating each with a single-component Metropolis step, and at the end of the loop, additional Metropolis-Hastings steps can optionally be added where candidates are generated by perturbing correlated parameters simultaneously.

The system does not yet contain ways of automatically generating step sizes for Metropolis steps, so in principle, a substantial amount of tuning of these step sizes could be necessary. We found that it was often straightforward to reduce the numbers of effective tuning parameters: for instance, we assumed that driver ability parameters could have step sizes that were a constant divided by the square root of the number of races in which the driver participated, and tuned the constant.

4.3 Convergence diagnostics

Addressing convergence of Markov Chains in complex models (large number of parameters) is non-trivial. To assess mixing, time series plots were created for each of the parameters, and all revealed good mixing of

the chains. Convergence of the chains were determined based on diagnostics proposed by Raftery and Lewis (1996). While these diagnostics cannot predict absolute convergence, the criteria set forth in Raftery and Lewis were all met completely for our study.

5. RESULTS

In this section we derive rankings for drivers based on their 2000 performances, estimate trends in drivers' abilities over time, and compare drivers in different racing series by taking advantage of the fact that some drivers participate in both. We also study track properties: whether tracks differ in predictability, which tracks are most alike, and whether there is a driver-track interaction.

5.1 Track-independent driver abilities

The simplest model for these data includes only an ability coefficient for each driver that does not depend on the race ($\theta_{ij} \equiv \theta_i$). We fit this model to the results of the 2000 seasons in all three series. This model allows us to ask whether a season convincingly demonstrates one race team's superiority to another or whether luck could explain their different degrees of success. For example, Jeff Gordon dominated the WC series in 1995-98 with forty wins and three championships, but he slipped to sixth in the season standings in 1999 and ninth in 2000 before winning another championship in 2001. We conclude that his 2000 season showed a significant decline in his racing team's ability parameter.

5.1.1 2000 Winston Cup season

Table 1 contains the results for the 2000 NASCAR Winston Cup season, in which Bobby Labonte won his first NASCAR championship. The hierarchical model employed here assumes that driver ability coefficients θ_i do not depend on the race and have normal prior distributions with mean zero and variance σ_θ^2 . The hyperprior for σ_θ is exponential with mean 1 as discussed in §4.1. The results are based on 10,000 MCMC iterations after 1000 iterations of burn-in. We note that correlations between parameters are modest. For purposes of comparison, we fit Stern's model as well as the attrition model. Thirty-one drivers missed two races or fewer, so comparisons among these drivers will not be affected by race selection bias.

Our results are similar to the WC series points standings, given in the "Official" column, with some

differences. Our model, the official standings, and Stern’s model differ in the extent to which they reward very good finishes as opposed to punishing very bad finishes. Our model is the most strongly influenced by victories and the least strongly affected by poor finishes, while Stern’s model is the reverse. In official NASCAR standings, winless Ricky Rudd edged out series win leader Tony Stewart for fifth place, while we place Stewart in third and Rudd in ninth, and Stern’s model lists Rudd in third and Stewart in eighth. (Racing fans and journalists frequently complain about how the NASCAR points system fails to encourage aggressive driving and winning races. Our model is not a serious candidate as a replacement, due to its relative complexity and the uncertainty involved with MCMC estimates. However, it might be promising to develop a simple system that approximates our ratings, by, say, regressing our posterior means on numbers of finishes in each position. The results could then be adjusted so that drivers always earn positive points by appearing in a race, so that there is no incentive to skip races.)

The biggest beneficiary of our method is Jeremy Mayfield, who is thirteen places lower in the official standings due to missing two races, being penalized for an illegal part in another, and otherwise having many very good and many very bad performances. An interesting question for future work is whether Mayfield has a higher variance in his performance than other drivers. Ron Hornaday had the highest ranking for a driver who drove in only one race; he finished thirteenth in that race. Ron Fellows and Kerry Earnhardt each finished last in the only race they drove in, but since all drivers have prior mean abilities of zero, a single last place does not translate to the lowest ranking (in posterior mean). (Fellows is a road course specialist, so there was a danger that we would overestimate his ability as he entered only a race on his best track, His poor finish solved this problem, unlike in the Busch season; see §5.1.2.) Darrell Waltrip, one of NASCAR’s all-time greats who retired after the season, had the lowest rating of any driver who drove in most of the races. The posterior mean and standard deviation of σ_θ were 0.82 and 0.09 respectively; σ_θ is smaller in the WC series than in the BGN or CT series, indicating that drivers are more evenly matched in the top series.

The eighth and ninth columns in Table 1 report results of 1000 simulated seasons where we recorded which drivers won races or season championships. To simulate a season, we pick a vector of θ s by choosing one of the MCMC iterations at random and obtaining race results according to the attrition model, assuming that the participants in each race are the same as those who were in the corresponding race in the real season.

These results help to address goodness of fit of the model, and the model matches well, with the average numbers of wins close to the observed values and with deviations in unsurprising directions (e.g. Stewart and Wallace won more races than expected, Earnhardt and Jarrett fewer). In the NASCAR ranking system, drivers accumulate points in each race toward the season championship, earning points for their finishing positions according to a piecewise linear function (175 for first, 150 for sixth, 130 for eleventh, and three points fewer for each spot below eleventh), plus five point bonuses for leading at least one lap or for leading more laps than any other drivers. Since we do not model leading laps, we approximated the leading lap points by awarding ten points to the race winner and none to anyone else. Simulated Bobby Labontes won a quarter of the simulated season championships, slightly more than simulated Jeff Burtons. We see that the probability of Dale Earnhardt winning the championship was about thirteen percent, which provides a starting point for discussions of the likelihood of his winning an unprecedented eighth championship had he survived. Jeff Gordon's θ posterior is not consistent with his performance as the dominant driver in 1996-8; something was genuinely different in 2000: in fact, his crew chief left the team in mid-1999. In 2001, Gordon and his new crew chief adjusted to each other as once again they convincingly won the championship.

