Competing Models*

José Luis Montiel Olea[†] Pietro Ortoleva[‡] Mallesh Pai[§] Andrea Prat[¶]

First version: October 2017
This version: June 25, 2019

Abstract

Different agents compete to predict a variable of interest related to a set of covariates via an unknown data generating process. All agents are Bayesian, but may consider different subsets of covariates to make their prediction. After observing a common dataset, who has the highest confidence in her predictive ability? We characterize it and show that it crucially depends on the size of the dataset. With *small data*, typically it is an agent using a model that is 'small-dimensional,' in the sense of considering fewer covariates than the true data generating process. With *big data*, it is instead typically 'large-dimensional,' possibly using more variables than the true model. These features are reminiscent of model selection techniques used in statistics and machine learning. However, here model selection does not emerge normatively, but positively as the outcome of competition between standard Bayesian decision makers. The theory is applied to auctions of assets where bidders observe the same information but hold different priors.

Keywords: Bayesian Linear Regression, competition, model selection, Akaike Information Criterion.

JEL: D81, C11, C52.

---

*We thank Sylvain Chassang, Kfir Eliaz, Annie Liang, George Mailath, Stephen Morris, Wolfgang Pesendorfer, Rani Spiegler, and the participants of many seminars and conferences for their useful comments and suggestions. Ortoleva gratefully acknowledges the financial support of NSF Grant SES-1763326. Pai gratefully acknowledges the financial support of NSF Grant CCF-1763349.
[†]Department of Economics, Columbia University. Email: jm4474@columbia.edu
[‡]Department of Economics and Woodrow Wilson School, Princeton University. Email: pietro.ortoleva@princeton.edu.
[§]Department of Economics, Rice University. Email: mallesh.pai@rice.edu.
[¶]Columbia Business School and Department of Economics, Columbia University. Email: andrea.prat@columbia.edu

1 Introduction

Consider a setting where decision makers with different priors compete to predict a variable y. To fix ideas, suppose that agents use a (possibly misspecified) statistical model that treats y as a linear function of a number of possible covariates $\{x_i\}_{i\in\{1,\dots,k\}}$ plus a noise term, i.e., $y = \sum \beta_i x_i + \epsilon$. For example, y could be a country's GDP growth, which agents are trying to predict using a long list of variables x. Both the β_i's and the variance of ϵ, are unknown.

Agents share the same quadratic loss function about their prediction, but use different *models*—different subsets of covariates as relevant to the prediction. In the GDP example, some may believe that relevant factors include education level and net trade surplus; others may also consider monetary supply and climate change data. Suppose all agents are Bayesian and update their prior after observing a common dataset: n draws of y and x from the unknown data generating process. What are the characteristics of the model of the agent that, after observing the data, has the highest confidence in its predictive ability, i.e., has the lowest posterior expected loss?

We provide a complete characterization and show that this depends both on the model dimension of the agent, i.e., how many variables are considered, as well as on the size of the dataset n. In particular, we show that with *small samples* the most confident agent is one using a model that is small-dimensional, possibly smaller than one that is properly calibrated. In contrast, with *big data*, the most confident agent instead uses a 'large-dimensional' model, with possibly more variables than the true data generating process. These two results are reminiscent of well-known properties of model selection techniques suggested in Statistics and widely used in Econometrics and Machine Learning. There is however one crucial difference in the approach: in statistics and machine learning, model selection criteria emerge *normatively*, as the optimal procedure with respect to some objective function; here instead they emerge *positively*, from a competition mechanism where the most confident Bayesian agent is selected.

Our results characterize the agent with the highest confidence in her predictive ability (lowest expected loss) according to her own (possibly misspecified) statistical model and prior. Studying the expected loss from the point of view of each agent is a key departure from the literature, which instead typically considers the objec-

tive expected loss, based on the *true* data generating process. As motivation, note that there are many competitive situations in which more confident agents acquire prominence: these are the agents who are willing to stake the most on their ability to forecast.

A practical example is a second-price auction in which agents are bidding to acquire a productive asset. The value of this asset to an agent depends on their ability to predict a given variable using a set of covariates. The asset could be ad-space on an online platform, the value of which depends on the sellers' ability to infer customers' preferences using their observable characteristics; or it could be a company, whose future value depends on how accurately the new owner is able to predict market conditions. All bidders observe the same data, but may use different variables to make this prediction, as they may have different priors, or models. The auction is clearly won by the agent who is most confident in her predictive ability, according to her own posterior after observing the dataset. In general, our results are useful to characterize the winner (and their model) in competitive situations in which a leading position is taken by those who are most confident in their predictive ability. Other examples include political competition, or board meetings, where the ability to credibly show one's confidence may lead to selection.

Summary of Results and Intuition. Our first result characterizes the expected posterior loss of an agent who has prior π and observes data D_n. We show that this loss can be decomposed as the sum of two components which we term: 1) *model fit*: the agent's posterior expectation of the variance of ϵ and 2) *model estimation uncertainty*: the degree of uncertainty that the agent has about each of the coefficients in its regression model. Crucially, we show that the latter in turn depends on the model dimension. This implies that while a Bayesian agent uses the Bayesian prior to compute the best action and does not care about the dimension of the model she is using, this very dimension affects her *confidence* in her own predictive ability.

This characterization has two immediate implications, depending on the size of the dataset. For clarity, consider first the case in which the dataset is large. In this case, the 'model estimation uncertainty' term vanishes: agents will have no uncertainty about their fitted parameters, even if they are using the wrong model. The comparison is therefore based only on *model fit*. As a result, incorrectly specified

models, i.e., models which omit an observable that is relevant for prediction, never prevail. At the same time, we show that larger models that contain additional observables that are irrelevant to the true data generating process (DGP) may continue to win, even asymptotically. Even though these larger models will converge to the properly calibrated ones, for any finite sample they remain strictly different, and we show that their probability of winning remains strictly above zero. Our results show that the prior does not vanish asymptotically: it continues to affect a large model's probability of winning even with infinite data.

Our second set of results pertain to the case of small datasets. Here 'model estimation uncertainty' plays a critical role. We show that the agent with the lowest loss will be one with a model that is of smaller dimension than the true DGP. This is because while agents with misspecified models may have a lower model fit, they will also have a lower model estimation uncertainty (as they have less parameters to estimate).

In order to establish the aforementioned result, we make additional assumptions. We assume that all agents' priors take the normal-inverse gamma form. We also assume that with no data all agents have the same expected prior loss. This guarantees that all heterogeneity comes from the different ways in which agents' confidence is affected by data – eliminating the possibility that differences in prior confidence drive our results.

First, we prove that when the dataset consists of a single data point, the winning model always involves exactly 1 observable. Deriving more general results is challenging, as with small samples we cannot exploit the distributional approximations adopted in the large-sample analysis. Small samples have two features: the dependence on specific data realizations, and the fact that the prior remains relevant. In our analysis, we want to preserve the second feature, but circumvent the first – the source of the difficulty in analytical tractability. To this is end, we use a non-standard asymptotic framework that allows the prior to 'drift' with the sample size.[1] We present asymptotic results in which we let the dataset grow but at the

[1] 'Non-standard asymptotics' that allow for the parameters of a statistical model to be indexed by the sample size have been used extensively in econometrics. The typical goal of an alternative asymptotic framework is to provide better approximations to finite-sample distributions of estimators, tests, and confidence intervals, while exploiting Laws of Large Numbers and Central Limit Theorems. For example, the local-to-unity asymptotics of Phillips (1987) studies auto-regressive models that are close to being nonstationary; the local-to-zero asymptotics of Staiger and Stock

same time we make the priors more dogmatic. This allows us to use the Law of Large Numbers and avoid issues pertaining to specific realizations while at the same time maintaining the relevance of the prior. Using this approach, we show that indeed small-dimensional models — which are possibly misspecified as they use fewer observables that then true DGP — prevail.

Our main results above follow from a simple intuition. Suppose Dr. A and Dr. B are both trying to predict y using a set of covariates $\{x_i\}_{i\in\{1,\dots,100\}}$. Dr. A believes that only x_1 matters—she assigns probability zero to the event that any other variable is related to y. Dr. B, instead, considers all 100 covariates. Suppose the true DGP is such that the best linear predictor of the outcome variable includes all variables: thus, Dr. B has a 'correct' model, while Dr. A does not. Lastly, normalize the priors so that, if no data is revealed, Drs. A and B have the same expected loss. After n data points are revealed, who is more confident?

Suppose first that n is small, e.g. $n = 5$. In this case, Dr. A will believe she has a good grasp of the data generating process—she is trying to fit only one parameter with 5 data points; her confidence will be high. Dr. B, instead, will make little headway in estimating her model. Fitting 100 parameters using 5 observations; her confidence will be low. Further, since the amount of data is "small," both agents' posterior estimates of σ^2_ϵ are close to their prior and therefore the competition is mainly over who believes they have a good grasp of the data generating process—i.e., Dr. A. Therefore, even if Dr. A has a misspecified model that omits 99 out of the 100 relevant variables, and even if the agents confidence without data is normalized to be the same, when n is small she will nevertheless have higher confidence in her predictive ability.

What happens then as data accumulates? A tradeoff emerges. While Dr. A will be able to estimate the parameters of her model well, she will also observe that it has a poor fit on the data. After all, she must attribute all the explanatory power of $x_2 \dots x_{100}$, which she does not consider in her model, to noise, therefore leading her to increase her estimate of σ^2_ϵ. Dr. B instead will take longer to estimate the parameters of her model, but she will be able to fit the data with a lower σ^2_ϵ. When n is small, the first effect dominates, and Dr. A will be more confident. As n grows, however,

(1997) studies Instrumental Variables models that are close to being unidentified; and Cattaneo et al. (2018) studies models where possibly many covariates are included for estimation and inference.

the second effect will acquire prominence, and Dr. B will become more confident.

This trade-off is the core of our results with small samples. A small number of observations increases confidence *faster* for agents with small-dimensional models. It is only as n grows larger that the confidence of agents with larger-dimensional models may catch up. As this may happen also when the true DGP is large-dimensional, when the dataset is relatively small agents with small-dimensional models are thus *overconfident* about their predictive abilities – and may thus be the most confident of all.

Relation to Statistics, Econometrics, and Machine Learning. A large literature studies model-selection techniques (see Claeskens and Hjort (2008) and Burnham and Anderson (2003) for textbook overviews). These include, for example, the C_p criterion of Mallows (1973), the Akaike Information Criterion (AIC) of Akaike (1974), and the Bayes Information Criterion (BIC) of Schwarz (1978). A key feature, common to all such techniques, is that they penalize large-dimensional models in small datasets. This is motivated normatively by the need to avoid over-fitting: large-dimensional models may be too flexible and give an illusion of fitting the data.

Our results have aspects reminiscent of these approaches. Small-dimensional models may prevail in small samples; in large samples, incorrect models are not selected, but larger models can continue to be selected with positive probability. As mentioned above, however, while results in the literature are justified normatively, our model selection criterion emerges *positively* from a framework in which different purely-Bayesian decision makers use different models, and the selected model is the one of the agent who is most confident in her own predictive ability. It is the competition between these agents—the selection of the most confident one—that generates the model selection.

One implication of our results is that in a competitive environment such as the auction we described, if we observe the use of smaller-dimensional models, it may not be possible to determine whether this is due to the use of model-selection techniques prescribed in Statistics or from the competitive selection between fully-Bayesian agents. To illustrate, it may be worth highlighting a parallel with the selection of entrepreneurs. In a context of heterogeneous priors, it is often observed that entrepreneurs hold more optimistic beliefs. The causality, however, may not be that

being an entrepreneur leads agents to become optimistic, but rather that agents whose priors are more optimistic are those that tend to become entrepreneurs. Similarly, in our context, small dimensional-models may be used because agents adopt them normatively; but our results also show that it could be that every agent is purely Bayesian, but it is the agents who have a smaller-dimensional model who are over-confident in their predictive ability, and thus acquire a prominent position – e.g., win the auction.

The remainder of the paper is organized as follows: Section 2 outlines the formal model and notation. Section 3 characterizes the expected posterior loss of a single agent, the foundation of our results. Section 4 collects our main results characterizing the winning model under competition: Section 4.1 for the case when n is large, and Section 4.2 for the case when n is small. Section 5 considers some extentions and implications of our results. Section 6 concludes and discusses the related literature in further detail. All proofs appear in the Appendix.

2 Model

A group of agents is competing to provide a forecast for a real-valued variable y as a function of k real-valued covariates $x \in \mathbb{R}^k$. In this section, we describe the relationship between y and x postulated by each of the agents, the data available, the agents competing, and the competition process itself.

Data Generating Process. A true Data Generating Process (DGP), denoted $\mathbb{P}$, determines the relationship between y and x. All agents assume there is a linear relation between the variable y and the covariates $x \in \mathbb{R}^k$, i.e.,

$$\begin{aligned} y &= x'\beta + \epsilon, \\ \text{where} \quad \epsilon|x &\sim \mathcal{N}_1(0, \sigma_\epsilon^2), \quad \beta \in \mathbb{R}^k. \end{aligned} \tag{1}$$

That is, agents believe that the DGP is a homoskedastic linear regression with Gaussian errors. For simplicity of exposition, we assume that the agents treat the distribution of the observables x as known, and denote it by P. We assume that under this distribution $\mathbb{E}_P[xx']$ is a full rank matrix. Let $\Theta := \mathbb{R}^k \times \mathbb{R}_+$, with $\theta = (\beta, \sigma_\epsilon^2)$ defin-

ing the unknown parameters of interest. As we discuss below, agent have (possibly different) priors π over Θ. Fixing P, $\theta = (\beta, \sigma_\epsilon^2)$ fully defines the DGP according to agents, denoted by Q_θ. We assume, for simplicity, that Q_θ has a probability density function $q(x, y|\theta)$, which holds whenever P has a probability density function.

