4F13 Coursework 2: Probabilistic Ranking

Part a)

Gibbs sampling facilitates estimating the posterior distribution of model parameters in Bayesian settings. It is a Markov Chain Monte Carlo (MCMC) method which is especially useful for complex models where direct sampling is challenging or impossible. Figure 1 shows skill samples drawn using a Gibbs sampling process from an 1801 game database for four arbitrary, 2011 ATP men's tennis players. A player's 'skill' quantitatively defines their ability, providing a way of ranking players (or at least a different way than the current ATP points system) and predicting match outcomes.

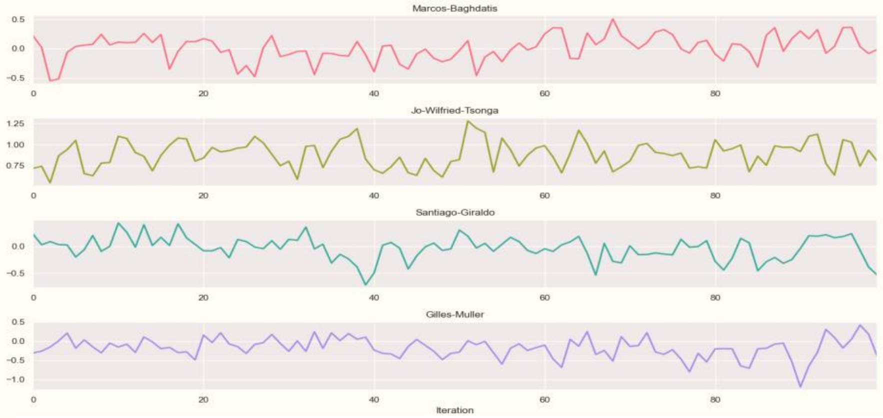

Figure 1: Skill samples across 4 players from the Gibbs sampling process against iteration

There is an initial period of uncertainty until the samples move into higher probability distribution regions. This is because Gibbs sampling can give biased estimates until it reaches a steady-state. As such, we want to discard samples drawn before the chain converges, known as 'burning-in'. Burn-in time refers to the time taken, or Gibbs iterations, for the chain's transitionary phase from its initial state to a state roughly in equilibrium with the target distribution. Finding the optimal burn-in period is crucial for an accurate chain: too short and the samples maybe be biased or simply unrepresentative, too long and there may be unnecessary computational expense.

We can overlap numerous skill graphs and deduce an appropriate burn-in time by inspection, overpredicting it if necessary. Additionally, analysing the stability of the joint log-probability against iteration of the chain should lead us to a similar conclusion (ref. fig.2 and fig.3). In our case, we come to a value of approximately 10 iterations before the chain moves into this higher probability region, although to provide margin for error and given our sample size, we could easily increase this to 50 steps for a minimal additional computational cost.

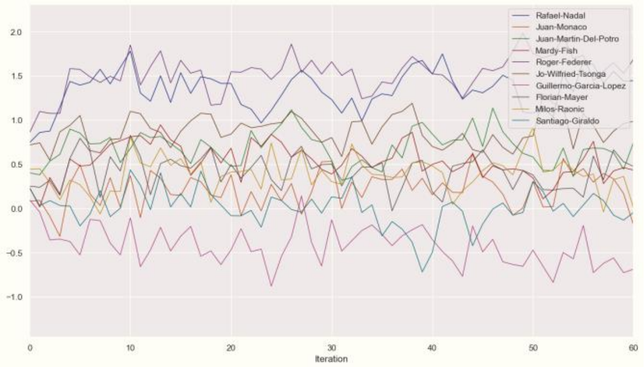

Figure 3: Skill samples for 10 players, noting initial transitionary period until approximately 10 iterations.

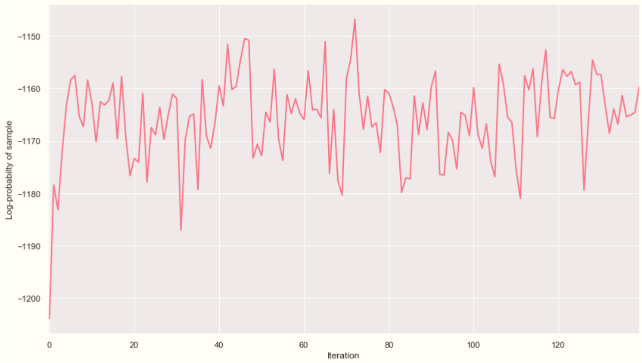

Figure 3: Log-probability of sample given iteration, noting the rapid rise until approximately 10 iterations.