For comparison, results from Stern's model are also given in Table 1, including the posterior mean, average number of wins in simulated seasons, and number of season championships won in 1000 simulated seasons. Bobby Labonte's rating is identical to ours to two decimal places, but Stern's model gives a much larger gap between Labonte and his pursuers: Labonte won a surprisingly high 77% of simulated championships while winning a surprisingly low number of races. In the actual season, Labonte had only one finish of 22nd or worse, explaining his dominance according to metrics that punish bad finishes, as Stern's model does. The numbers of wins for top drivers listed in Table 1 match simulated wins numbers from our model much better than for Stern's: the excess wins in Stern's model go to drivers such as Sterling Marlin or Darrell Waltrip who had mediocre or worse seasons in 2000. (Dale Earnhardt Jr.'s two wins looks like the biggest potential sign of lack of fit for our model, as we estimate his expected number of wins to be 0.22. However, one out of 48 Poisson(0.22) variables will be two or larger, and it is not surprising to have one outlier of this magnitude among 74 drivers.) In 283 of the 1000 simulated seasons under Stern's model, no driver won more than three races. At least since 1972, every WC series season has featured at least one driver who won five or more

races. For the attrition model, fourteen out of 1000 simulations called for four wins by the leader, and never did the leader have three or fewer. The attrition model's right tail for the maximum wins statistic is still too short compared to history: in thirteen out of thirty-one seasons, a driver had ten or more wins, while the attrition model predicts that this should have happened only 4.4 times. This is unsurprising since the drivers were more evenly matched in the 2000 data than in most years. These comparisons raise interesting questions about choosing the best predictive model. If our model predicts better, then drivers with high variability in their finishes should be expected to do better in the future than drivers who are consistent but rarely outstanding. This heuristic would indicate that it is easier for a racing team to overcome reliability and/or accident difficulties than to add enough speed to win.

Table 1. Selected results from simplest model, fit to 2000 Winston Cup Results. Shown are the driver's rank according to our posterior mean and according to the season points system, the posterior mean and standard deviation of θ_i for our model, the posterior mean for Stern's model, and the numbers of races the driver participated in and won in the 2000 season. The ninth and eleventh columns contain results from simulations based on our model: the average number of wins by the driver per simulated season, and the number of championships won by the driver in 1000 simulated seasons. The tenth and twelfth columns contain simulation results using Stern's model.

<i>Atr</i>	<i>WC</i>	<i>Driver</i>	<i>Post. Mean</i>	<i>SD</i>	<i>Races</i>	<i>Actual Wins</i>	<i>Sim. Wins</i>	<i>Sim. Titles</i>			
			Atr	Stern			Atr	Stern	Atr	Stern	
1	1	Bobby Labonte	1.64	1.64	0.21	34	4	4.86	3.00	266	770
2	3	Jeff Burton	1.59	0.98	0.21	34	4	4.48	1.50	231	27
3	6	Tony Stewart	1.53	0.73	0.22	34	6	3.84	1.20	155	1
4	2	Dale Earnhardt	1.47	1.28	0.20	34	2	3.64	2.10	126	141
5	4	Dale Jarrett	1.40	0.98	0.20	34	2	2.99	1.60	78	22
6	7	Rusty Wallace	1.34	0.85	0.21	34	4	2.62	1.40	51	7
7	8	Mark Martin	1.25	0.39	0.20	34	1	2.16	0.80	30	0
8	9	Jeff Gordon	1.25	0.80	0.21	34	3	2.12	1.20	39	3
9	5	Ricky Rudd	1.21	1.03	0.20	34	0	1.94	1.70	22	27
10	10	Ward Burton	0.97	0.53	0.20	34	1	1.02	1.00	3	2
11	24	Jeremy Mayfield	0.96	-0.16	0.21	32	2	0.98	0.50	0	0
16	16	Dale Earnhardt Jr.	0.51	0.21	0.21	34	2	0.22	0.70	0	0
21	54	Casey Atwood	0.28	0.28	0.47	3	0	0.02	0.07	0	0
22	19	Sterling Marlin	0.27	0.21	0.20	34	0	0.10	0.70	0	0
23	61	Ron Hornaday	0.21	0.16	0.62	1	0	0.02	0.02	0	0
32	48	Kurt Busch	-0.08	0.02	0.37	7	0	0.01	0.10	0	0
61	72	Ron Fellows	-0.61	-0.67	0.81	1	0	0.01	0.01	0	0
63	73	Kerry Earnhardt	-0.64	-0.61	0.80	1	0	0.01	0.01	0	0
71	36	Darrell Waltrip	-0.85	-0.47	0.21	29	0	0.00	0.30	0	0
74	53	Jeff Fuller	-1.14	-0.54	0.41	7	0	0.00	0.07	0	0

5.1.2 2000 Busch Series season

We also perform the same analyses on the Busch series Grand National Division results, as seen in Table 2. The participants in this series are quite different in that there are a number of series regulars who drive in every race and compete for the championship, but since the races are often colocated with the Winston Cup races, several drivers from the stronger series participate in many Busch races and generally do well. Mark Martin was easily the highest rated driver in 2000, with five wins and thirteen top-six finishes in thirteen races. Jeff Burton and Matt Kenseth are also Cup drivers who were strong in Busch races. Jeff Green won the Busch championship by a record-setting margin of over 600 points; we find the posterior probability that his θ is larger than that of the second best Busch regular, Jason Keller, is about 0.988. In a later section we investigate the relative strengths of the three series in order to describe where Green's Busch performance would place him relative to Winston Cup drivers.

Ron Fellows and Butch Leitzinger each drove in only one race, at the road course at Watkins Glen, New York, and finished first and second in that race. For that performance they receive sixth and twelfth place rankings from our method. This points out a weakness in the simple model, which assumes that drivers are equally skilled on all tracks. Drivers that enter only a few races are likely to choose the tracks for which they are best suited, and that is certainly what happened with road racing specialists Fellows and Leitzinger. Again we used the unit exponential hyperprior distribution for σ_θ . Here the posterior mean and standard deviation for σ_θ were 0.93 and 0.09. Since σ_θ is larger for Busch races, there is a larger range of abilities for drivers in Busch races than in Cup races, as one expects to find given the presence of stars like Mark Martin.