Two comments are in order. First, about the linearity assumption: note that, because the covariates in x can be correlated, the linearity assumption is only mildly restrictive. For example, if one wished to express the non-linear DGP $y = 3\frac{x_1^3}{\sqrt{x_5}} + \epsilon$, one can simply define a new observable equal to $\frac{x_1^3}{\sqrt{x_5}}$. While not all non-linear DGPs can be expressed this way, good approximations can always be achieved. Thus, our framework allows the agents to have a wide family of non-linear relations as DGP.

Second, note that the assumptions above only concern the agent's perceived DGP, which is allowed to be misspecified: it may be that Q_θ differs from $\mathbb{P}$ at every θ – for example, errors may be heteroskedastic in the true DGP. We discuss the implications when they arise.

Data. Before making a prediction, each agent observes a dataset, denoted D_n, composed of n i.i.d. draws according to the true DGP $\mathbb{P}$. We denote the data as $D_n = (Y, X)$ where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times k}$. We assume that all agents observe the *same* data: this will be relevant for our application—as we shall see, in an auction setting this will avoid Winner's curse type concerns.

Actions and utility. Agents make a prediction of y given the covariates x, which formally means that they construct a prediction function f that maps x into y, i.e., $f : \mathbb{R}^k \to \mathbb{R}$. Their utility is maximized by minimizing a standard quadratic loss function, equal to the square of the difference between the true y and their forecast f, i.e., $-(y - f)^2$.

All agents are Bayesians, and thus choose a prediction function f that minimizes their Expected Loss given their Bayes-updated posterior beliefs. Define $L(f, \theta)$ as the agent's loss under prediction function f assuming the true DGP is Q_θ, i.e.

$$L(f, \theta) := \mathbb{E}_{Q_\theta}[-(y, f(x))^2]. \tag{2}$$

The loss captures the average quadratic error incurred in predicting y using $f(x)$,

assuming (x, y) are drawn randomly according to Q_θ. If π is the agent's prior over θ, and D_n the observed data, then the optimal action for the agent is to choose a prediction function $f^*_{(\pi,D_n)} : \mathbb{R}^k \to \mathbb{R}$ such that

$$f^*_{(\pi,D_n)} \in \operatorname*{argmin}_f \ \mathbb{E}_\pi[L(f, \theta)|D_n]. \tag{3}$$

For convenience, we denote by $L^*(\pi, D_n)$ the expected posterior loss of an agent who has prior π, observes data D_n, and uses the optimal predictor defined above, that is

$$L^*(\pi, D_n) := \mathbb{E}_\pi[L(f^*_{(\pi,D_n)}, \theta)|D_n] = \min_f \ \mathbb{E}_\pi[L(f, \theta)|D_n]. \tag{4}$$

2.1 "Models" and Competition

A key ingredient in our setting, as foreshadowed in the introduction, is that different agents may have with different priors over the unknown parameters in θ. Of particular interest will be the case in which these agents consider different subset of observables as relevant for their prediction – they have different "models" of the world.

If $\{1, 2 \ldots, k\}$ label the observables, for any i of them, if the agent's prior on β_i is degenerate at 0 it is easy to see that the agent is bound never to consider observable i in its prediction. Denote by $J(\pi)$ the set of observables that are instead considered by an agent with prior π. Formally, if π_i denotes the marginal over β_i of prior π and δ_0 is a Dirac measure at zero,

$$J(\pi) := \{i \in 1, \ldots, k : \pi_i(\beta_i) \neq \delta_0\}.$$

In what follows, we sometimes use simply $J \subset \{1, \ldots, n\}$ to denote a model – understood as the set of observables considered to make a prediction. Lastly, for a given vector β, denote by β_J the subvector consisting solely of the components in the set $J \subseteq \{1, \ldots, k\}$. Define x_J as the analogous subvector of x, and X_J as the corresponding submatrix of X.

Example: Normal-Inverse Gamma prior A convenient example is when the prior on $\beta|\sigma_\epsilon^2$ is normally distributed over the coordinates of the covariates that belong to a set J, and degenerate at zero otherwise, while the prior over σ_ϵ^2 is an inverse gamma distribution. This is the typical prior used for the Bayesian analysis of the Normal linear regression model.

Definition 1. *We say that the agent's prior π has Normal-Inverse Gamma form with* hyperparameters (γ, a_0, b_0) *if*

$$\beta_{J(\pi)}|\sigma_\epsilon^2 \sim \mathcal{N}_{|J(\pi)|}\left(0, \frac{\sigma_\epsilon^2}{\gamma|J(\pi)|}\mathbb{I}_{|J(\pi)|}\right) \qquad \sigma_\epsilon^2 \sim \textit{Inv-Gamma}(a_0, b_0).$$

In this special case, all agents differ on the subset of covariates $J(\pi)$ they consider and on the covariance of the slope coefficients.

Note also that if covariates have the same variance, the priors above are automatically normalized so that they all have the same expected loss before data, i.e., for all π, π', $L^*(\pi, \emptyset) = L^*(\pi', \emptyset)$.

The Competition Mechanism. As we discussed, agents compete through a mechanism that selects the agent with the *lowest posterior expected loss* given *her own prior*. Our analysis applies to any mechanism that leads to this selection. To give a concrete example, the following is a simple game in which the dominant strategy equilibrium results in this selection.

Consider a second-price auction, where, like in Atakan and Ekmekci (2014), the winner of the auction gets to choose an action that affects the value of the asset. Specifically, the action has a value that depends on her ability to predict a given variable, as in the examples given in the introduction. Formally, fixing the environment defined above (DGP, agents etc), consider a game with the following timing:

1. Nature draws $\theta \in \Theta$;
2. All agents see a common dataset D_n drawn according to Q_θ;
3. Agents submit bid in a sealed-bid second-price auction;

4. The winner observes x randomly drawn according to P and chooses an real-values action a;

5. The winner gets a lump sum payoff of $M-(y-a)^2$, where M is a large positive number.

Every bidder seeks to minimize the expected value $M-(y-a)^2$, leading to the expected loss function discussed above.

Because agents see a common data set, an agent with prior π has an expected value of $M - L^*(\pi, D_n)$ for winning. In the standard dominant equilibrium, the winning agent is the one with the highest value: since M is common across agents, the winner is thus the agent with the lowest expected loss (according to her own prior) given the observed data. Notice that since all agents observe the same dataset, and thus there is no asymmetric information – only heterogenous priors – no winner-curse-type consideration apply.

3 Characterizing the Posterior Expected Loss

We begin by characterizing (i) the optimal prediction function of a single Bayesian agent, and (ii) her expected posterior loss (henceforth, posterior loss) conditional on choosing the optimal prediction function using her own belief. The latter plays a crucial role in our environment.

3.1 Optimal Prediction

Characterizing the optimal prediction is a standard problem. The agent chooses f to minimize, $\mathbb{E}_\pi[L(f,\theta)|D_n]$, that can be rewritten as:

$$\mathbb{E}_\pi[\sigma_\epsilon^2|D_n] + \mathbb{E}_\pi\mathbb{E}_P[(x'\beta - f(x))^2|D_n]. \tag{5}$$

The first term does not depend on f. The second term involves the average error incurred in predicting $x'\beta$ using $f(x)$.[2] With standard arguments (i.e., exchanging

[2]The inner expectation averages over values of x. The outer expectation averages over the values of β.

the order of integration and taking first order conditions), we can see that the inner expectation of the second term is minimized by the function:

$$f^*_{(\pi,D_n)}(x) := x' \ \mathbb{E}_\pi[\beta|D_n] = x'_{J(\pi)}\mathbb{E}_\pi[\beta_{J(\pi)}|D_n]. \tag{6}$$

Thus, a Bayesian decision maker with a posterior $\pi|D_n$, model $J(\pi)$, and a square loss function, forecasts y at x as her Bayesian posterior mean of $x'\beta$. Again, this is a standard result.

3.2 Posterior Loss

We now turn to characterizing the agent's posterior loss computed using her own belief and conditional on her adopting an optimal forecast. This measures how confident each agent is of her predictive ability, and it will be the central driver of the dynamic of our competition between agents. Most importantly, the key driving forces of our results will already be evident from this simple analysis.

The following Lemma shows that the agent's posterior loss $L^*(\pi, D_n)$ can be decomposed into the sum of two parts: one that we interpret as *model fit*, i.e., how well is the agent's model fitting existing data; and one that we interpret this as the agent's *model's estimation uncertainty* according to her own prior.

Lemma 1. *The agent's posterior expected loss from her Bayes predictor is:*

$$L^*(\pi, D_n) = \mathbb{E}_\pi\left[\sigma^2_\epsilon|D_n\right] + \mathrm{Tr}\left(\mathbb{V}_\pi[\beta_J|D_n]\ \mathbb{E}_P[x_J x'_J]\right), \tag{7}$$

where $\mathbb{V}(\cdot)$ *is the variance-covariance operator,* Tr *is the trace operator, and* J *denotes the agent's model* $J(\pi)$*.*

The Lemma above shows that the agent's expected posterior loss $L^*(\pi, D_n)$ can be characterized as made of two components. The first is standard: the posterior expectation of the variance of the error— the agent's estimate of the irreducible noise in the system, in turns related to *model fit*, i.e., how well is the agent's model fitting existing data, because the agent must ascribe all unexplained variation to noise.

The second term, $\mathrm{Tr}\left(\mathbb{V}_\pi[\beta|D_n]\ \mathbb{E}_P[xx']\right)$, is the trace of the variance-covariance matrix of the coefficients of the model (adjusted by $\mathbb{E}_P[xx']$). This is a measure of

how uncertain is the agent how her estimation of her model— thus capturing the to *model estimation uncertainty* faced by the agent according to her own prior. For an intuition consider the simpler case in which observables are independent and have the same variance (i.e., orthornormal). In this case, the second term reduces to $\mathrm{Tr}\left(\mathbb{V}_\pi[\beta|D_n]\right)$, i.e., $\sum_{i=1}^k \mathbb{V}_\pi[\beta_i|D_n]$. By comparison, directly evaluating the loss of the Bayes estimator, i.e., substituting (6) in (5), the second term in (5) equals

$$\begin{aligned}
&\mathbb{E}_P\mathbb{E}_\pi[(x'\beta - f^*_{(\pi,D_n)}(x))^2|D_n] \\
=&\mathbb{E}_P\mathbb{E}_\pi[(x'\beta - x'\mathbb{E}_\pi[\beta|D_n])^2|D_n] \\
=&\mathbb{E}_P\mathbb{E}_\pi[(x'(\beta - \mathbb{E}_\pi[\beta|D_n])^2|D_n] \\
=&\sum_{i=1}^k \mathbb{V}_\pi[\beta_i|D_n]
\end{aligned}$$

where the last equality follows from the definition of variance and the assumption that the x's are orthonormal. Thus, the second part of the loss function is in this case simply the sum of the variances of the parameters β, indeed a measure of model estimation uncertainty. The exact formula in (7) extends this to cover the case of observables with a general variance-covariance matrix.

4 Competing Models

Lemma 1 helps us understand the loss a single agent expects given her posterior. We now apply this Lemma to understand the model of the 'winning agent,' i.e., the agent with the lowest expected posterior loss among a collection of agents.

As we foreshadowed, we identify the winner both in the case that the dataset is large (i.e. the number of observations n is "large" relative to the number of observables k), and the case that it is small. We will handle each in turn. We show that when the dataset is big, the true (or larger) model prevail. Our results apply to general priors (modulo some technical assumptions to ensure posteriors are well behaved enough) and general data generating processes. Conversely, when the dataset is small, smaller models may take a lead. Our small-sample results below are for the specific case when all agents have priors of the Normal-Inverse Gamma form introduced earlier. For this reason, we begin our discussion with the large-data analysis.

The main building block of our results is Lemma 1, which contains the key intuition. When n is large, the model estimation uncertainty component of the posterior loss vanishes: each agent, even those with a 'wrong' model, will reduce the uncertainty about the parameters to zero. All that matters is the model fit. Then, it is easy to see that agents that use models that exclude relevant variables are bound to have a higher expected loss, since they must estimate a higher σ_ϵ^2 to account for the variation that they are disregarding. Put differently: with large data, agents whose models are misspecified by excluding relevant variables will *not* win our competition. Whether agents who consider more variables may win is a separate question, as in this case the model fit achieved by two competing models will be the same. In what may be less intuitive, we show that agents' prior continue to affect the model competition even with infinite data.

When n is small, agents with small-dimensional models have instead an advantage, because they are going to have smaller model estimation uncertainty, the second part of the expected loss as characterized in Lemma 1. Even though all agents start with the same expected loss with no data, when the data revealed is (relatively) small, the expected loss decreases *faster* for agents with small-dimensional model. Thus, agents who hold models that are misspecified in that they exclude relevant variable may end up being more confident in their predictive ability. The example discussed in the introduction (of Dr. A and B) may provide further intuition. To recap, *ceteris paribus*, trying to estimate more parameters from the same amount data will result in more model uncertainty, i.e., less concentrated posteriors. This uncertainty will therefore be reflected in the agent's expected loss.