When determining how long we need to run the Gibbs sampler for before the results are reliable, we need to consider the auto-correlation times as well as the burn-in times. Simply put, this is the number of steps required for the correlation between adjacent samples to decay to a predefined near-zero, value. A short time indicates samples become uncorrelated quickly, leading to more efficient sampling. From figure 4, we see that an iteration lag of approximately 10 steps is sufficient for auto-correlation to drop to zero. Using similar logic to above, we can conclude 50 Gibbs samples as sufficient burn-in for reliability.

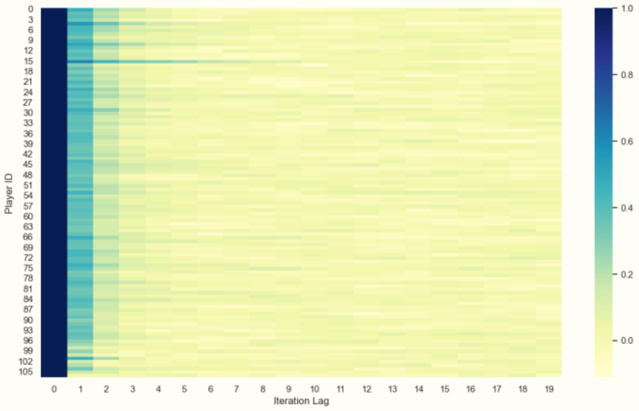

Figure 4: Auto-correlation of skill samples across all sampled ATP players against the iteration lag.

Part b)

In a Gibbs sampler, convergence occurs when the algorithm has mixed well with the state-space and the Markov chain reaches a stage where further samples are being drawn from the true distribution, or hence the sampler converges to the joint posterior distribution of the model parameters. Methods for judging Gibbs sampler convergence are described above. On the other hand, in Message Passing (MP) algorithms which incorporate graphical models, convergence relates to the iterative propagation of messages, or beliefs, between the nodes of the model until a state of equilibrium is reached. The model converges when following iterations do not result in significant change in belief values. Consider the maximum absolute change between subsequent estimates for the mean and precision against the number of iterations, as per figure 5. Once the change goes below a sufficiently small, arbitrary threshold, say 10^{-4}, then we can deem the process as having converged. As such, and in line with previous estimates for burn-in and autocorrelation times, we estimate that the MP algorithm converges after 50 steps. Also note the rapid decrease in error during the first 10 epochs.

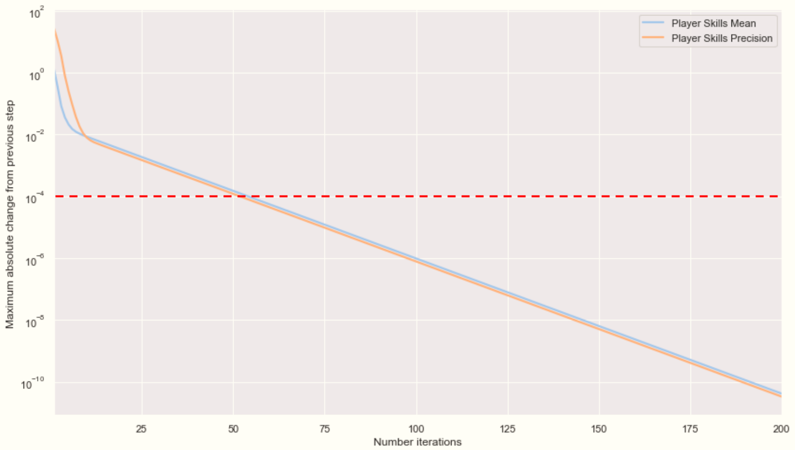

Figure 5: Maximum absolute change in mean and precision between consecutive steps of MP algorithm against iteration number. Tolerance line in red.