From the simulated seasons, we see that Jeff Green's probability of winning the championship was over 88%, and that his six wins were actually fewer than expected of such a dominant driver. We estimate that Fellows' probability of winning the single race he entered (and won) was about one in six. Another example of the difference between our ranking method and the official points standings is the case of Todd Bodine and Kevin Harvick. Harvick finished ahead of Bodine in the official standings, and even though he won more races and missed a race, factors that would ordinarily favor him in our model, we invert the ordering and rank Bodine (slightly) higher. The explanation is in the quantile-quantile plot of their finishing positions in Figure 1. Bodine had many more bad finishes compensated by many more top-five finishes (six additional

Table 2. Selected results from simplest model fit to 2000 Busch series Results. Shown are the driver's rank according to our posterior mean and according to the season points system, the posterior mean and standard deviation of θ_i , and the numbers of races the driver participated in and won in the 2000 season. The last two columns contain results from simulations based on our model: the average number of wins by the driver per simulated season, and the number of championships won by the driver in 1000 simulated seasons.

Rank	Official	Driver	Mean	SD	Races	Wins	Wins	Titles
1	27	Mark Martin	2.746	0.349	13	5	4.72	
2	29	Jeff Burton	2.427	0.327	14	4	3.52	
3	17	Matt Kenseth	2.089	0.281	20	4	3.51	
4	1	Jeff Green	2.041	0.232	32	6	7.41	884
5	57	Jeff Gordon	1.587	0.456	5	1	0.29	
6	84	Ron Fellows	1.519	0.670	1	1	0.16	
8	2	Jason Keller	1.405	0.219	32	1	2.52	71
10	4	Todd Bodine	1.254	0.213	32	1	1.80	32
12	87	Butch Leitzinger	1.197	0.628	1	0	0.10	
11	3	Kevin Harvick	1.215	0.230	31	3	1.52	9
16	5	Ron Hornaday	0.911	0.211	32	2	0.77	4
30	8	Casey Atwood	0.419	0.204	32	0	0.16	
52	21	Buckshot Jones	-0.007	0.212	32	0	0.03	
55	22	Dick Trickle	-0.098	0.202	32	0	0.03	
62	100	Randy Tolsma	-0.200	0.728	1	0	0.011	
66	28	Lyndon Amick	-0.245	0.213	29	0	0.008	
109	117	Ken Schrader	-0.803	0.922	1	0	0.006	
120	38	P. J. Jones	-0.983	0.254	20	0	0	
124	55	Derrick Gilchrist	-1.109	0.366	9	0	0.001	

top 5's and six additional finishes of thirtieth or worse).

5.1.3 2000 Craftsman Truck season

In Table 3 we present the results for the third major NASCAR series, the Craftsman Truck series. Again our results are similar to the season points standings, and it is more obvious than in the Busch series because Winston Cup drivers participate in truck races much less frequently. We were somewhat less impressed than the points system with Andy Houston's results and rated Mike Wallace and Jack Sprague slightly higher than Houston, because Houston had no finished lower than 26th, while Wallace was 30th or worse twice and Sprague three times. Poor finishes are not influential in our model, which we chose since any driver can have early mechanical failure or be in an early crash. Ron Fellows finished third in the Watkins Glen race after a series of mishaps.

Phil Bonifield is in a class by himself. His expected θ of -2.498 compares to the second lowest expected θ of -1.314 for Robert Hillis. According to our MCMC iterations, the probability that Bonifield is better

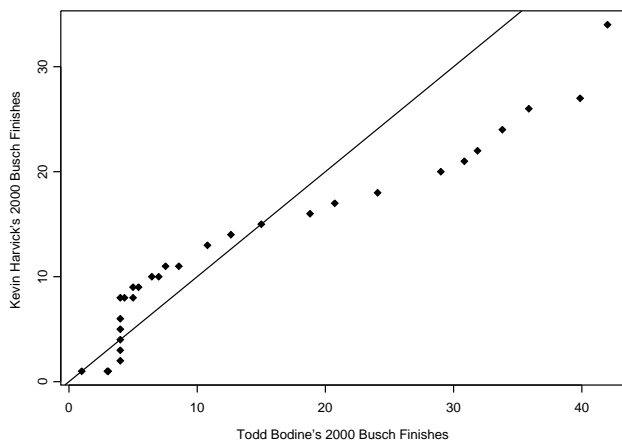


Figure 1. Quantile-quantile plot of finish positions for Todd Bodine and Kevin Harvick in 2000 Busch Series races. Our method ranks Bodine slightly higher despite Harvick's higher points finish.

than Hillis is less than 10%, while the probability that Bonifield is worse than all other drivers is 67%, a remarkably high probability given that there are many drivers with large uncertainty in their θ s as a result of driving in only a few races. In eight races Bonifield finished ahead of a total of ten drivers.

Again we placed a standard exponential hyperprior on σ_θ . The posterior mean and standard deviation of σ_θ are 0.99 and 0.10, so that the truck series drivers appear to be less evenly matched than the Busch series drivers. From the simulated results we see that Biffle was easily the most likely driver to win the championship, and Fellows' third place finish in his only race indicates that he had a 9% chance of winning that race. Bonifield won none of his 8000 simulated races.

5.2 Who are the future stars? Trends in driver ability

Fitting the models of the previous section on other individual years or on several years simultaneously makes it clear that driver abilities are not constant over time. Berry, Reese and Larkey (1999) use hierarchical modeling to estimate typical ability trends as a function of age in baseball, golf, and ice hockey. Since we have only five years of data available and since drivers often have long careers, sometimes performing at top levels into their fifties, we will be more modest and look for linear trends in ability over this period of time. We focus on the CT series, since it tends to have many young, rapidly improving drivers.

Table 4 contains the results, based on 10000 MCMC iterations after 1000 iterations of burn-in. After collapsing all 83 drivers who drove in only one race into a single "Miscellaneous" driver, 224 drivers remained

Table 3. Selected results from simplest model fit to 2000 Craftsman Truck Series results. Shown are the driver's rank according to our posterior mean and according to the season points system, the posterior mean and standard deviation of θ_i , and the numbers of races the driver participated in and won in the 2000 season. The last two columns contain results from simulations based on our model: the average number of wins by the driver per simulated season, and the number of championships won by the driver in 1000 simulated seasons.