Before we dive into the results, let us introduce them with simulation evidence. Figure 1 shows simulation results in a setting where there are six observables in the dataset, $\{x_1, \ldots, x_6\}$, of which only the first five are relevant for prediction. We suppose we have 63 agents, one for each non-empty subset of $\{x_1, \ldots, x_6\}$, all with Normal-Inverse Gamma priors with the same shared hyperparameters. We simulate datasets of sizes $n = 1$ to 50, and plot the frequency of the size of the model of the agent with the lowest subjective expected loss. Two main features emerge. First, when n is "small" the winner tends to have a small model, indeed a model that we know to be misspecified (since we chose the DGP to depend on observables x_1 through x_5). Secondly, as n grows large, the true model wins more often. However,

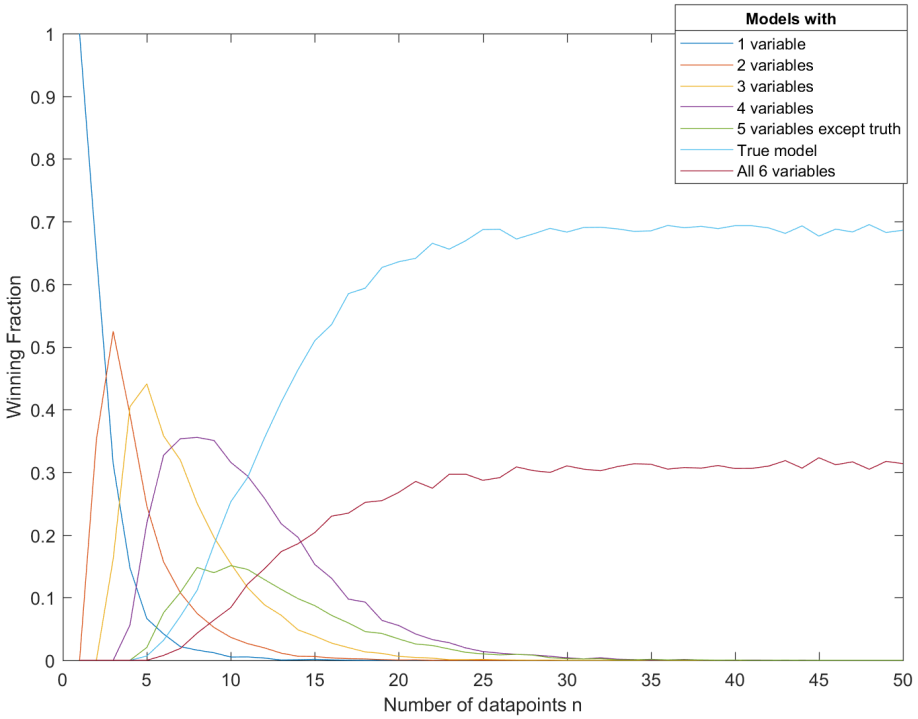

Figure 1: Winning rates for different models with Normal Inverse-Gamma priors and shared hyperparameters $(a_0, b_0, \gamma) = (2, 1, 0.001)$ on $5,000$ simulated datasets of size $n = 1$ to 50. 6 Covariates are distributed $x \sim N(0, \mathbb{I}_6)$. True d.g.p only depends on covariates 1–5, $(\beta_1 \dots \beta_5) \sim N(0, \mathbb{I}_5)$, $\beta_6 = 0$.

also the larger model, that includes the redundant variable x_6, continues to win, with relative frequencies that appear to converge to a steady state. In what follows, we give analytical foundations to each of these observations.

Finally, a little more notation will be useful. Note that a dataset D_n induces an order $\succ_{D_n}$ over priors according to the posterior loss they induce given D_n.

Definition 2. *Fixing a dataset D_n we define the order $\succ_{D_n}$ over priors as:*

$$\pi \succ_{D_n} \pi' \quad \iff \quad L^*(\pi, D_n) < L^*(\pi', D_n). \tag{8}$$

Definition 3. *Given a vector $\beta_0 \in \mathbb{R}^k$, J_0 is the set of indices of the coordinates of β_0 that are nonzero; i.e.,*

$$J_0 := \{\kappa | \beta_\kappa \neq 0\}.$$

Then define:

$$\begin{aligned}\mathcal{J}_0^S &:= \{J \in \mathcal{J} \mid J \subset J_0\},\\ \mathcal{J}_0^L &:= \{J \in \mathcal{J} \mid J_0 \subset J\},\\ \mathcal{J}_0^W &= \{J \in \mathcal{J} \mid J_0 \not\subset J\}.\end{aligned}$$

In words, if $\mathcal{J}_0$ the set of indexes useful to in the prediction, $\mathcal{J}_0^S$ are the strictly smaller set of indexes nested in J_0; $\mathcal{J}_0^L$ are the strictly larger ones that nest J_0; and $\mathcal{J}_0^W$ are "wrong" ones, i.e., those that rule out at least one explanatory variable that is non-zero in J_0. Note that $\mathcal{J}_0^S \subset \mathcal{J}_0^W$.

We also define the set of priors that give non-zero weight to indexes in $\mathcal{J}_0$,

$$\Pi_0 := \{\pi | J(\pi) = J_0\}.$$

The sets Π_0^S Π_0^L, Π_0^W are defined analogously

4.1 The winner with n 'large'

We characterize the winner for large n under the mildly technical regularity assumptions about the priors of the agents and a set of 'standard' high-level conditions on the true DGP, $\mathbb{P}$.

Assumption 1. *Each agent has a prior over θ characterized by a probability density function $\pi(\cdot)$ that is six times continuously differentiable and with full support over the set $(\beta_{J(\pi)}{}', \sigma_\epsilon^2)' \in \mathbb{R}^{|J(\pi)|} \times \mathbb{R}_+$.*[3]

Assumption 1 posits that agents' priors over the β_i's are either degenerate at 0, or full support. In the latter case, it requires a pdf to exist and to be suitably differentiable. (This naturally holds for Normal-Inverse Gamma.)

We now turn to assumptions on the true DGP $\mathbb{P}$. Before we do, recall that while the analysis conducted by each of our Bayesian agents is based on a linear regression model with Normal and homoskedastic errors (Eq. (1)), in the asymptotic results below we allow for the possibility of their likelihoods being misspecified. For example,

[3]By definition, the prior of agent j for any β_κ, $\kappa \notin J$, is degenerate at 0.

true errors may be heteroskedastic or non-Normal. We now impose assumptions on the true DGP $\mathbb{P}$.

Assumption 2. *Let $\mathbb{P}$ denote the joint distribution of (x, y). Let the data $D_n := ((x_1, y_1), \ldots, (x_n, y_n))$ denote an i.i.d. sample from $\mathbb{P}$. Then:*

1. *(Population Second Moments) The smallest eigenvalue of the matrix $\mathbb{E}_{\mathbb{P}}[xx']$ is strictly positive, and its largest eigenvalue is finite.*

2. *(Central Limit Theorem for covariates and residuals) Let β_0 denote the parameter that satisfies $\mathbb{E}_{\mathbb{P}}[x(y - x'\beta_0)] = 0$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i(y_i - x_i'\beta_0) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathbb{E}_{\mathbb{P}}[(y - x'\beta_0)^2 xx'])$$

3. *(Asymptotic behaviour of posterior variances for misspecified models) Let $g(x, y|\theta)$ denote the probability density function of the possibly misspecified parametric model for the distribution of (x, y) used by the agents. Let $K(\theta) := -\mathbb{E}_{\mathbb{P}}[\ln g(x, y|\theta)]$ and let $D^2K(\theta)$ denote the Hessian of $K(\theta)$. Then*

$$nV(\theta|D_n) \xrightarrow{p} \left(D^2K(\theta_0)\right)^{-1},$$

where θ_0 is the parameter that minimizes $K(\theta)$.

Let us briefly discuss the content of the assumption above. For the non-technical reader, it may suffice to note that this assumption is satisfied when the true DGP is well-behaved 'enough' that long run estimation/inference of this true DGP is possible for our misspecified Bayesian agents —- indeed conditions satisfied by most commonly used examples. Part (1) guarantees that the matrix of population second moments is both finite and invertible. This implies there is a unique parameter β_0 satisfying $\mathbb{E}_{\mathbb{P}}[x(y - x'\beta_0)] = 0$ and we interpret it as the true parameter.[4] Part (2) is a standard Central Limit Theorem, often invoked to obtain the asymptotic distribution of the

[4]This also implies that the population second moments can be consistently estimated from the sample second moments of the data. We will use this assumption to characterize the probability limit of the Maximum Likelihood Estimators based on the possibly misspecified likelihoods of the Bayesian agents. Note that in principle, we allow for the distribution of covariates assumed by the competing agents (denoted P) to be different from the distribution of covariates under $\mathbb{P}$.

Ordinary Least Squares (OLS) estimator in a linear regression model.[5] Part (3) can be thought of as imposing a particular aspect of the large sample behavior of posterior distributions captured by the Bernstein-Von Mises Theorem (BVMT).[6]

We are now ready to state our large-sample results.

Theorem 1. *Let $\mathbb{P}$ denote the true data generating process and let the data $D_n := ((x_1, y_1), \ldots, (x_n, y_n))$ denote an i.i.d. sample from $\mathbb{P}$. Define β_0 as the parameter such that $\mathbb{E}_{\mathbb{P}}[x(y - x'\beta_0)] = 0$ and let $\sigma_0^2 = \mathbb{E}_{\mathbb{P}}[(y - x'\beta_0)^2]$ and π_0 any element of Π_0. Let J_0 denote the associated true model for β_0. Then, for any priors satisfying Assumptions 1 and any $\mathbb{P}$ satisfying Assumption 2:*

i. If $\pi \in \Pi_0^L$: $\mathbb{P}[\pi \succ_{D_n} \pi_0] \to c(\pi, \pi_0, \beta_0, \sigma_0^2) \in (0, 1]$.

ii. If $\pi \in \Pi_0^W \cup \Pi_0^S$: $\mathbb{P}[\pi \succ_{D_n} \pi_0] \to 0$.

Theorem 1 gives two main takeaways. The first item tells us that a model which is larger than the true model, i.e., contains additional observables that are irrelevant for prediction, continues to win against the true model with a probability that is bounded away from zero, even with infinite data. The second item tells us that a wrong model, i.e., a model which rules out an observable that is relevant for prediction will eventually lose to the true model.

The second result is intuitive. By Lemma 1, we can decompose posterior loss into two terms: expected variance of the noise and model estimation uncertainty. The latter converges to zero for all agents (guaranteed by Assumption 2. The first term is the posterior expectation of the noise term. The former term will instead differ:

[5] We will use this assumption to characterize the asymptotic distribution of the difference in model fit for models that are larger than the true model. This assumption allows for conditional heteroskedasticity of regression residuals.

[6] If we assume that the agents' DGP, $g(y, x|\theta)$, is a correctly specified parametric statistical model, the BVMT implies that the posterior distribution of a parameter θ is approximately Normal, centered at the maximum likelihood estimator, and covariance matrix equal to

$$\left(D^2 K(\theta_0)\right)^{-1} / n, \tag{9}$$

where θ_0 denotes the true parameter generating the data. A similar result is available for misspecified models; see Bunke et al. (1998) and Kleijn et al. (2012). Instead of imposing the BVMT theorem for misspecified models as a high-level assumption (as, for example, Condition 1 in Müller (2013)) we only assume that the variance of the posterior of θ in a possibly misspecified model is approximately given by (9). In this case, θ_0 refers to the parameter that minimizes the Kullback-Leibler divergence between $\mathbb{P}$ and $g(x, y|\theta)$.

agents who rule out an observable that is relevant for prediction must attribute its explanatory power to noise. So, as n grows large, their posterior expectation of the variance of the noise term will be necessarily larger than that of an agent with the true model. Thus, he will always have a lower confidence.

But what about agents whose model is larger than the true model? This part, covered by part (i) of the theorem, is slightly more subtle. After all, this agent will also eventually learn the true data generating process: that is, her beliefs about the βs associated to redundant observables must converge to zero. But how will the confidence compare? For any fixed n, the agent with more observables in her model will have a less concentrated posterior on β. On the other hand, she will also have slightly smaller posterior expectation of the variance of the noise term: she will mistakenly attribute some explanatory power to these superfluous observables. Which of these two effects dominate, both of which can be shown to be $O_p(\frac{1}{n})$, determines the likelihood of winning. Part (i) of Theorem 1 says that the probability of the larger model winning is bounded away from zero, even in the limit; at the same time, this probability need not necessarily converge to being identical to that of the correctly specified model.

In fact, we show that as the sample size grows large, the comparison between model fit and model uncertainty behaves as the probability of some positive random variable (coming from the difference between the estimated variances of the smaller and larger model) exceeding some constant (coming from the difference in model uncertainty). The following corollary gives a concrete characterization of this relation, assuming conditional homoskedasticity assumption and that agents have the correct specification of the distribution of covariates.

Corollary 1. *Let $\theta_0 := (\beta_0, \sigma_0^2)$ where β_0 and σ_0^2 are defined as in Theorem 1. Suppose that the data is conditionally homoskedastic; that is $\mathbb{E}[(y - x'\beta_0)^2 xx'] = \sigma_0^2 \mathbb{E}_{\mathbb{P}}[xx']$. Suppose also that the distribution P assumed by the agents is correctly specified. Then, under the assumptions of Theorem 1:*

$$c(\pi, \pi_0, \beta_0, \sigma^2) = P(\chi^2_{|J(\pi)-J(\pi_0)|} > 2(\eta_{\pi_0}(\theta_0(\pi_0)) - \eta_\pi(\theta_0(\pi))),$$

where $\eta_\pi(\theta)$ denotes the elasticity of the prior π with respect to σ^2 at θ.

A possibly less intuitive implication of Corollary 1 is that a large date set does

not completely 'wash out' the priors. Indeed, one may argue that aspects of priors beside having the right states in the support typically should matter for Bayesian agents with infinite datasets. But this is not the case here: we show that even in large samples, the specific priors π and π_0 affect the competition. Interestingly, our result is very concrete about the feature of the prior that matters: the elasticity of the prior density with respect to the variance parameter. This, along with the model's dimension, is the key aspect that affects the probability that a large model defeats the correct one.