Part c)

We can use the posterior skills distribution approximated by the independent Gaussian from the MP algorithm to predict probabilities relating to skill and match outcome. The probability that player p_1 has a higher skill than player p_2 can be represented as:

$$p(\boldsymbol{w}_{p1} > \boldsymbol{w}_{p2}) = \Phi\left(\frac{\mu_{p1} - \mu_{p2}}{\sqrt{\lambda_{\mathrm{p1}}^{-1} + \lambda_{\mathrm{p1}}^{-1}}}\right)$$

Furthermore, the probability that a player p_1 wins against player p_2 can be represented as:

$$p(\boldsymbol{w}_{p1} - \boldsymbol{w}_{p2} + n > 0) = \Phi\left(\frac{\mu_{p1} - \mu_{p2}}{\sqrt{\lambda_{\mathrm{p1}}^{-1} + \lambda_{\mathrm{p1}}^{-1} + 1}}\right)$$

Where $n \sim \mathrm{N}(0,1)$ is the performance noise, accounting for external factors, μ_i is a mean, λ_i is a precision and Φ is the cumulative probability function.

We choose to focus on the top 4 players in 2011 across all ATP tournaments. We come to two tables, table 1 shows the probability that one player has a higher skill than another, while table 2 shows the probability of a player winning against another. Clearly, the player with higher skill is always expected to win, but our confidence when predicting a winner is much lower than when discussing who is more skilled. From above, the performance noise n will greatly influence the game outcome and generates uncertainty in our estimate, better reflecting the real outcomes of matches played between similarly skilled opponents, say Federer and Nadal.

Table 1: Probability that player 1 has a higher skill than player 2, using Message Passing (MP) algorithm.

		PLAYER 2			
PLAYER 1		Novak Djokovic	Rafael Nadal	Roger Federer	Andy Murray
	Novak Djokovic	-	0.940	0.909	0.985
	Rafael Nadal	0.0602	-	0.427	0.767
	Roger Federer	0.0911	0.573	-	0.811
	Andy Murray	0.0147	0.233	0.189	-

Table 2: Probability that player 1 wins against player 2, using Message Passing (MP) algorithm.

		PLAYER 2			
PLAYER 1		Novak Djokovic	Rafael Nadal	Roger Federer	Andy Murray
	Novak Djokovic	-	0.655	0.638	0.720
	Rafael Nadal	0.345	-	0.482	0.573
	Roger Federer	0.363	0.518	-	0.591
	Andy Murray	0.280	0.427	0.409	-

Part d)

When estimating a player's skill using Gibbs sampling, it is important to consider what method of approximation we are using. Three different approaches include: (1) assuming the skills are independent and fitting a Gaussian to each marginal probability, (2) assuming dependence and hence covariance, thus fitting a joint Gaussian to both probabilities or (3) approximating directly from the samples. Using these three methods, table 3 shows estimates of skill for only Nadal and Federer, more specifically the probability that Nadal's skill is greater than Federer's.

Table 3: Probability of #2 ranked Nadal's skills being higher than #3 ranked Feder, using approaches defined in part d)

	APPROACH (1)	**APPROACH (2)**	**APPROACH (3)**
$\boldsymbol{p(w_{Nadal} > w_{Federer})}$	0.597	0.615	0.619

Method (2) would yield better results than method (1) if the covariance between skills is non-zero. We show this is the case in figure 6, clearly the Gaussian-shaped marginals of both players lead to a symmetrical and strong pattern when plotted in a heatmap, indicating covariance.

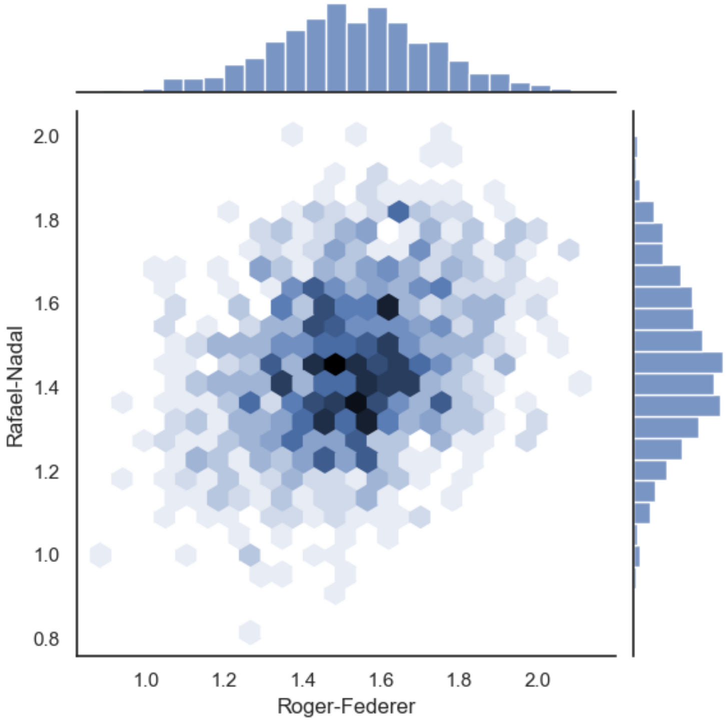

Figure 6: Heatmap of Gibbs samples for Nadal and Federer, with marginal distributions on the sides