Rank	Official	Driver	Mean	SD	Races	Wins	Wins	Titles
1	1	Greg Biffle	2.214	0.274	24	5	5.25	432
2	2	Kurt Busch	2.076	0.263	24	4	4.13	245
3	4	Mike Wallace	1.832	0.257	24	2	2.62	85
4	5	Jack Sprague	1.825	0.257	24	3	2.55	97
5	3	Andy Houston	1.808	0.258	24	2	2.55	92
6	6	Joe Ruttman	1.623	0.259	24	3	1.73	33
7	80	Ron Fellows	1.555	0.623	1	0	0.09	
8	7	Dennis Setzer	1.474	0.244	24	1	1.15	8
9	8	Randy Tolsma	1.432	0.245	24	1	1.05	10
10	63	Lyndon Amick	1.233	0.579	2	0	0.08	
11	9	Bryan Reffner	1.178	0.248	24	1	0.54	
12	42	Bobby Hamilton	1.142	0.468	5	1	0.15	
13	10	Steve Grissom	1.086	0.237	24	0	0.45	2
15	29	Ricky Hendrick	1.028	0.397	6	0	0.10	
19	20	Scott Riggs	0.838	0.261	17	0	0.14	
39	16	Lance Norick	0.067	0.238	24	0	0.005	
40	21	B. A. Wilson	0.019	0.257	21	0	0.004	
57	19	Randy MacDonald	-0.250	0.237	24	0	0.002	
102	25	Ryan McGlynn	-0.980	0.269	18	0	0	
106	39	Phil Bonifield	-2.498	0.495	8	0	0	

in the data set. (The Miscellaneous driver includes such stars as Bobby Labonte, Rusty Wallace, and Tony Stewart, as well as many far less talented drivers.) The model assumes that driver i 's ability in year 1998 is equal to α_i , and that driver i improves by an amount β_i each year: $\theta_{ij} = \alpha_i + \beta_i(y_j - 1998)$, where y_j is the year in which race j took place. The α_i s have a $N(0, 1)$ prior distribution, while the β_i s have a $N(0, 0.3^2)$ prior distribution, independent of the α_i s. The table contains posterior means and standard deviations of α and β for selected drivers, the rank of each driver's posterior mean ability for 1996, 1998, and 2000, and the number of races driven in each of the five seasons.

The fastest improving driver was Greg Biffle, who did not race in the CT series in 1996 or 1997, finished eighth in the standings in 1998, second in 1999, and comfortably first in 2000. Still more interesting was the second fastest improving driver, Kevin Harvick, who left the CT series after a 12th place finish in 1999. Had we fit this model at the end of 1999, we would have been much less surprised by Harvick's third place finish in the BGN series standings; that performance was far better than he had done in the CT series, but his improvement in CT racing was steady and fast. In 2001, Harvick had a strong rookie year in the WC series (he took over for Dale Earnhardt), while also winning the BGN series championship. We might have expected Biffle to do well in the BGN series in 2001 (he finished a strong fourth), and Andy Houston, who had the third highest improvement rate, to do well in the WC series (he did not, being released by his team after qualifying for only seventeen races with a best finish of 17th). The best way to earn a highly negative β coefficient was to perform brilliantly in 1996 and to mostly stop racing in the CT series afterwards, as 1996 third place finisher and 1997-2000 WC series driver Mike Skinner did. (Of course, if we included Cup results in this analysis, we might no longer claim that Skinner was regressing, since he has performed solidly in his new series.) Joe Ruttman, now in his late 50s, remained largely consistent over the five years.

5.3 Comparing the three series

Since our model specifies how finishing order is determined for a race containing any subset of drivers, it is natural to analyze races from several different series simultaneously so long as one is willing to make the assumption that a driver's ability in a Winston Cup car is the same as his ability in a truck. By doing so we can address questions such as where Jeff Green's record-setting performance in the 2000 Busch series

Table 4. Trends in driver ability over time. From data for Craftsman Truck series, 1996-2000. The table lists posterior means and standard deviations for α_i and β_i , where the ability for driver i in a race in year j is $\alpha_i + (j - 1998)\beta_i$. We also list the numbers of races by the drivers in each of the five seasons, and the ranks for the posterior means in 1996, 1998, and 2000. The drivers shown are listed in order of estimated 2000 ability $(\alpha + 2\beta)$.

Driver	$\widehat{E}(\alpha)$	$SD(\alpha)$	$\widehat{E}(\beta)$	$SD(\beta)$	Races	Ranks
Greg Biffle	1.53	0.20	0.49	0.15	00/00/25/25/24	46/11/01
Kurt Busch	1.61	0.54	0.28	0.25	00/00/00/00/24	23/06/02
Ernie Irvan	1.83	0.36	0.16	0.24	03/01/02/01/00	10/03/03
Stacy Compton	1.39	0.14	0.32	0.14	00/22/25/25/00	38/15/04
Rich Bickle	1.69	0.27	0.14	0.18	24/22/04/02/00	12/05/05
Jack Sprague	2.10	0.11	-0.07	0.08	24/22/25/25/24	04/01/06
Mike Wallace	1.24	0.14	0.35	0.11	00/14/25/25/24	49/18/07
Ron Fellows	1.43	0.31	0.21	0.20	00/04/04/04/01	30/14/08
Andy Houston	1.11	0.17	0.36	0.13	00/04/25/25/24	59/19/10
Kevin Harvick	0.78	0.14	0.49	0.15	04/12/24/25/00	123/35/13
Joe Ruttman	1.60	0.11	0.01	0.08	24/22/25/19/24	09/07/14
Ron Hornaday	2.02	0.13	-0.22	0.10	24/22/25/25/00	02/02/15
Randy Tolsma	0.94	0.13	0.23	0.10	01/16/24/25/24	54/29/16
Mike Bliss	1.46	0.13	-0.09	0.10	24/22/25/25/00	07/13/18
Mark Martin	1.77	0.68	-0.31	0.27	02/00/00/00/00	03/04/20
Mike Skinner	1.56	0.41	-0.45	0.21	24/03/02/00/00	01/09/42
Tammy Jo Kirk	0.06	0.25	-0.07	0.24	00/16/12/00/00	71/91/100
Mike Swaim	-0.13	0.59	-0.00	0.28	00/00/00/03/00	106/112/111
Dave Rezendes	0.55	0.29	-0.40	0.18	24/19/06/00/00	13/53/136
Miscellaneous	-0.55	0.13	-0.02	0.08	**/**/**/**/**	174/175/175
Jerry Glanville	-0.63	0.31	-0.05	0.18	05/02/02/04/00	180/189/194
Mike Hurlbert	-1.69	0.57	0.31	0.26	12/01/00/00/00	224/223/209
Phil Bonifield	-2.56	0.43	-0.28	0.25	00/00/00/13/08	223/224/224