To give a more concrete sense of Corollary 1, consider the example of Normal-Inverse Gamma prior. In this case, the elasticity of the prior density with respect to the variance can be shown to equal:

$$\eta_\pi(\theta) = \frac{|J(\pi)|}{2}\left(\frac{\gamma\beta'_{J(\pi)}\beta_{J(\pi)}}{\sigma^2_\epsilon} - 1\right) - (a_0 + 1) + \frac{b_0}{\sigma^2_\epsilon}.$$

Two implications follow. First, consider the competition between two agents π_L and π_0, both with Normal-Inverse Gamma priors with the same parameters (a_0, b_0) and a diffuse prior on β ($\gamma = 0$). In large samples the probability of $\pi_L \succ_{D_n} \pi_0$ becomes

$$P(\chi^2_{|J(\pi_L)|-|J(\pi_0)|} > |J(\pi_L)| - |J(\pi_0)|).$$

This function is increasing in $|J(\pi_L)| - |J(\pi_0)|$ and asymptotes to 50%, meaning that a larger model can defeat the true model at most half of the time. This is intuitive as the models become identical.

However, consider now the competition between the same Normal-Inverse Gamma agents, but allow them to have different parameters (a_π, b_π). Algebra shows that if b_π is large enough (meaning that the variance of the prior over σ^2_ϵ is large), then the probability that the larger model defeats a smaller model can become arbitrarily close to 1. That is: the larger model 'beats' the correct one even in the limit, with a probability that can be made close to 1.

4.2 The winner with n "small"

We are now ready to discuss the properties of the winner model when the number of observations n is relatively small. We have already seen in Section 3.2 how in this case there are advantages given to be smaller-dimensional models, and the winner may indeed be a model smaller-dimensional than the true DGP. We will now provide additional formal results to strengthen this understanding, aiming to characterize when this is the case.

In this subsection, we assume that agents have Normal-Inverse Gamma priors for tractability.[7] We also assume that all agents' priors share the same hyperparameters: as we discussed in the introduction, this ensures that all agents have the same prior expected loss before data. Differences in posterior expected loss arise only from the fact that the posterior evolves differently for models of different sizes given the same model. Lastly, in some cases we will also assume that covariates are i.i.d., i.e., $\mathbb{E}_P[x'x] = \mathbb{I}_k$.

The winner with 1 data point. We start with an extreme but stark result for the case in which agents observe only one datapoint.

Proposition 1. *Suppose all agents have Normal-Inverse Gamma priors with shared hyper-parameters (a_0, b_0, γ) and that $\mathbb{E}_P[x'x] = \mathbb{I}_k$. If the dataset consists of a single observation, i.e. $n = 1$, then the winner is always some agent with a single variable model, i.e., an agent with a prior π s.t. $|J(\pi)| = 1$.*

Note that this result holds independently of the true DGP: even when that is high-dimensional, with only one point it is always a 1-dimensional model to win. Numerical simulations suggest that a generalization of this result appears to hold: with n observations the winner is n-dimensional or smaller. We were not able to

[7] As these are a conjugate priors for the Normal linear regression model, posteriors have simple analytical forms. Algebra shows that

$$\mathbb{E}_\pi[\sigma^2_\epsilon|D_n] = \frac{\frac{2b_0}{n} + \frac{1}{n}\min_{\beta\in\mathbb{R}^{|J(\pi)|}}(y - X_{J(\pi)}\beta)'(y - X_{J(\pi)}\beta) + (\gamma|J(\pi)|)\,||\beta||^2}{\frac{2a_0}{n} + 1 - \frac{2}{n}}, \quad (10)$$

$$\mathbb{V}_\pi[\beta_{J(\pi)}|D_n] = \mathbb{E}_\pi\left[\sigma^2_\epsilon|D_n\right](X'_{J(\pi)}X_{J(\pi)} + (\gamma|J(\pi)|)\mathbb{I}_{|J(\pi)|})^{-1}. \quad (11)$$

formalize such an observation, as in finite samples the expressions for $\mathbb{E}_\pi[\sigma_\epsilon^2|D_n]$ and $\text{Tr}(\mathbb{V}_\pi[\beta|D_n]))$ are algebraically less tractable: the reason is that they depend on the inverse of a matrix of specific data realizations, which is hard to operate with.

A novel approach for small n analysis. We now suggest a novel way of approaching the problem of small sample analysis that allows us to obtain further results despite the analytical limitations discussed above. This general approach may be of interest independently from the specifics of our problem.

The initial observation is that small samples appear to be distinct from large ones for two basic properties: *i*) that the prior remains relevant instead of being partially 'washed away' by the data; and *ii*) that specific data realizations matter, instead of only the population average mattering. It is the latter characteristic that leads to the analytical difficulties we encountered above. In large samples these issues do not arise because laws of large numbers can be invoked, circumventing the analytical concerns as they allow us to replace specific observation with population averages.

But what if we find a way to maintain the first property of small samples—that the prior still matters—while dispensing with the second, problematic one—that specific realizations matter? To do this, we let n grow to infinity, thus allowing us to use the law of large number, but at the same time vary the hyperparameters of priors to simultaneously make them become more and precise, at a pace such that they maintain their relevance. Such 'alternative asymptotics' framework, has been used to study different inference problems in econometrics.[8]

The next result uses this approach to show that as long as the prior remains relevant, smaller models have an advantage.

Theorem 2. *Suppose all the agents have Normal-Inverse Gamma prior with shared hyper-parameters (a_o, b_n, γ), where $b_n \in O\left(n^{2+\upsilon}\right)$, for some $\upsilon > 0$. Let $\mathbb{P}$ denote the true data generating process and let the data $D_n := ((x_1, y_1), \ldots, (x_n, y_n))$ denote an i.i.d. sample from $\mathbb{P}$. Define β_0 as the parameter such that $\mathbb{E}_\mathbb{P}[x(y - x'\beta_0)] = 0$ and let $\sigma_0^2 = \mathbb{E}_\mathbb{P}[(y - x'\beta_0)^2]$. Let J_0 denote the associated true model for β_0. If P is correctly specified, then for any $\mathbb{P}$ satisfying Assumption 2:*

[8] See the local-to-zero asymptotics of Staiger and Stock (1997) for the analysis of instrumental variable regression with a weak instrument, the local-to-unity framework of Phillips (1987) for the analysis of inference in a autoregressive model with autocorrelation close to 1.

i. If $\pi \in \Pi_0^L$, $\pi_0 \in \Pi_0$*:* $\mathbb{P}[\pi \succ_{D_n} \pi_0] \to 0$ *as* $n \to \infty$.

ii. If $\pi \in \Pi_0^S$, $\pi_0 \in \Pi_0$*:* $\mathbb{P}[\pi \succ_{D_n} \pi_0] \to 1$ *as* $n \to \infty$.

Proof. See Section A.4 of the main Appendix. □

In words, this result shows that if the prior concentrates fast enough, the results are the *converse* of the large data case (i.e., Proposition 1): models that are larger-dimensional than the true DGP never win, and instead the winner is always smaller-dimensional than the truth.

5 Extensions and Implications

We conclude our formal analysis with a discussion of two variants of our model, both of which provide the same stark prediction of the "small data" world: agents with "simple" models always win. Indeed, both of these strengthen our small data results.

Known Variance. What happens when the variance σ^2 of the noise-term ϵ is commonly known among the agents? This is an extreme special case of our analysis above; it may be realistic in some situations, but not in others.[9]

Proposition 2. *Suppose agents have Normal priors with shared hyper-parameter* γ *. Fix a prior* π *with* $|J(\pi)| = k$*. For any* $k' < k$*, and any dataset* D_n *for* $n > 0$*, there exists a prior* π' *such that* $J(\pi') \subseteq J(\pi)$ *with* $|J(\pi')| = k'$ *and such that* $\pi' \succ_{D_n} \pi$*.*

In short, for any model $J(\pi)$, and *any* dataset of *any* size, some smaller model with a subset of the explanatory variables will have a lower posterior loss.

[9]Indeed, there is an aspect that makes this assumption problematic in some environments. When variance is not uncertain, agents with incorrect models of the world will, as data accrues, observe that their model has an empirical error higher than the (known) σ^2, because the model disregards some observables relevant for prediction. For n large, this disparity in the empirical error and the (known) σ^2 should lead them to question their underlying model. However, as is standard with Bayesians with dogmatic beliefs (here they have degenerate beliefs on σ^2) they do not. When the dataset is not too large, however, such issues will not arise.

Bidding Before seeing the data. A different extension is to a setting where agents know that they will see exactly n data points, but expected confidence *before* they view the data: this is the case, for example, when agents have to submit a bid before seeing the data, but know that they will see n data points before making their prediction. Put differently, we study the expectation before seeing the data of the expected loss after n datapoints.[10] This situation may be not unusual in reality, as often new data is revealed after bidding but before predictions needs to be made.

A stark result holds in this case: smaller models always win. In fact, in this case the result is even stronger than previous ones, as we explain below.

Proposition 3. *Suppose agents have Normal Inverse-Gamma priors with shared hyper-parameters* (a_0, b_0, γ)*, and that* $\gamma = 0$*. Suppose further that* $x \sim \mathcal{N}_k(0, \mathbb{I}_k)$ *independently of* ϵ*. Fix a prior* π*. For any prior* π'*, such that* $|J(\pi')| < |J(\pi)|$*, we have that*

$$\mathbb{E}_{m(\pi')}[L^*(\pi', D_n)] < \mathbb{E}_{m(\pi)}[L^*(\pi, D_n)],$$

whenever $n > |J(\pi)| + 1$*. Here the outer expectation is taken over the agents' 'marginal' distribution of the data* $m(\pi) := \int q_\theta(D_n)\pi(\theta)d\theta$*.*[11]

Proposition 3 shows that when confidence is computed before data is realized, not only smaller models 'beat' the correctly specified one, but this holds for *any* smaller model, not just *some* of the smaller models, as was the case in some of the previous results; moreover, this holds for any size of the dataset n.

For an intuition, consider again the decomposition of posterior loss obtained through Lemma 1,

$$L^*(\pi, D_n) = \mathbb{E}_\pi\left[\sigma_\epsilon^2 | D_n\right] + \mathrm{Tr}\left(\mathbb{V}_\pi[\beta|D_n]\,\mathbb{E}_P[xx']\right).$$

Depending on the realized D_n, the first term, model fit, can be larger or smaller than the prior expectation of it before data is realized. Indeed, this is the complicating factor in the analyses of Propositions 1 and 2. However, in the case of Proposition 3, we take expectation over all possible datasets, and the first term reduces to its prior

[10] Note that since different agents have different beliefs about the data generating process, they take expectations with respect to different probability distributions over the space of datasets D_n.

[11] Hence, the expression $\mathbb{E}_{m(\pi')}[L^*(\pi', D_n)]$ is the *Bayes risk* of the *Bayes Predictor*. See Equation 1.14 in Chapter 1.6 in Ferguson (1967).

expectation. So we can focus only on the second term, model estimation uncertainty. But then, for reasons analogous to the previous propositions, the residual model uncertainty is smaller in expectation for smaller models. Proposition 3 follows.

Connection with the Akaike information criterion. A different way to understand our results is to relate the model section induced by competing models to the Akaike Information Criterion, a well-studied model selection criterion in Econometrics and Statistics. In what follows, we illustrate that the loss function of an agent with Normal-Inverse Gamma prior is "close" to the AIC for the linear regression model.

Definition 4 (Akaike Information Criterion)**.** *Given a dataset $D_n = (y, X)$ with n data points and k possible covariates, the Akaike information criterion for linear regression evaluates a model based on X_J as:*

$$L_{Akaike}(J, n, D_n) = \ln \widehat{\sigma}^2(J, n, D_n) + \frac{2|J|}{n},$$

where

$$\widehat{\sigma}^2(J, n, D_n) = \frac{1}{n} \min_{\beta \in \mathbb{R}^{|J|}} (y - X_J\beta)'(y - X_J\beta).$$

In words, consider a model J with $|J|$ observables. The expression $\widehat{\sigma}^2(J, n, D_n)$ is the OLS estimator of the residual variance based on a model with covariates X_J in the dataset D_n with n observations. As is well understood, selecting a model with a lower estimated variance may not favor the model with the best out of sample performance. This is because selecting based on average residuals favors models that have more covariates (i.e., regressions which "overfit" the data). The Akaike Information Criterion (AIC) compensates for this by adding a penalty term equal to $\frac{2|J|}{n}$, i.e., twice the ratio of the number of covariates in the model and the number of data points. Algebra shows that if agents have an uninformative Normal-Inverse Gamma prior ($\gamma = 0$), then the posterior loss is approximately equal to

$$\ln\left(\widehat{\sigma}^2(J, n, D_n)\right) + \ln\left(1 + \frac{1}{n}\mathrm{Tr}\left(\left(\frac{X_J'X_J}{n}\right)^{-1} \mathbb{E}_p[x_J x_J']\right)\right).$$

Thus, if the sample size is large and the agents' distribution of covariates is well-

specified, the posterior loss of an agent with prior π will be approximately equal to the Akaike Information criterion (with a penalty of $|J|/n$ instead of $2|J|/n$).

The prevalence of larger models in the model competition can then associated to the 'conservativeness' of the Akaike Criterion for model selection. Our Theorem 1, however, makes it clear that the relation is only qualitative: larger models will indeed prevail in large samples, but the probability of a larger model being selected will continue to be affected by the prior.

Finally, it is worth reiterating that the foundations of the AIC are normative: the criterion was proposed as a way to select a model that avoids overfitting. Conversely, our analyis provides a *positive* foundation for the AIC: we study the outcomes when Bayesian agents compete in a way that selects the agent with the lowest posterior expected loss.