In comparison, method (3) would estimate the probability directly from the data, without attempting to describe the distribution, making it the least biased and hence best method in this situation. Whilst the skills distribution for both Nadal and Federer may look Gaussian, method (3) does not rely on this assumption, averaging instead instances where $w_{Nadal} > w_{Federer}$ over N datapoints:

$$p(w_{Nadal} > w_{Federer})_{Method(3)} \approx \frac{1}{N} \sum_{i} w_{Nadal}^{i} > w_{Federer}^{i}$$

Finally, we can compute a new skills table using our chosen Gibbs sampler, method (3), and thus compare it to the Message Passing (MP) algorithm from part c) (ref. table 4).

Table 4: Probability that player 1 has a higher skill than player 2, using Gibbs samples.

		PLAYER 2			
PLAYER 1		Novak Djokovic	Rafael Nadal	Roger Federer	Andy Murray
	Novak Djokovic	-	0.957	0.934	0.989
	Rafael Nadal	0.043	-	0.381	0.740
	Roger Federer	0.066	0.619	-	0.808
	Andy Murray	0.011	0.260	0.192	-

Clearly both Gibbs sampling and the MP algorithm return very similar results but ultimately, the simplicity and capability for handling high-dimensional problems of Gibbs sampling is more ideal.

Part e)

Tying in all aspects of the above, we aim to rank the 2011 ATP players based on three different methods of inference: (1) empirical game outcome averages, (2) Gibbs sampling predictions and (3)

MP algorithm predictions. Method (1) will rely solely on observed historical data without consider the underlying models or uncertainties. Clearly, there are limitations when attempting to rank players with fewer matches played, or who have won very few/none of their matches. We would require more observations to address this, which is not always possible e.g. in a season/calendar year with a fixed maximum number of games.

Gibbs sampling (GS) and MP predictions are based on probabilistic models and as such are more sophisticated, accounting for uncertainties and model complexities. A notable difference between GS and MP is the context in which ranking is based on. Whilst GP considers player skills and accounts for random match-by-match variations, it places too much emphasis on solely the outcome: winning or losing, rather than the outcome with respect to the opponent. Should beating a higher skilled opponent push your ranking higher than say, beating two lower skilled opponents? MP models similar to those following the TrueSkill graphing can account for interdependencies between players and historical game outcomes and is arguably a better ranking system. Finally, methods (2) and (3) can also express the uncertainty in its ranking, as mentioned above. By comparing the mean variance across all players (0.271 vs 0.263), we further justify the proficiency of Message Passing algorithms.

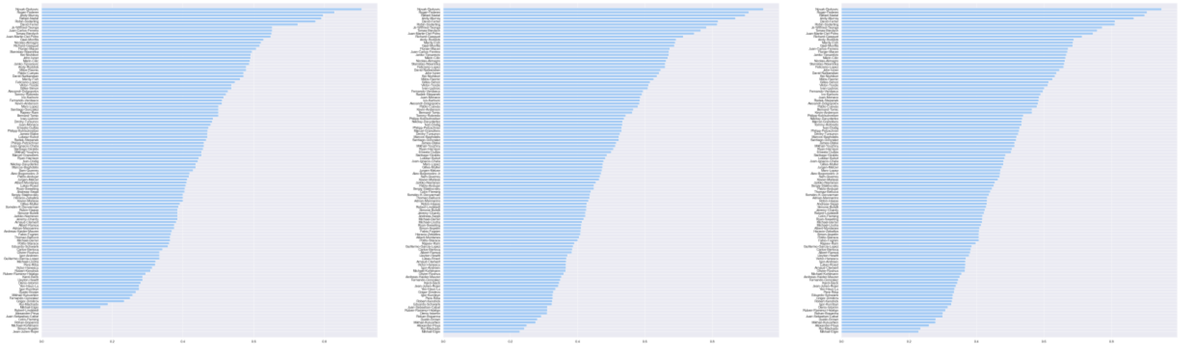

Figure 7a,b,c): Ranking of ATP tennis players based on expected skill and a) empirical game averages, b) expected output with MP algorithm, c) expected outcome with Gibbs sampling.