places him relative to Winston Cup drivers, how many Cup drivers are likely to be more skilled than Trucks champion Greg Biffle, and so on. The success of this modeling depends on to what extent the same drivers participate in each series. The Busch and Cup series can be compared reliably, because there are many weekends in which the two series have races on the same track on Saturday and Sunday, respectively, so that several drivers, including Mark Martin, Jeff Burton, and Matt Kenseth, drive in both. The Trucks feature less moonlighting, so to obtain reliable relative ratings, it is necessary to use multiple years of data containing promotions from the Trucks to the other series. In order to use multiple years of data, it is important to allow for improvement, otherwise Kevin Harvick's 12th place trucks finish in 1999 and his 3rd place Busch finish in 2000 would lead to an overestimate of the strength of the Truck series.

Again, the prior distribution for the α s was taken to be standard normal, while the prior distribution for

the β s was mean-zero normal with standard deviation 0.3.

Table 5 contains the results of estimating driver abilities using the 1998 through 2000 seasons for all three series, allowing driver abilities to change linearly between years. The results are presented in order of estimated ability in 2000. Rankings within series are largely similar as in previous analyses. Tony Stewart is ranked even higher than in the analysis with the 2000 data alone, presumably because he drove in the Busch series in 1998 with indifferent results before an exceptional Winston Cup rookie season in 1999, suggesting rapid improvement.

Jeff Green, who won the Busch championship by a record margin, achieves an eighth-place ranking overall, which is perhaps higher than we think appropriate. Other 2000 Busch series notables are runner-up Jason Keller (18th overall), fourth place Todd Bodine (22nd), third place Kevin Harvick (24th), and fifth place Ron Hornaday (34th). Truck series stars rank as follows: champion Greg Biffle (13th), runner-up Kurt Busch (19th), fifth place Jack Sprague (26th), seventh place Dennis Setzer (28th), fourth place Mike Wallace (29th) and third place Andy Houston (35th). The overall conclusion is that the Busch series is slightly stronger than the Truck series: for example, thirteen 2000 Busch regulars are rated higher in 2000 than the tenth highest rated Trucks regular. Many of these highly ranked drivers in the Busch and Truck series are driving Winston Cup cars in 2001, and it is interesting to compare our estimated rankings to their standings positions; the two rankings should be close, with a slight tendency for drivers to be ranked better in the 2001 standings if good drivers are still in the lower series. The positions of these drivers in the final 2001 Winston Cup standings were 29th for Todd Bodine, 9th for Harvick, 38th for Hornaday, 27th for Kurt Busch, 34th for Mike Wallace, and 46th for Houston. Our rankings are generally close to the 2001 season rankings; there appears to be a slight tendency for us to overrate Busch and Truck drivers overall. Harvick did unexpectedly well because he had the unusual opportunity to join a championship caliber team (Earnhardt's) as a rookie. We estimated Jeff Green to be the fastest improving driver; he improved by moving from an underfunded Winston Cup team to a strong Busch team.

These rankings are based on the assumption that drivers are equally good in each series in which they participate. A reason this assumption might be faulty is that if a driver is a regular in one series, he is likely to have a solid team in that series, while his ride in a couple of races in another series might be significantly

Table 5. Comparing the strengths of drivers in various series. Uses data from all three series from 1998-2000, allowing drivers to improve or regress linearly in time. Drivers are ranked in order of 2000 ability. 224 drivers are in the data set, including the Miscellaneous driver, a combination of all drivers with three or fewer races. The last column lists number of races in the Winston Cup, Busch series, and Craftsman Truck series respectively.

Rank	Driver	$E(\alpha)$	$SD(\alpha)$	$E(\beta)$	$SD(\beta)$	Races
1	Bobby Labonte	2.42	0.12	0.25	0.13	101/6/0
2	Tony Stewart	1.95	0.13	0.60	0.14	68/22/0
3	Jeff Burton	2.39	0.10	0.15	0.12	101/41/0
4	Dale Jarrett	2.36	0.11	0.09	0.12	101/13/0
5	Mark Martin	2.46	0.10	-0.03	0.12	101/42/0
6	Dale Earnhardt	1.97	0.11	0.38	0.12	101/0/0
7	Jeff Gordon	2.60	0.13	-0.43	0.13	101/11/0
8	Jeff Green	1.33	0.11	0.79	0.14	23/70/0
12	Matt Kenseth	1.67	0.11	0.12	0.11	40/83/0
13	Greg Biffle	1.40	0.14	0.37	0.16	0/0/74
16	Dale Earnhardt Jr.	1.70	0.13	-0.20	0.13	39/63/0
18	Jason Keller	0.86	0.12	0.54	0.13	0/95/0
19	Kurt Busch	1.21	0.34	0.11	0.28	7/0/24
22	Todd Bodine	1.09	0.11	0.16	0.12	26/77/0
24	Kevin Harvick	0.76	0.12	0.46	0.13	0/32/49
26	Jack Sprague	1.37	0.13	-0.18	0.14	0/5/74
28	Dennis Setzer	0.91	0.13	0.19	0.15	7/3/61
29	Mike Wallace	0.81	0.12	0.26	0.13	3/15/74
34	Ron Hornaday	1.19	0.12	-0.24	0.13	3/40/50
35	Andy Houston	0.82	0.12	0.12	0.14	5/2//74
50	Robert Pressley	0.49	0.10	0.19	0.11	92/21/0
51	Randy Tolsma	0.56	0.13	0.11	0.14	0/1/73
52	Casey Atwood	0.59	0.12	0.04	0.15	3/76/0
210	Miscellaneous	-0.91	0.08	-0.17	0.08	9/78/138

less reliable.