6 Discussion and Conclusion

We analyze a novel model of competition between agents. A variable of interest is related to a vector of covariates. Agents have different models of these relationship: in particular they rule in/ rule out different $x's$ as being potentially related to prediction. All agents observe a common dataset of size n, drawn from the true data generating process. The winner is the agent with the lowest expected loss, expectations taken with respect to their own subjective posterior. This winner corresponds to the winner under a stylized auction model we formally define and analyze, but may also be of interest more generally in situations where subjective confidence in predictions lead to selection. We study the relationship between the true data generating process and the model of the winner, and how this relationship changes as a size of the available dataset, n. We show two stark results.

Firstly, when n is large, the winner is qualitatively similar to the model with the lowest value of the Akaike Information Criterion. Misspecified models (i.e., models that rule out an observable which is relevant for prediction) never win, but overly large models may continue to win even as data grows unboundedly large. The prior is not completely 'washed out' by the large sample. The elasticity of the prior density with respect to the variance parameter continues to affect the model competition even

with infinite data. This result is established for a very general class of priors and true data generating processes.

Secondly, when n is small, we show that 'simple' models, i.e. models that employ few observables, take the lead, even if the true data generating process is rich. To establish this result, we used a 'drifting' Normal-Inverse Wishart prior; where we allowed the elasticity of the prior density with respect to the variance to increase with the sample size.

There are several natural avenues to future research. An obvious one is a setting in which agents each observe a private dataset: this complicates our analysis because now a notion of the winner's curse applies. Each agent must consider whether they are beating the others because their model is truly performing well on the data, or because their dataset is non-representative. Another one is to consider dynamic variants: if agents got feedback or could invest to acquire more data, what kinds of models would be selected?

6.1 Related Literature

As we mentioned in the Introduction, there is a large body of literature in statistics and econometrics that studies model selection methods and provides normative foundations. That literature is too vast to comprehensively cite here, we refer the reader to a textbook account in Burnham and Anderson (2003). Our large data results are closely connected to the Akaike Information Criterion (AIC) introduced in Akaike (1974). Asymptotic properties of the AIC were studied in the seminal paper of Nishii (1984), which our results closely track.

In terms of the connections to the literature in economic theory, since one natural application of our model is an auction, our results are related to Atakan and Ekmekci (2014), who consider the competitive sale of assets whose value depends on how they are utilized.[12] The successful bidder chooses an action that determines, together with the state of the world, the payoff generated by the asset. They focus on a setting where bidders have a common prior but observe private signals. Their main result is the possibility of (complete) failure of information aggregation. Our results are similar in that in our applications as well the value of the object depends on an action taken

[12] Bond and Eraslan (2010) study a trading environment with a similar feature.

by the agent. However, our paper considers a complementary environment where all bidders observe the same information but they have different priors. Information aggregation is ruled out by assumption, and our key theme is model selection.

We assume that agents have different priors and are fully aware they have different priors: that is to say our agents agree to disagree. This assumption has been used in economic theory at least since Harrison and Kreps (1978). We refer the reader to Morris (1995) for a discussion of the common and heterogeneous prior traditions in economic theory. Heterogenous priors have been used in a number of applications in bargaining (Yildiz, 2003), trade (Morris, 1994), financial markets (Scheinkman and Xiong, 2003; Ottaviani and Sørensen, 2015) and more.

We limit ourselves in the rest of this section to connections with the growing literature in economic theory that attempts to understand outcomes in economic settings when agents have misspecified models.

A large literature has studied models of model misspecification in individual decision-making, with famous examples like overconfidence and correlation neglect. A few recent theoretical contributions to this enormous literature include Heidhues et al. (2018) and Ortoleva and Snowberg (2015), to which we refer for further references.

A novel approach to modeling misspecification in economic theory is the directed acyclic graph approach; see Pearl (2009). This is exploited in a single person decision framework in Spiegler (2016), which studies a single decision maker with a misspecified causal model and large amounts of data. The paper shows that the decision maker may evaluate actions differently than their long-run frequencies, and exhibit artifacts such as "reverse causation" and coarse decision making. This approach is then used in Eliaz and Spiegler (2018), which proposes a model of competing narratives. A narrative is a causal model that maps actions into consequences, including other random, unrelated variables. An equilibrium notion is defined, and the paper studies the distribution of narratives that obtains in equilibrium.

In strategic settings, Esponda and Pouzo (2016) defines a learning-based solution concept ('Berk-Nash Equilibrium') for games in which agents' beliefs are misspecified. More broadly, solution concepts have been posited for settings where agents suffer from some sort of misspecification, including well-known examples like analogy-based equilibrium (Jehiel, 2005) and cursed equilibrium (Eyster and Rabin, 2005).

Recent works have studied the implications of agents with misspecified models in various strategic settings. For instance, Bohren (2016), Bohren and Hauser (2017), Frick et al. (2019b) and Frick et al. (2019a) study social learning when agents have misspecified models that cause them to misinterpret other agents' actions. Mailath and Samuelson (2019) study a stylized prediction market where Bayesian agents have different models of the world (defined there as different partitions of a common state space), and discuss the possibility of information aggregation.

There are several works that consider outcomes when some agents behave in a way that can be construed as coming from a misspecified model. For instance in Spiegler (2006) or Spiegler (2013) society misunderstands the relationship between outcomes and the actions of strategic agents, which affects the actions the latter take in equilbrium and resulting outcomes (in the former, in the context of a market for quacks, in the latter with implications to the reforms taken by a policitican). Liang (2018) studies outcomes in games of incomplete information where agents behave like statisticians and have limited information.[13]

Finally, the understanding that agents should be cognizant that their models may be misspecified has also led to new approaches in mechanism design, where the designer accounts for misspecification in various ways. The literature on robust mechanism design (beginning with the seminal Bergemann and Morris 2005) provides foundations for using stronger solution concepts. Madarász and Prat (2017) shows that an optimal mechanism may perform very poorly if the planner's model is even slightly misspecified, and identifies a class of near optimal mechanisms that degrade gracefully. Works such as Chassang (2013) and Carroll (2015) develop optimal 'robust' contracts in general settings and contrast to classical optimal contracting.

References

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2016): "Fragility of asymptotic agreement under Bayesian learning," *Theoretical Economics*, 11, 187–225.

[13]There is a larger literature which studies the outcomes when agents are modeled as statisticians or machine learners, e.g., Al-Najjar (2009), Al-Najjar and Pai (2014), Acemoglu et al. (2016) and Cherry and Salant (2018).

Akaike, H. (1974): "A new look at the statistical model identification," *IEEE transactions on automatic control*, 19, 716–723.

Al-Najjar, N. I. (2009): "Decision makers as statisticians: Diversity, ambiguity, and learning," *Econometrica*, 77, 1371–1401.

Al-Najjar, N. I. and M. M. Pai (2014): "Coarse decision making and overfitting," *Journal of Economic Theory*, 150, 467–486.

Atakan, A. E. and M. Ekmekci (2014): "Auctions, actions, and the failure of information aggregation," *American Economic Review*, 104.

Bergemann, D. and S. Morris (2005): "Robust mechanism design," *Econometrica*, 73, 1771–1813.

Bohren, J. A. (2016): "Informational herding with model misspecification," *Journal of Economic Theory*, 163, 222–247.

Bohren, J. A. and D. Hauser (2017): "Bounded rationality and learning: A framework and a robustness result," *Working Paper, University of Pennsylvania*.

Bond, P. and H. Eraslan (2010): "Information-based trade," *Journal of Economic Theory*, 145, 1675–1703.

Bunke, O., X. Milhaud, et al. (1998): "Asymptotic behavior of Bayes estimates under possibly incorrect models," *The Annals of Statistics*, 26, 617–644.

Burnham, K. P. and D. R. Anderson (2003): *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Science & Business Media.

Carroll, G. (2015): "Robustness and linear contracts," *American Economic Review*, 105, 536–63.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350–1361.

Chassang, S. (2013): "Calibrated incentive contracts," *Econometrica*, 81, 1935–1971.

Cherry, J. and Y. Salant (2018): "Statistical Inference in Games," Tech. rep., mimeo.

Claeskens, G. and N. Hjort (2008): "Model selection and model averaging," *Cambridge Books*.

Eliaz, K. and R. Spiegler (2018): "A Model of Competing Narratives," *CEPR Discussion Paper No. DP13319*.

Esponda, I. and D. Pouzo (2016): "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84, 1093–1130.

Eyster, E. and M. Rabin (2005): "Cursed equilibrium," *Econometrica*, 73, 1623–1672.

Ferguson, T. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, vol. 7, Academic Press New York.

Frick, M., R. Iijima, and Y. Ishii (2019a): "Dispersed Behavior and Perceptions in Assortative Societies," .

——— (2019b): "Misinterpreting Others and the Fragility of Social Learning," *Cowles Foundation Discussion Paper*.

Harrison, J. M. and D. M. Kreps (1978): "Speculative investor behavior in a stock market with heterogeneous expectations," *The Quarterly Journal of Economics*, 92, 323–336.

Heidhues, P., B. Kőszegi, and P. Strack (2018): "Unrealistic expectations and misguided learning," *Econometrica*, 86, 1159–1214.

Horn, R. A. and C. R. Johnson (1990): *Matrix analysis*, Cambridge university press.

Jehiel, P. (2005): "Analogy-based expectation equilibrium," *Journal of Economic theory*, 123, 81–104.

Kass, R., L. Tierney, and J. B. Kadane (1990): "The validity of posterior expansions based on Laplaces method," in *Bayesian and Likelihood Methods in*

Statistics and Econometrics, ed. by S. Geisser, J. Hodges, S. Press, and A. Zellner, vol. 7, 473.

Kleijn, B., A. Van der Vaart, et al. (2012): "The Bernstein-von-Mises theorem under misspecification," *Electronic Journal of Statistics*, 6, 354–381.

Liang, A. (2018): "Games of Incomplete Information Played by Statisticians," *Working paper, University of Pennsylvania*.

Madarász, K. and A. Prat (2017): "Sellers with misspecified models," *The Review of Economic Studies*, 84, 790–815.

Mailath, G. J. and L. Samuelson (2019): "The Wisdom of a Confused Crowd: Model-Based Inference," *Cowles Foundation Discussion Paper*.

Mallows, C. L. (1973): "Some Comments on C_P," *Technometrics*, 15, 661–675.

Morris, S. (1994): "Trade with heterogeneous prior beliefs and asymmetric information," *Econometrica: Journal of the Econometric Society*, 1327–1347.

——— (1995): "The common prior assumption in economic theory," *Economics & Philosophy*, 11, 227–253.

Müller, U. K. (2013): "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix," *Econometrica*, 81, 1805–1849.

Nishii, R. (1984): "Asymptotic properties of criteria for selection of variables in multiple regression," *The Annals of Statistics*, 758–765.

Ortoleva, P. and E. Snowberg (2015): "Overconfidence in political behavior," *American Economic Review*, 105, 504–35.

Ottaviani, M. and P. N. Sørensen (2015): "Price reaction to information with heterogeneous beliefs and wealth effects: Underreaction, momentum, and reversal," *American Economic Review*, 105, 1–34.

Pearl, J. (2009): *Causality*, Cambridge university press.

Phillips, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

Scheinkman, J. A. and W. Xiong (2003): "Overconfidence and speculative bubbles," *Journal of political Economy*, 111, 1183–1220.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461–464.

Spiegler, R. (2006): "The market for quacks," *The Review of Economic Studies*, 73, 1113–1131.

——— (2013): "Placebo reforms," *American Economic Review*, 103, 1490–1506.

——— (2016): "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 131, 1243–1290.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Yildiz, M. (2003): "Bargaining without a common prior—an immediate agreement theorem," *Econometrica*, 71, 793–811.

A Main Appendix

A.1 Proof of Lemma 1

Proof. Fix a data set D_n. We need to analyze

$$\mathbb{E}_\pi \left[\mathbb{E}_P \left[(x'\beta - f^*_{(\pi, D_n)}(x))^2 \right] \middle| D_n \right].$$

Substituting f^* from (6), we have that this term

$$= \mathbb{E}_\pi \left[\mathbb{E}_P \left[((\beta - \mathbb{E}_\pi[\beta|D_n])'x)^2 \right] \middle| D_n \right].$$

Recalling that for a scalar a, $a = \mathrm{Tr}(a)$, we have

$$= \mathbb{E}_\pi \left[\mathbb{E}_P \left[\mathrm{Tr}[((\beta - \mathbb{E}_\pi[\beta|D_n])'x)^2] \right] \middle| D_n \right],$$

and then by symmetry and linearity of the trace operator, we can conclude,

$$\begin{aligned}
&= \mathbb{E}_\pi \left[\mathbb{E}_P \left[\mathrm{Tr}[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'xx'] \right] \middle| D_n \right] \\
&= \mathbb{E}_\pi \left[\mathrm{Tr}[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'\mathbb{E}_P[xx']] \middle| D_n \right] \\
&= \mathrm{Tr} \left[\mathbb{E}_\pi \left[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])' \middle| D_n \right] \mathbb{E}_P[xx'] \right].
\end{aligned}$$

Finally, by the definition of variance, we have the desired form

$$= \mathrm{Tr}(\mathbb{V}_\pi(\beta|D_n) E_P[xx']]$$ □

A.2 Proof of Theorem 1

Proof of (i): Consider two agents, one with prior $\pi_L \in \Pi_0^L$ and another with prior $\pi_0 \in \Pi_0$. Given dataset D_n, the agent with prior π_L defeats the agent with prior π_0 whenever

$$L^*(\pi_L, D_n) < L^*(\pi_0, D_n).$$

By Lemma 1 this happens if and only if

$$\mathbb{E}_{\pi_0}[\sigma^2|D_n] - \mathbb{E}_{\pi_L}[\sigma^2|D_n] \tag{12}$$

is strictly larger than

$$\text{Tr}\left(V_{\pi_L}\left(\beta_{J(\pi_L)}|D_n\right)\mathbb{E}_P[x_{J(\pi_L)}x_{J(\pi_L)}{}']\right) - \text{Tr}\left(V_{\pi_0}\left(\beta_{J(\pi_0)}|D_n\right)\mathbb{E}_P[x_{J(\pi_0)}x_{J(\pi_0)}{}']\right). \tag{13}$$

The proof has four main steps.

Step 0 (Maximum Likelihood Estimators): Since an agent with prior π only uses covariates with indices in $J(\pi)$, this agent's posterior can be obtained using the likelihood

$$f(Y|X_{J(\pi)};\beta_{J(\pi)},\sigma^2) := \frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}(Y - X_{J(\pi)}\beta_{J(\pi)})'(Y - X_{J(\pi)}\beta_{J(\pi)})\right). \tag{14}$$

Let $\widehat{\beta}(\pi)$ and $\widehat{\sigma}^2(\pi)$ denote the parameters $\theta(\pi) := (\beta_{J(\pi)},\sigma^2)$ that maximize such likelihood. It is well known that under Assumption 2:

$$\begin{aligned}\widehat{\theta}(\pi_L) &:= (\widehat{\beta}(\pi_L),\widehat{\sigma}^2(\pi_L)) \xrightarrow{p} (\beta_{0,J(\pi_L)},\sigma_0^2),\\ \widehat{\theta}(\pi_0) &:= (\widehat{\beta}(\pi_0),\widehat{\sigma}^2(\pi_0)) \xrightarrow{p} (\beta_{0,J(\pi_0)},\sigma_0^2).\end{aligned}$$

This happens because $J(\pi_L)$ nests the true model $J(\pi_0)$ and the true model estimates the coefficients of the best linear prediction of y given x. Moreover, standard algebra of linear regression[14] shows that

$$n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L)) = (\sqrt{n}R\widehat{\beta}(\pi_L))'[R(X'_{J(\pi_L)}X_{J(\pi_L)}/n)^{-1}R']^{-1}(\sqrt{n}R\widehat{\beta}(\pi_L)),$$

where R is the $|J(\pi_L) - J(\pi_0)| \times |J(\pi_L)|$ matrix that selects the entries of $\beta_{J(\pi_L)}$ that are zero under the model specified by π_0 and $|J|$ denotes the cardinality of the set J. Under Assumption 2

$$n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L))/\widehat{\sigma}^2(\pi_0) \xrightarrow{d} \zeta \equiv \xi'[R(\mathbb{E}_{\mathbb{P}}[x_{J(\pi_L)}x'_{J(\pi_L)}])^{-1}R']^{-1}\xi/\sigma_0^2,$$

where

$$\xi \sim \mathcal{N}_{|J(\pi_L)-J(\pi_0)|}(0, R\mathbb{E}_{\mathbb{P}}[(y - x'\beta_0)^2(x_{J(\pi_L)}x'_{J(\pi_L)})^{-1}]R').$$

[14] See Theorem 3.4 and Theorem 3.5 in Greene (2003).

Under conditional homoskedasticity ζ has a $\chi^2_{|J(\pi_L)-J(\pi_0)|}$ distribution and consequently.

$$n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L))/\widehat{\sigma}^2(\pi_0) \xrightarrow{d} \chi^2_{|J(\pi_L)|-|J(\pi_0)|}. \tag{15}$$

More generally, ζ is just a quadratic form of multivariate normal random variables.

One additional piece of notation. We define the scaled log-likelihood function for an agent with prior π as

$$h_n(\theta(\pi)) := \frac{1}{n} \ln f(Y|X_{J(\pi)}; \theta(\pi)).$$

The (i, j) component of the matrix of second derivatives of $h_n(\theta(\pi))$ with respect to $\theta(\pi)$ (the Hessian of the scaled log-likelihood) will be denoted as $h_{ij}(\cdot)$. We omit the dependence on n, unless confusion arises. The components of the inverse of the Hessian will be written as $h^{ij}(\cdot)$. Finally, $h_{rsj}(\cdot)$ denotes the partial derivative of h_{rs} with respect to the j-th component of $\theta(\pi)$.

Step 1 (Asymptotic Expansions of posterior moments): Kass et al. (1990) provide "large n" asymptotic expansions for posterior moments around the maximizer of the likelihood used to compute the posterior.

In the linear regression model, Theorem 4 and 5 in Kass et al. (1990) imply that for any prior π satisfying Assumption 1, $\mathbb{P}$ satisfying Assumption 2, and for any six-times differentiable positive real-valued function the posterior of $g(\theta)$ can be expanded as

$$\begin{aligned} E_\pi[g(\theta)|D_n] &= g(\widehat{\theta}(\pi)) + \frac{1}{n} \sum_{1\leq i,j\leq \dim(\theta(\pi))} \left(\frac{\partial g}{\partial \theta_i}(\widehat{\theta}(\pi))\right) h^{ij}(\widehat{\theta}(\pi)) \left\{ \left(\frac{\partial \pi}{\partial \theta_j}(\widehat{\theta}(\pi))\right) \cdot \right. \\ &\quad \left. \frac{1}{\pi(\widehat{\theta}(\pi))} - \frac{1}{2} \sum_{1\leq r,s\leq \dim(\theta(\pi))} h^{rs}(\widehat{\theta}(\pi)) h_{rsj}(\widehat{\theta}(\pi)) \right\} \\ &+ \frac{1}{2n} \sum_{1\leq i,j\leq \dim(\theta(\pi))} h^{ij}(\widehat{\theta}(\pi)) \left(\frac{\partial g}{\partial \theta_i \theta_j}(\widehat{\theta}(\pi))\right) \\ &+ O\left(\frac{1}{n^2}\right). \end{aligned}$$

See equation 2.6 in p. 481 of Kass et al. (1990).

Consider the positive function

$$g(\theta(\pi)) = g(\beta_{J(\pi)}, \sigma^2) = \sigma^2.$$

Because

$$\frac{\partial g}{\partial \sigma^2}(\widehat{\theta}(\pi)) = 1 \quad \text{and} \quad \frac{\partial g}{\partial \theta_i}(\widehat{\theta}(\pi)) = 0,$$

for any $i < |J(\pi)| + 1$, the expansion above simplifies to

$$\begin{aligned} E_\pi[\sigma^2|D_n] &= \widehat{\sigma}^2(\pi) + \frac{1}{n} \sum_{1 \le j \le |J(\pi)|+1} h^{(|J(\pi)|+1)j}(\widehat{\theta}(\pi)) \left\{ \left(\frac{\partial \pi}{\partial \theta_j}(\widehat{\theta}(\pi)) \right) \cdot \right. \\ &\quad \left. \frac{1}{\pi(\widehat{\theta}(\pi))} - \frac{1}{2} \sum_{1 \le r,s \le \dim(\theta(\pi))} h^{rs}(\widehat{\theta}(\pi)) h_{rsj}(\widehat{\theta}(\pi)) \right\} \\ &+ O\left(\frac{1}{n^2}\right). \end{aligned}$$

Moreover, the Hessian matrix of $h_n(\theta(\pi))$ equals

$$\begin{pmatrix} \frac{-1}{n\sigma^2} {X_{J(\pi)}}' X_{J(\pi)} & -\frac{1}{n\sigma^4} {X_{J(\pi)}}'(Y - {X_{J(\pi)}}'\beta_{J(\pi)}) \\ -\frac{1}{n\sigma^4}(Y - {X_{J(\pi)}}'\beta_{J(\pi)})' X_{J(\pi)} & \frac{1}{2\sigma^4} - \frac{1}{n\sigma^6}(Y - {X_{J(\pi)}}'(Y - X_{J(\pi)}), \end{pmatrix} \tag{16}$$

and the inverse Hessian evaluated at $\widehat{\theta}(\pi)$ is

$$\begin{pmatrix} -\widehat{\sigma}^2(\pi) \left({X_{J(\pi)}}' X_{J(\pi)}/n \right)^{-1} & \mathbf{0} \\ \mathbf{0} & -2\widehat{\sigma}^4(\pi), \end{pmatrix} \tag{17}$$

This further simplifies the expansion to

$$
\begin{aligned}
E_\pi[\sigma^2|D_n] &= \widehat{\sigma}^2(\pi) - \frac{2\widehat{\sigma}^4(\pi)}{n}\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\widehat{\theta}(\pi))\right)\cdot\frac{1}{\pi(\widehat{\theta}(\pi))}\right. \\
&- \left.\frac{1}{2}\sum_{1\le r,s\le|J(\pi)+1|} h^{rs}(\widehat{\theta}(\pi))h_{rs(|J(\pi)|+1)}(\widehat{\theta}(\pi))\right\} + O\left(\frac{1}{n^2}\right).
\end{aligned}
$$

Finally, the terms

$$
h^{r(|J(\pi)|+1)}, h^{(|J(\pi)|+1)s}
$$

are both 0 for any $r, s < |J(\pi)| + 1$. Algebra shows that

$$
\begin{aligned}
\sum_{1\le r,s\le|J(\pi)|+1} h^{rs}(\widehat{\theta}(\pi))h_{rs(|J(\pi)|+1)}(\widehat{\theta}(\pi)) &= \sum_{1\le r,s\le|J(\pi)|} h^{rs}(\widehat{\theta}(\pi))h_{rs(|J(\pi)|+1)}(\widehat{\theta}(\pi)) \\
&+ h^{(|J(\pi)|+1)(|J(\pi)|+1)}(\widehat{\theta}(\pi))\cdot \\
&\quad h_{(|J(\pi)|+1)(|J(\pi)|+1)(|J(\pi)|+1)}(\widehat{\theta}(\pi)) \\
&= -\widehat{\sigma}^{-2}(\pi)|J(\pi)| \\
&- 4\hat{\sigma}^{-2}.
\end{aligned}
$$

We conclude that the Kass-Tierney-Kadane expansion of $E_\pi[\sigma^2|D_n]$ equals

$$
\widehat{\sigma}^2(\pi) - \frac{2\widehat{\sigma}^4(\pi)}{n}\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\widehat{\theta}(\pi))\right)\cdot\frac{1}{\pi(\widehat{\theta}(\pi))}\right\} - \widehat{\sigma}^2(\pi)\frac{|J(\pi)|+4}{n} + O_P\left(\frac{1}{n^2}\right). \quad (18)
$$

STEP 2 (COMPARISON OF MODEL FIT): The expansion in (18) implies

$$\begin{aligned} n\left(E_{\pi_0}[\sigma^2|D_n] - E_{\pi_L}[\sigma^2|D_n]\right) &= n\left(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L)\right) \\ &- 2\sigma_0^4\left\{\left(\frac{\partial \pi_0}{\partial \sigma^2}(\theta_0(\pi_0))\right)\cdot\frac{1}{\pi_0(\theta_0(\pi_0))}\right. \\ &- \left.\left(\frac{\partial \pi_L}{\partial \sigma^2}(\theta_0(\pi_L))\right)\cdot\frac{1}{\pi_L(\theta_0(\pi_L))}\right\} \\ &- \sigma_0^2\left(|J(\pi_0)| - |J(\pi_L)|\right) \\ &+ O_P\left(\frac{1}{n}\right). \end{aligned}$$

STEP 3 (COMPARISON OF MODEL UNCERTAINTY): Let β_0, σ_0^2 denote the true parameters of the model as defined in the statement of Theorem 1. Under Assumption 2.4, for $\pi \in \{\pi_0, \pi_L\}$ we have

$$nV_\pi\left(\beta_{J(\pi)}|D_n\right) \xrightarrow{p} \sigma_0^2\mathbb{E}_{\mathbb{P}}[x_{J(\pi)}x'_{J(\pi)}]^{-1}.$$

Consequently:

$$n\mathrm{Tr}\left(V_\pi\left(\beta_{J(\pi)}|D_n\right)\mathbb{E}_P[x_{J(\pi)}x_{J(\pi)}']\right) \xrightarrow{p} \sigma_0^2\mathrm{Tr}(\mathbb{E}_{\mathbb{P}}[x_{J(\pi)}x'_{J(\pi)}]^{-1}\mathbb{E}_P[x_{J(\pi)}x_{J(\pi)}']).$$

STEP 4 (MODEL FIT VS. MODEL UNCERTAINTY): π_L defeats π_0 if the gain in model fit in equation (12) is larger than the increase in model uncertainty, as captured by (13).

Define

$$\eta_\pi(\theta_0(\pi)) := \left(\frac{\partial \pi_0}{\partial \sigma^2}(\theta_0(\pi_0))\right)\cdot\frac{\sigma_0^2}{\pi_0(\theta_0(\pi_0))}.$$

This parameter denotes the elasticity of the prior π with respect to the parameter σ^2 at the true parameter $\theta_0(\pi)$.