These results are based on 5000 iterations after 1000 iterations of burn-in, and the convergence diagnostics proposed by Raftery and Lewis were satisfied.

This methodology could potentially be used to recommend to car owners whom they should hire from a lower series to drive for them. NASCAR's web site also retains race results from series such the Winston West, Goody's Dash, and Featherlite Southwest, and these are proving grounds for future stars.

A complicating factor that we do not pretend to address is that when a driver changes series, one expects their performance to fall off relative to their ability as they learn a new vehicle type and new tracks. The drivers will also get new crew chiefs, engine makers, and other sources of variability.

5.4 Do different tracks have different variances?

We now explore track effects. Before modeling track-driver interactions in the next section, we first ask the question of whether certain tracks' results are more predictable than others'. Let $T(j)$ denote the track for race j , and let $\lambda_{ij} = \exp(-\phi_{T(j)}\theta_{ij})$. The $\phi_{T(j)}$ parameters are then precision parameters, which are large if races on track $T(j)$ are relatively predictable. For the purposes of this analysis we took driver abilities to be independent of the tracks ($\theta_{ij} \equiv \theta_i$), and to have the standard Gaussian prior. We took the prior distribution for the ϕ 's to be exponential with mean b_ϕ , where b_ϕ has a standard exponential hyperprior.

We studied this model using five years of WC series data (1996-2000). We made this choice because we wanted to limit the number of drivers and to assume that their abilities remained constant over the races in the data set. Table 6 contains the results for the 22 tracks in the data set. We ran the chain for 5000 iterations after 1000 iterations of burn-in. There are a total of 136 parameters including 113 driver abilities, 22 track predictabilities, and b_ϕ . The posterior mean and standard deviation for b_ϕ were 2.42 and 0.61. In this analysis, the driver abilities were also compressed relative to previous analyses, so that the ϕ s can all be larger than one. The standard deviation of the set of posterior means of driver coefficients is 0.36, while this quantity was about one in the previous analyses.

There are some potential identifiability issues with this model. The functions of θ and ϕ that are estimable from the data are of the form $\phi_j(\theta_{i_1} - \theta_{i_2})$. Another way of saying this is that the likelihood is invariant to increasing the scale of the ϕ s by a positive constant and decreasing the scale of the θ s by the same constant. Consequently, the prior alone determines whether, for example, all the ϕ s should be very large while all the θ s are close together. Identifiability issues like these can have two consequences: first, the parameters can be hard to interpret. In this case, ratios of posterior means in Table 6 are more interpretable than the posterior means themselves, and all of the ϕ s are positively correlated so the marginal standard deviations lead to overestimates of the variability in ratios of ϕ s. Second, poor identifiability can lead to poor convergence properties of MCMC algorithms with componentwise parameter updates, and indeed we did observe poor mixing of, for example, the mean of the ϕ s and the variance of the θ s when we used only componentwise Metropolis updates. One way to address this problem is to insist that the average value of the ϕ s is always equal to one and modify the prior distribution accordingly. Instead, we chose to use simple prior distributions

and a new MCMC algorithm with componentwise Metropolis steps as before, and additional Metropolis steps designed to improve mixing of the MCMC algorithm. Specifically, we added the following two steps: in the first step, our Metropolis candidate is obtained by generating a single Gaussian random variable with specified standard deviation, and adding this common value to all of the θ_i 's. This keeps the average value of the θ s around zero and hence consistent with the prior. In the second step, we generate a random Gaussian W , and construct the new candidate values of the parameters as follows: let $\bar{\theta}$ be the mean of the current values of the θ_i 's, and let $\theta_i^C = \bar{\theta} + \exp(W)(\theta_i - \bar{\theta})$, $\phi_t^C = \exp(-W)\phi_t$, and $b_\phi^C = \exp(-W)b_\phi$ where the superscript C refers to "candidate." The parameters are then changed to their candidate values with the appropriate Metropolis-Hastings probability. These two moves do not change the value of the likelihood, only the prior. The Java system we are using makes it easy to add Metropolis-Hastings steps in which multiple parameters are modified at once. In the present example, we do obtain very good mixing, according to the Raftery and Lewis convergence diagnostics: only one parameter was suspect, the ϕ parameter for North Wilkesboro Raceway, for which we only had one race of data. The ϕ_t 's are all estimated to be greater than their prior mean value of one, because the prior distribution for the θ s pulls them toward each other and hence encourages the ϕ s to be large.

Most of the tracks have similar values of ϕ , but some interesting things emerge. The most and fourth-most unpredictable tracks, Watkins Glen and Sears Point, are the only road courses in the data set. Road courses differ from other tracks in that they include more curves, including some *right* turns. Common perception is that driving talent differs for road and oval tracks (in fact, teams often hire specialists to substitute for their normal drivers on road courses). The second most unpredictable track, Talladega Superspeedway in Alabama, is genuinely unpredictable. It is NASCAR's fastest track with qualifying speeds in excess of 190 mph for a lap, and the track tends to equalize the cars because it is easier to build a car with good handling there. It is one of two tracks (Daytona being the other) where carburetor restrictor plates are used, which reduces rates of acceleration. Races at Talladega feature all the cars bunched together for essentially the entire race, where cars frequently can gain or lose many ranks rapidly if they line up together or fail to do so. Also, crashes at Talladega can easily include ten or more cars due to the close packing and high speeds.

One of our hypotheses is that the most challenging tracks will be most predictable. Of the predictable

Table 6. Predictability measures for Winston Cup tracks, listed from most to least predictable. The table contains the posterior mean and standard deviation for the predictability parameters ϕ_t , together with information about the tracks' characteristics: their lengths in miles, numbers of degrees of banking in the corners (not listed for road courses), and speed as measured in miles per hour for the fastest lap ever run in qualifying.