Let

$$\begin{aligned}\Delta(\pi_L, \pi_0) &\equiv \text{Tr}(\mathbb{E}_{\mathbb{P}}[x_{J(\pi_L)}x'_{J(\pi_L)}]^{-1}\mathbb{E}_P[x_{J(\pi_L)}x_{J(\pi_L)}{}'])\\ &- \text{Tr}(\mathbb{E}_{\mathbb{P}}[x_{J(\pi_0)}x'_{J(\pi_0)}]^{-1}\mathbb{E}_P[x_{J(\pi_0)}x_{J(\pi_0)}{}'])\end{aligned}$$

Step 2 and Step 3 imply that the probability of the event in which π_L defeats π_0 can be approximated in large samples by

$$\begin{aligned}n\left(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L)\right)/\sigma_0^2 &> \Delta(\pi_L, \pi_0)\\ &- (|J(\pi_L)| - |J(\pi_0)|)\\ &- 2(\eta_{\pi_L}(\theta_0(\pi_L)) - \eta_{\pi_0}(\theta_0(\pi_L)))\\ &+ O_p(1).\end{aligned}$$

We have shown that under Assumption 2

$$n\left(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L)\right)/\sigma_0^2 \xrightarrow{d} \zeta.$$

Consequently:

$$\mathbb{P}[\pi_L \succ_{D_n} \pi_0] \to c(\pi_L, \pi_0, \theta_0),$$

where $c(\pi_L, \pi_0, \theta_0)$ is the function

$$P\left(\zeta > \Delta(\pi_L, \pi_0) - (|J(\pi_L)| - |J(\pi_0)|) + 2(\eta_{\pi_0}(\theta_0(\pi_0)) - \eta_{\pi_L}(\theta_0(\pi_L)))\right).$$

Proof of (ii): We show that an agent with prior $\pi \in \Pi_0^W \cup \Pi_0^S$ can never defeat an agent with prior $\pi_0 \in \Pi_0$. Given dataset D_n, the agent with prior π_0 is victorious over π whenever

$$L^*(\pi_0, D_n) < L^*(\pi, D_n).$$

Using Lemma 1 this happens if and only if

$$\mathbb{E}_{\pi}[\sigma^2|D_n] - \mathbb{E}_{\pi_0}[\sigma^2|D_n] \tag{19}$$

is strictly larger than

$$\text{Tr}\left(V_{\pi_0}\left(\beta_{J(\pi_0)}|D_n\right)\mathbb{E}_P[x_{J(\pi_0)}x_{J(\pi_0)}{}']\right) - \text{Tr}\left(V_{\pi}\left(\beta_{J(\pi)}|D_n\right)\mathbb{E}_P[x_{J(\pi)}x_{J(\pi)}{}']\right). \tag{20}$$

Assumption 2.4 implies that (20) converges in probability to zero. In addition, using the expansion of Kass et al. (1990) we can write (19) as

$$\widehat{\sigma}^2(\pi) - \widehat{\sigma}^2(\pi_0) + O_P\left(\frac{1}{n}\right).$$

It is well-known that the probability limit of the difference

$$\widehat{\sigma}^2(\pi) - \widehat{\sigma}^2(\pi_0)$$

is strictly positive: under our assumptions, the misspecified model has strictly larger residual variance than the true model.

A.3 Proof of Proposition 1

Proof. Denote the single datapoint as $D_1 = (Y, X)$, where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^{1\times k}$ (k is the number of covariates), and $X = x'$. First, observe that for any agent j with a single explanatory variable κ in his model (denoted x_κ). By Lemma 2

$$\begin{aligned} L^*(\pi_j, D_1) &= \frac{b_0 + \frac{1}{2}\left(y^2 - \frac{y^2 x_\kappa^2}{x_\kappa^2+\gamma}\right)}{a_0 - \frac{1}{2}}\left(1 + \frac{1}{x_\kappa^2+\gamma}\right) \\ &= \frac{b_0 + \frac{1}{2}\frac{y^2\gamma}{x_\kappa^2+\gamma}}{a_0 - \frac{1}{2}}\left(1 + \frac{1}{x_\kappa^2+\gamma}\right). \end{aligned}$$

The winning agent among the single variable models will therefore clearly be the agent with the variable κ that maximizes x_κ. Without loss of generality, call this variable 1.

To economize on notation, now consider the full model with all the explanatory

variables, it will be clear from the logic that this argument will work for any model larger than a single variable. For an agent j with all k variables, we know that

$$L^*(\pi_j, D_1) = \frac{b_0 + \frac{y^2}{2}\left(1 - X(X'X + \gamma k \mathbb{I}_k)^{-1}X'\right)}{a_0 - \frac{1}{2}} \left(1 + \text{Tr}\left[(X'X + \gamma k \mathbb{I}_k)^{-1}\right]\right)$$

To show that this model always loses, we need to show that this model's loss is always larger than the "best" single variable model. To do this, it is sufficient to show that:

$$\left(1 - X(X'X + \gamma k \mathbb{I}_k)^{-1}X'\right) \geq \frac{\gamma}{x_1^2 + \gamma},$$
$$\text{Tr}\left[(X'X + \gamma k \mathbb{I}_k)^{-1}\right] \geq \frac{1}{x_1^2 + \gamma}.$$

We will handle each of these separately. Let's start with the second. Recall that for any matrix A, $\text{Tr}(A)$ equals the sum of eigenvalues of A. Further, the eigenvalues of A^{-1} are the reciprocals of the eigenvalues of matrix A for an invertible matrix. Finally if A is positive definite, all the eigenvalues are strictly positive.

By the Gershgorin circle theorem (see e.g. Theorem 6.1.1 of Horn and Johnson (1990)), all the eigenvalues of a matrix A lie within $\bigcup_{\kappa=1}^{k}[a_{\kappa,\kappa} - R_\kappa, a_{\kappa,\kappa} + R_\kappa]$ where R_κ is the sum of the absolute values of the non-diagonal terms on row κ, and $a_{\kappa,\kappa}$ is the κ diagonal element.

Consider the matrix $(X'X + \gamma k \mathbb{I}_k)$. Observe that R_κ in this case $= |x_\kappa|(\sum_{\kappa' \neq \kappa} |x_{\kappa'}|)$, while $a_{\kappa,\kappa} = x_\kappa^2 + k\gamma$. Therefore the largest possible eigenvalue is $|x_1|(\sum_\kappa |x_\kappa|) + k\gamma$, which in turn is small than $k(x_1^2 + \gamma)$.

Therefore for the matrix $(X'X + \gamma k \mathbb{I}_k)^{-1}$, all eigenvalues are larger than $\frac{1}{k(x_1^2+\gamma)}$, and therefore the sum of eigenvalues is at least $\frac{1}{(x_1^2+\gamma)}$ (since there are k eigenvalues)!

We can therefore conclude that

$$\text{Tr}\left[(X'X + \gamma k \mathbb{I}_k)^{-1}\right] \geq \frac{1}{x_1^2 + \gamma},$$

as desired.

We are left to prove that:

$$\begin{aligned}(1 - X(X'X + \gamma k\mathbb{I}_k)^{-1}X') &\geq \frac{\gamma}{x_1^2 + \gamma},\\ \iff X(X'X + \gamma k\mathbb{I}_k)^{-1}X' &\leq \frac{x_1^2}{x_1^2 + \gamma}.\end{aligned}$$

Now, observe that $X(X'X + \gamma k\mathbb{I}_k)^{-1}X'$ is a scalar. We know that for a scalar, $a = \text{Tr}(a)$. Therefore we have that

$$\begin{aligned}&X(X'X + \gamma k\mathbb{I}_k)^{-1}X',\\ =&\text{Tr}[X(X'X + \gamma k\mathbb{I}_k)^{-1}X'],\\ =&\text{Tr}[(X'X + \gamma k\mathbb{I}_k)^{-1}X'X]\\ =&\text{Tr}[(\frac{1}{\gamma k}X'X + \mathbb{I}_k)^{-1}\frac{1}{\gamma k}X'X]\end{aligned}$$

Denote $\frac{1}{\gamma k}X'X$ as A. Substituting

$$=\text{Tr}[(A + \mathbb{I}_k)^{-1}A]$$

Now, observe that if λ is an eigenvalue of A, then $\frac{\lambda}{1+\lambda}$ is an eigenvalue of $(A+\mathbb{I}_k)^{-1}A$. To see this, suppose v is an eigenvector of A with eigenvalue λ. Then,

$$\begin{aligned}&Av = \lambda v\\ \implies &(A + \mathbb{I}_k)v = (\lambda + 1)v\\ \implies &(A + \mathbb{I}_k)^{-1}v = \frac{1}{1 + \lambda}v\\ \implies &(A + \mathbb{I}_k)^{-1}Av = \frac{\lambda}{1 + \lambda}v\end{aligned}$$

Substituting this in, we have

$$\begin{aligned}&\text{Tr}[(A + \mathbb{I}_k)^{-1}A]\\ =&\sum_{i=1}^{k}\frac{\lambda_i}{1 + \lambda_i}\end{aligned}$$

Therefore we are left to show that

$$\sum_{i=1}^{k} \frac{\lambda_i}{1+\lambda_i} \leq \frac{x_1^2}{x_1^2+\gamma}$$

Here λ_i's are the eigenvalues of $\frac{1}{\gamma k} X'X$. This implies that $\sum_i \lambda_i = \frac{1}{\gamma k} \sum_i x_i^2$.

Note that $X'X$ is not full rank, indeed, its null space is of dimension $k-1$. Therefore it has $k-1$ multiplicity eigenvalue of 0. The unique non-zero eigenvalue must then be $\frac{1}{\gamma k} \sum_i x_i^2$.

Substituting in, we have

$$\begin{aligned}
\sum_{i=1}^{k} \frac{\lambda_i}{1+\lambda_i} =& \frac{\frac{1}{\gamma k} \sum_i x_i^2}{\frac{1}{\gamma k} \sum_i x_i^2 + 1} \\
=& \frac{\frac{1}{k} \sum_i x_i^2}{\frac{1}{k} \sum_i x_i^2 + \gamma} \\
\leq& \frac{x_1^2}{x_1^2+\gamma}
\end{aligned}$$

where the last inequality follows since we assumed that $x_1^2 = \max_i \{x_i^2 : 1 \leq i \leq k\}$. □

A.4 Proof of Theorem 2

Proof. It is well known that for a prior π in the Normal-Inverse Gamma family:

$$\begin{aligned}
\mathbb{V}_\pi[\beta_{J(\pi)}|D_n] &= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] (X'_{J(\pi)} X_{J(\pi)} + \gamma |J(\pi)| \mathbb{I}_{|J(\pi)|})^{-1} \\
&= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] \frac{1}{n} \left(\frac{X'_{J(\pi)} X_{J(\pi)}}{n} + \frac{\gamma |J(\pi)| \mathbb{I}_{|J(\pi)|}}{n} \right)^{-1}
\end{aligned}$$

Under the Assumptions on $\mathbb{P}$ in Theorem 2

$$\left(\frac{X'_{J(\pi)} X_{J(\pi)}}{n} + \frac{\gamma |J(\pi)| \mathbb{I}_{|J(\pi)|}}{n} \right)^{-1} = \mathbb{E}_{\mathbb{P}}[x_{J(\pi)} x'_{J(\pi)}]^{-1} + o_{\mathbb{P}}(1).$$

Consequently,

$$\text{Tr}\left(\mathbb{V}_\pi[\beta_{J(\pi)}|D_n]\mathbb{E}_{\mathbb{P}}[x_{J(\pi)}x'_{J(\pi)}]\right) = \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right]\left(\frac{J(\pi)}{n} + o_{\mathbb{P}}\left(\frac{1}{n}\right)\right). \tag{21}$$

Algebra shows that for any priors π, π' in the Normal-Inverse Gamma family

$$L^*(\pi', D_n) > L^*(\pi, D_n)$$

if and only if

$$\left(\mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] - \mathbb{E}_{\pi'}\left[\sigma_\epsilon^2|D_n\right]\right)\left(1 + \frac{J(\pi')}{n} + o_{\mathbb{P}}\left(\frac{1}{n}\right)\right) \tag{22}$$

is strictly larger than

$$\mathbb{E}_\pi[\sigma_\epsilon^2|D_n]\left(\frac{J(\pi) - J(\pi')}{n}\right). \tag{23}$$

Proof of (i): It is well known that for a prior π in the Normal-Inverse Gamma family, the posterior mean of $\beta_{J(\pi)}$ is the 'Ridge estimator'

$$\widehat{\beta}_\pi := (X'_{J(\pi)}X_{J(\pi)} + \gamma|J(\pi)|\mathbb{I}_{J(\pi)})^{-1}X'_{J(\pi)}y,$$

which solves the problem

$$\min_{\beta\in\mathbb{R}^{|J(\pi)|}}(y - X_{J(\pi)}\beta)'(y - X_{J(\pi)}\beta) + (\gamma|J(\pi)|)\,||\beta||^2$$

Consider two priors π, π' such that $J(\pi') \subset J(\pi)$. In a slight abuse of notation let $\widehat{\beta}_{\pi'}$ denote the vector in $\mathbb{R}^{J(\pi)}$ with all the coordinates in $J(\pi)\backslash J(\pi')$ equal to zero. Also, let J be used to abbreviate $J(\pi)$

Equation (10) implies that for any such two priors π, π'

$$n(\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_\pi[\sigma_\epsilon^2|D_n])$$

is proportional to the sum of

$$(y - X_J\widehat{\beta}_{\pi'})'(y - X_J\widehat{\beta}_{\pi'}) - (y - X_J\widehat{\beta}_\pi)'(y - X_J\widehat{\beta}_\pi) \tag{24}$$

and

$$\gamma \left(|J(\pi')| \, ||\widehat{\beta}_{\pi'}||^2 - |J| \, ||\widehat{\beta}_{\pi}||^2 \right). \tag{25}$$

where the proportionality constant is $c_n := (2a_0/n + 1 - 2/n)^{-1}$.