Rank	Track	$E(\phi)$	$SD(\phi)$	Length	Banking	Speed
1	Charlotte	3.22	0.63	1.50	24	186
2	Dover	3.03	0.59	1.00	24	160
3	Rockingham, NC	2.97	0.59	1.02	22	158
4	Michigan	2.94	0.58	2.00	18	191
5	Darlington, SC	2.92	0.58	1.37	23-25	174
6	Indianapolis	2.81	0.62	2.50	31	181
7	Bristol, TN	2.80	0.57	0.53	36	126
8	Pocono, PA	2.78	0.57	2.50	6-14	172
9	Richmond, VA	2.76	0.56	0.75	14	126
10	Las Vegas	2.68	0.65	1.50	12	173
11	North Wilkesboro, NC	2.61	0.80	0.63	14	?
12	California	2.56	0.61	2.00	14	186
13	Atlanta	2.54	0.55	1.54	24	197
14	Phoenix	2.51	0.54	1.00	11	135
15	Homestead-Miami	2.47	0.69	1.50	8	156
16	Martinsville, VA	2.45	0.52	0.52	12	95
17	Daytona	2.34	0.48	2.50	31	210
18	New Hampshire	2.18	0.46	1.06	12	132
19	Sears Point, CA	2.07	0.45	1.95	Road	99
20	Texas	2.04	0.52	1.50	24	192
21	Talladega, AL	1.79	0.39	2.66	33	213
22	Watkins Glen, NY	1.42	0.40	2.45	Road	121

tracks, Lowe's Motor Speedway in Charlotte, NC is famous for changing its properties during races as temperatures and cloud cover change, which would favor experienced teams and drivers, but it is also the site of NASCAR's longest race (600 miles), and a priori one expects that longer races provide more time for good drivers to assert their superiority regardless of track difficulty. The second, third, and fifth most predictable tracks, Dover Downs ("The Monster Mile"), North Carolina Speedway in Rockingham, and Darlington, SC ("The Track Too Tough to Tame") are also known for their difficulty; Rockingham and Darlington are difficult in part because their abrasive surfaces wear out tires quickly.

Another of our hypotheses is that short tracks are less predictable than longer tracks, because more congested racing can lead to more frequent crashes, even by good drivers. The data contain no evidence of this phenomenon (or its reverse), as the four tracks in the data set of length shorter than a mile (Bristol, Richmond, North Wilkesboro, and Martinsville) were all estimated to be of essentially average predictability.

5.5 How strongly do drivers' abilities depend on tracks? Which tracks are most similar?

Home field advantages are factors in most sports. In racing, this concept is much richer, since drivers potentially have a different ability on each track. Rusty Wallace seems to be the strongest driver on short tracks, while the late Dale Earnhardt was unmatched on large, fast superspeedways where restrictor plates are used and where drafting is critical. In this section we fit ability coefficients that are different for each (track, driver) combination. The hierarchical form of the model encourages drivers that have been successful in general to have large abilities even on tracks where they have been unsuccessful in limited opportunities. We can also post-process the track-driver interactions to explore which tracks are most similar, by measuring which drivers tend to do well on both tracks.

This is a much more general model, in which $\theta_{ij} = \gamma_{iT(j)}, (\gamma_{i1}, \dots, \gamma_{iT}) | \theta_i \sim N(\theta_i, b^2)$, and $\theta_i \sim N(0, b_\theta^2)$. (Recall that $T(j)$ denotes the track of race j .) The respective prior distributions for b and b_θ can measure our a priori opinions about the importance of track-driver interactions. We took b_θ to have a gamma distribution with mean 2 and standard deviation 0.5 and b a gamma distribution with mean 1 and standard deviation 0.1, so that a priori the ratio of across-driver standard deviation to within-driver, across-track standard deviation would be about 2:1. Another approach would be to assume that $\theta_1, \dots, \theta_T$ have a multivariate normal distribution with unknown covariance matrix, where the off-diagonal elements of the covariance matrix show which tracks are most similar to each other. We chose to instead analyze similarity of tracks by post-processing the sequences of $\gamma_{iT(j)}$'s obtained by the MCMC.

The results, based on the 1996-2000 WC data and 5000 MCMC iterations after 1000 iterations of burn-in, are presented in Table 7. We organize the results by track. We estimated that Jeff Gordon is in fact the best driver on every track. This hints that track-driver interactions may not be particularly important, but there are indications based on correlations between tracks that track-driver interactions are real but require many races to estimate because of the large amount of noise in racing. We list Gordon's posterior mean ability on each track, then the driver we estimated to be the second best at each track, and the difference between Gordon's posterior mean and his. We then list the driver who has the largest posterior mean difference between his track ability and his overall ability ($\gamma_{iT(j)} - \theta_i$), together with this difference. These results are satisfying since they agree with our perceptions of who has been successful at particular tracks. (The

authors have been using these results in a fantasy racing league, with reasonable results).

We also list, for each track, the track whose driver ability coefficients are most similar, by constructing the matrix of posterior means of $(\gamma_{iT(j)} - \theta_i)$'s and computing the correlation matrix of tracks. These results hint that the track-driver interaction phenomenon is real, since tracks we expected to be similar are estimated to be similar. The two restrictor plate superspeedways, Daytona and Talladega, tend to favor the same drivers. The track most like the road course at Watkins Glen is the other road course at Sears Point. The slowest oval, the short track at Martinsville, is similar to these road courses, as is the short track at Bristol. Another pair of tracks that we a priori expected to be similar were Darlington and Rockingham, two old Southern tracks that wear out tires quickly. The other track similarities are more difficult to explain, but no obvious nonsense emerges. At the time we completed the analysis we believed that the biggest surprise was the similarity between the three-quarter mile track at Richmond and mile track at Phoenix, but both of these tracks are relatively flat, and we have since heard driver Jimmy Spencer say that Phoenix was the most logical place to expect success from his car that won both BGN races at Richmond in 2001. The posterior mean of the ratio b_θ/b (a measure of importance of driver-to-driver variation relative to track-driver interactions) was 3.63, with standard deviation 0.53, so the probability that this ratio exceeds its prior mean of two is quite high. That is, track-driver interactions are smaller than we expected a priori.

5.6 Bayes Factors

In Section 5.1.1 we presented results for our model and the corresponding results for Stern's model and argued for our model based on the results themselves. We can use Bayes factors (Kass and Raftery 1995) to compare the two models in a more quantitative sense. Bayes factors are ideal for comparing two models p_1 (ours) and p_2 (Stern's) by computing

$$B_{12} = \frac{\int p_1(y|\Omega_1)\pi_1(\Omega_1)d\Omega_1}{\int p_2(y|\Omega_2)\pi_2(\Omega_2)d\Omega_2}, \quad (3)$$

where Ω_1 and Ω_2 are the parameters that identify the prior distributions π_1 and π_2 for models p_1 and p_2 , respectively. Using an optimal bridge sampling (Meng and Wong 1997) approach to calculate the numerator and the denominator, we find $\log B_{12} \approx 13.24$, so that the data suggest our model fits the 2000 WC series

Table 7. Track-driver interactions. We estimated Jeff Gordon to be the best driver on all tracks, so for each track we list his posterior mean $\gamma_{iT(j)}$. We also list the second best driver on each track together with the difference between Gordon’s abilities and theirs: these can be converted into probabilities that they beat Gordon using the function $1/(1 + \exp(x))$. The next two columns contain the names of the “specialists”, the drivers whose track-specific abilities are above their overall abilities by the largest amount, and the size of this deviation. Finally for each track we list the track most similar to it as measured by correlations between drivers’ posterior means. Refer back to Table 6 for descriptions of these tracks.

Track	Gordon	Driver#2	<Gordon	Specialist	Δ	Most Similar
Atlanta	1.82	Dale Jarrett	0.08	Bobby Labonte	0.31	Charlotte
Bristol	1.95	Mark Martin	0.30	Rusty Wallace	0.39	Sears Point
California	1.94	Dale Jarrett	0.38	Jeremy Mayfield	0.26	Pocono
Charlotte	1.87	Dale Jarrett	0.20	Ward Burton	0.19	Dover
Darlington	2.04	Dale Jarrett	0.30	Jeff Burton	0.25	Rockingham
Daytona	1.79	Dale Jarrett	0.08	Dave Marcis	0.24	Talladega
Dover	1.88	Mark Martin	0.20	Matt Kenseth	0.24	Homestead
Homestead	1.83	Tony Stewart	0.19	Tony Stewart	0.18	Dover
IndianapolisMS	1.87	Dale Jarrett	0.12	Bobby Labonte	0.17	Michigan
Las Vegas	1.80	Mark Martin	0.21	Jeff Burton	0.21	Darlington
Martinsville	1.95	Tony Stewart	0.47	Bobby Hamilton	0.34	Sears Point
Michigan	1.90	Dale Jarrett	0.10	Ernie Irvan	0.32	Indianapolis MS
New Hampshire	1.84	Jeff Burton	0.22	Joe Nemechek	0.33	Talladega
North Wilkesboro	1.96	Dale Jarrett	0.33	Terry Labonte	0.16	Rockingham
Phoenix	1.75	Dale Jarrett	0.13	Bobby Hamilton	0.20	Richmond
Pocono	2.01	Dale Jarrett	0.25	Jeremy Mayfield	0.28	Michigan
Richmond	1.85	Dale Jarrett	0.09	Dale Earnhardt Jr.	0.20	Phoenix
Rockingham	1.92	Dale Jarrett	0.10	Ricky Craven	0.34	Darlington
Sears Point	2.02	Mark Martin	0.37	Darrell Waltrip	0.24	Martinsville
Talladega	1.79	Dale Jarrett	0.12	Kenny Wallace	0.32	Daytona
Texas	1.69	Dale Jarrett	0.07	Dale Earnhardt Jr.	0.20	California
Watkins Glen	1.95	Mark Martin	0.41	Robby Gordon	0.36	Sears Point

data relatively well. We note that assuming the model that finishing position is purely a random arrangement of drivers versus our model provides a Bayes factor of $\log B_{12} \approx 3900$.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated a new class of models for permutations that have natural applicability in the analysis of auto racing results. One feature of our Bayesian formulation is that interaction terms can be estimated in a fully hierarchical manner. This hierarchical specification allows “borrowing of strength,” making interaction estimation in this framework particularly appealing. The results provide a basis for judging not only driving teams’ abilities, but the variation of driving teams’ abilities on different tracks. This has implications for sponsors who may choose whether to sponsor a team based on that team’s ability on those types of tracks.

Several points were made clearer through this analysis. First, the ability measured here is really team

ability and not necessarily driver ability. Most people associate cars with drivers, but it is all but impossible to separate the driver and driving team effects. The second point is that while we found one driver (Jeff Gordon) to be the best on every track, there is evidence that drivers have different abilities on different tracks; that is, there is a track-driver interaction. The third point is that our modeling approach permits us to identify “track specialists.” These may not be the best drivers on a particular track, but they are at their personal best at that track. This finding has implications for sponsors, as they may want to choose a specialist if they are going to underwrite an entire race. A potential danger for our methodology is that we can overestimate the ability of a driver if he is successful in entering races where his chances of a good finish are better than on other tracks. We think this is only a serious problem in the case of road course specialists. At the Winston Cup level, regular drivers miss races because they are hurt (and injuries happen at random for this purpose), because they lose their jobs (normally after a string of poor performances, in which case their ability is unlikely to be overestimated), or because they fail to qualify for the race (though this is rare due to provisional starting rules). In recent history, it is rare for drivers to preferentially enter races on particular types of oval-shaped tracks.

With regard to future work, we note that race weekends generate more measures of driver performance than just the finishing positions, including speed (and rank of these speeds) in qualifying and in two or three practice sessions. It would be of interest to model the relationships between these various rankings. It is not clear how to do this with our current model; in contrast, using the model of Johnson et al. (2000), one could model race performance as a Gaussian deviation from mean ability from which the ranks are observed. One could then estimate the properties of several correlated ranking mechanisms, although as mentioned before, asymmetric error distributions are probably needed to model racing. In any case, the correlations between these measures are likely to be low, and these other rankings probably do not help estimate driver abilities. However, the data also contain the indicators of whether drivers led laps in races and whether they led more laps than any of the other drivers, and these measures are likely to be helpful in estimating driver ability were we to find joint distributions for modeling finishing positions along with these indicators. An extreme example of modeling correlated rankings would be modeling the standings in a race as it progresses; the authors obtained such data at the time scale of roughly a minute for several 2000 races.