Algebra shows that the expression in (24) equals

$$-2(y - X_J\widehat{\beta}_{\pi})'X_J(\widehat{\beta}_{\pi'} - \widehat{\beta}_{\pi}) + (\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'})'X_J'X_J(\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'})$$

and the expression in (25)

$$\gamma|J(\pi')|(\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'})'(\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'}) - \gamma(|J| - |J(\pi')|)\widehat{\beta}_{\pi}'\widehat{\beta}_{\pi} + 2\gamma|J(\pi')|\widehat{\beta}_{\pi}'(\widehat{\beta}_{\pi'} - \widehat{\beta}_{\pi}).$$

The first-order conditions defining the Ridge estimator imply

$$-2(y - X_J\widehat{\beta}_{\pi})'X_J + 2\gamma|J|\widehat{\beta}_{\pi}' = 0.$$

Therefore, in any finite sample

$$\begin{aligned} n(\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_{\pi}[\sigma_\epsilon^2|D_n]) &= c_n\big((\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'})'(X_J'X_J + \gamma|J(\pi')|\mathbb{I}_J)(\widehat{\beta}_{\pi} - \widehat{\beta}_{\pi'}) \\ &+ \gamma(|J| - |J(\pi')|)\widehat{\beta}_{\pi}'\widehat{\beta}_{\pi} \\ &- 2\gamma(|J| - |J(\pi')|)\widehat{\beta}_{\pi'}'\widehat{\beta}_{\pi}\big). \end{aligned}$$

Under the assumptions of Theorem (2) and letting $\pi' \in \Pi_0$:

$$n(\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_{\pi}[\sigma_\epsilon^2|D_n]) = O_{\mathbb{P}}(1).$$

However, under the same assumptions

$$\mathbb{E}_{\pi}[\sigma_\epsilon^2|D_n] = \frac{2b_n}{n} + o_{\mathbb{P}}(1).$$

Since $b_n \in O\left(n^{v+2}\right)$. This implies

$$\mathbb{P}\left(n(\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_{\pi}[\sigma_\epsilon^2|D_n]) \left(1 + \frac{J(\pi')}{n} + o_{\mathbb{P}}(1)\right) > \mathbb{E}_{\pi}[\sigma_\epsilon^2|D_n])(J(\pi) - J(\pi')) \right)$$

converges to zero. We conclude that $\pi \in \Pi_0^L$, $\pi' \in \Pi_0$ implies

$$\mathbb{P}[\pi \succ_{D_n} \pi'] \to 0.$$

Proof of (ii): Consider the same framework as above, but let π now denote an element of Π_0 and π' an element of Π_0^S. The probability that the smaller model, π', is defeated by π is

$$\mathbb{P}\left((\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_\pi[\sigma_\epsilon^2|D_n])\left(1 + \frac{J(\pi')}{n} + o_{\mathbb{P}}(1)\right) > \frac{\mathbb{E}_\pi[\sigma_\epsilon^2|D_n])}{n}(J(\pi) - J(\pi'))\right),$$

where $J(\pi) > J(\pi')$. Under the assumptions of the theorem

$$(\mathbb{E}_{\pi'}[\sigma_\epsilon^2|D_n] - \mathbb{E}_\pi[\sigma_\epsilon^2|D_n]) = O_{\mathbb{P}}(1).$$

However,

$$\frac{\mathbb{E}_\pi[\sigma_\epsilon^2|D_n])}{n} = \frac{b_n}{n^2} + o_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Since $b_n \in O\left(n^{\upsilon+2}\right)$, for $\upsilon > 0$. We conclude that if $\pi' \in \Pi_0^S$ and $\pi \in \Pi_0$:

$$\mathbb{P}[\pi' \succ_{D_n} \pi] \to 1.$$

□

B Supplementary Material

B.1 Posterior Loss for Normal-Inverse Gamma Priors

We derive the specific formula of the posterior loss in the case of Normal-Inverse Gamma priors.

Lemma 2. *Suppose the agent has a Normal-Inverse gamma prior π with hyper-parameters (γ, a_0, b_0). Then, if the observed dataset is $D_n = (y, X)$ we have that her log posterior expected loss can be written as:*

$$\ln(L^*(\pi, D_n)) = \ln\left(\frac{\frac{2b_0}{n} + \frac{1}{n}\min_{\beta\in\mathbb{R}^{|J(\pi)|}}\left((y - X_{J(\pi)}\beta)'(y - X_{J(\pi)}\beta) + (\gamma|J(\pi)|)||\beta||^2\right)}{\frac{2a_0}{n} + 1 - \frac{2}{n}}\right) + \ln\left(1 + \mathrm{Tr}\left[\left((X'_{J(\pi)}X_{J(\pi)} + \gamma|J|\mathbb{I}_{|J(\pi)|}\right)^{-1}\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}]\right]\right) \tag{26}$$

Proof. We break the proof into two steps. Step 1 shows provides an expression for the posterior mean of σ^2_ϵ. Step 2 plugs-in this expression into the formula for the posterior loss.

Step 1 First we show that the posterior mean of σ^2_ϵ in a regression model with a Normal-Inverse Gamma prior with hyperparameters (γ, a_0, b_0) is given by:

$$\mathbb{E}_\pi[\sigma^2_\epsilon|D_n] = \frac{\frac{2b_0}{n} + \frac{1}{n}\min_{\beta\in\mathbb{R}^k}(y - X\beta)'(y - X\beta) + (\gamma k)\,||\beta||^2}{\frac{2a_0}{n} + 1 - \frac{2}{n}} \tag{27}$$

It is known that

$$\sigma^2_\epsilon|D_n \sim \text{Inv-Gamma}\Big(a_0 + \frac{n}{2}\,,\, b_0 + \frac{1}{2}(y'y - \widehat{\beta}'_{\mathrm{R}}(\gamma k)(X'X + (\gamma k)\mathbb{I}_k)\widehat{\beta}_{\mathrm{R}}(\gamma k))\Big).$$

where $\widehat{\beta}_{\mathrm{R}}(\gamma k)$ is the ridge estimator with penalty parameter γk. Since the mean of a random variable distributed as Inv-Gamma(a, b) is $\frac{b}{a-1}$, to show (10) it is sufficient to show that:

$$\min_{\beta\in\mathbb{R}^k}(y - X\beta)'(y - X\beta) + (\gamma k)\,||\beta||^2 = y'y - \widehat{\beta}_{\mathrm{R}}(\gamma k)'(X'X + (\gamma k)\mathbb{I}_k)\widehat{\beta}_{\mathrm{R}}(\gamma k). \tag{28}$$

To condense notation, let $\widehat{\beta}_{\mathrm{R}} \equiv \widehat{\beta}_{\mathrm{R}}(\lambda)$, where $\lambda = \gamma k$ is fixed. Note that:

$$\begin{aligned} y'X\widehat{\beta}_{\mathrm{R}} =& y'X(X'X + \lambda \mathbb{I}_k)^{-1}X'y \\ =& y'X(X'X + \lambda \mathbb{I}_k)^{-1}(X'X + \lambda \mathbb{I}_k)(X'X + \lambda \mathbb{I}_k)^{-1}X'y \\ =& \widehat{\beta}_{\mathrm{R}}'(X'X + \lambda \mathbb{I}_k)\widehat{\beta}_{\mathrm{R}} \\ =& \widehat{\beta}_{\mathrm{R}}'X'X\widehat{\beta}_{\mathrm{R}} + \lambda \widehat{\beta}_{\mathrm{R}}'\widehat{\beta}_{\mathrm{R}}. \end{aligned}$$

This implies that

$$\begin{aligned} & (y - X\widehat{\beta}_{\mathrm{R}})'(y - X\widehat{\beta}_{\mathrm{R}}) \\ =& y'y - 2y'X\widehat{\beta}_{\mathrm{R}} + \widehat{\beta}_{\mathrm{R}}'X'X\widehat{\beta}_{\mathrm{R}} \\ =& y'y - \widehat{\beta}_{\mathrm{R}}'X'X\widehat{\beta}_{\mathrm{R}} - 2\lambda\widehat{\beta}_{\mathrm{R}}'\widehat{\beta}_{\mathrm{R}}. \end{aligned}$$

Therefore:

$$\begin{aligned} & y'y - \widehat{\beta}_{\mathrm{R}}'(X'X + \lambda \mathbb{I}_k)\widehat{\beta}_{\mathrm{R}} \\ =& y'y - \widehat{\beta}_{\mathrm{R}}'(X'X)\widehat{\beta}_{\mathrm{R}} - \lambda\widehat{\beta}_{\mathrm{R}}'\widehat{\beta}_{\mathrm{R}} \\ =& (y - X\widehat{\beta}_{\mathrm{R}})'(y - X\widehat{\beta}_{\mathrm{R}}) + \lambda\widehat{\beta}_{\mathrm{R}}'\widehat{\beta}_{\mathrm{R}} \end{aligned}$$

Comparing, (28) follows, concluding our proof of (10).

Step 2 From Lemma 1, we have that the posterior loss

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma_\epsilon^2|D_n] + \int_0^\infty \mathrm{Tr}(\mathbb{V}_\pi(\beta|D_n, \sigma_\epsilon^2)\mathbb{E}_P[xx'])\pi(\sigma_\epsilon^2|D_n)d\sigma_\epsilon^2.$$

It is known that

$$\mathbb{V}_\pi(\beta|D_n, \sigma_\epsilon^2) = \sigma_\epsilon^2(X'X + (\gamma k)\mathbb{I}_k)^{-1},$$

This implies that

$$\begin{aligned} L^*(\pi, D_n) =& \mathbb{E}_\pi[\sigma_\epsilon^2|D_n] + \int_0^\infty \text{Tr}(\sigma_\epsilon^2(X'X + (\gamma k)\mathbb{I}_k)^{-1})\mathbb{E}_P[xx'])\pi(\sigma_\epsilon^2|D_n)d\sigma_\epsilon^2, \\ =& \mathbb{E}_\pi[\sigma_\epsilon^2|D_n] + \mathbb{E}_\pi[\sigma_\epsilon^2|D_n]\text{Tr}((X'X + (\gamma k)\mathbb{I}_k)^{-1}\mathbb{E}_p[xx']). \end{aligned}$$

Taking logs on both sides and using the formula for the posterior mean of σ_ϵ^2 from Step 1, we obtain the desired formula. □

B.2 Posterior Loss for Normal-Inverse Gamma priors in large samples

Observation 1. *Suppose the agent has a Normal-Inverse Gamma prior. Then, for n large, we have*

$$\ln\left(L^*(\pi, D_n)\right) \approx \underbrace{\ln\left(\mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right]\right)}_{\textit{Model Fit}} + \underbrace{\ln\left(1 + \frac{|J|}{n}\right)}_{\textit{Model Dimension}}. \tag{29}$$

Proof. The posterior upon observing dataset D_n is

$$\begin{aligned} \beta_J|D_n, \sigma_\epsilon^2 \sim \mathcal{N}_{|J|}(\widehat{\beta}_{J,\text{Ridge}}, \sigma_\epsilon^2(X_J'X_J + (\gamma|J|)\mathbb{I}_{|J|})^{-1}), \\ \implies \mathbb{V}_\pi[\beta_J|D_n] = \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right](X_J'X_J + (\gamma|J|)\mathbb{I}_{|J|})^{-1}. \end{aligned}$$

Therefore, substituting back, we have that

$$\begin{aligned} \text{Tr}\left(\mathbb{V}_\pi[\beta_J|D_n]\,\mathbb{E}_P[x_Jx_J']\right) &= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right]\text{Tr}\left((X_J'X_J + (\gamma|J|)\mathbb{I}_{|J|})^{-1}\,\mathbb{E}_P[x_Jx_J']\right), \\ &= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right]\text{Tr}\left(\frac{1}{n}\left(\frac{1}{n}\left(X_J'X_J + (\gamma|J|)\mathbb{I}_{|J|}\right)\right)^{-1}\,\mathbb{E}_P[x_Jx_J']\right), \end{aligned}$$

which for n large, by the law of large numbers

$$\begin{aligned}
&\approx \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] \operatorname{Tr}\left(\frac{1}{n}\left(\mathbb{E}_P[x_J x_J']\right)^{-1}\ \mathbb{E}_P[x_J x_J']\right), \\
&= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] \operatorname{Tr}\left(\frac{1}{n}\mathbb{I}_{|J|}\right), \\
&= \mathbb{E}_\pi\left[\sigma_\epsilon^2|D_n\right] \frac{|J|}{n}.
\end{aligned}$$

Thus for n large, (29) follows. □

B.3 Proof of Proposition 2

Proof. Suppose the known variance of ϵ is σ_ϵ^2. Then for any agent with prior π, upon seeing data D_n, the posterior expected loss evaluates to:

$$L^*(\pi_j, D_n) = \sigma_\epsilon^2 + \operatorname{Tr}(\mathbb{V}_\pi[\beta|D_n]),$$

where we have assumed that $\mathbb{E}_p[xx'] = \mathbb{I}$.

Without loss of generality, suppose the larger model J' is the entire set of observables of size k. We need to show that there exists a model J of size $|J|$ such that

$$\operatorname{Tr}(X'X + \gamma k\mathbb{I}_k)^{-1} \geq \operatorname{Tr}(X_J'X_J + \gamma|J|\mathbb{I}_{|J|})^{-1}.$$

In particular let J be such that $\sum_{j\in J} e_j(X'X+\gamma k\mathbb{I}_k)^{-1}e_j \leq \sum_{j\in J''} e_j(X'X+\gamma k\mathbb{I}_k)^{-1}e_j$ for any J'' such that $|J''| = |J|$. Then, it must be the case that

$$\operatorname{Tr}(X'X + \gamma k\mathbb{I}_k)^{-1} \geq \frac{k}{|J|}\sum_{j\in J} e_j(X'X + \gamma k\mathbb{I}_k)^{-1}e_j$$

Therefore it is sufficient to show that for this model J, we have

$$\frac{k}{|J|}\sum_{j\in J} e_j(X'X + \gamma k\mathbb{I}_k)^{-1}e_j \geq \operatorname{Tr}(X_J'X_J + \gamma|J|\mathbb{I}_{|J|})^{-1}.$$

Without loss we can renumber the indices so that $J = \{1, 2, \ldots, |J|\}$. Let L denote

the set of remaining indices, i.e. $L = \{|J|+1, \ldots, k\}$. We can thus write the left hand size of the inequality as:

$$\frac{k}{|J|}\sum_{j\in J} e_j \begin{pmatrix} X_J'X_J + \gamma k \mathbb{I}_{|J|} & X_J'X_L \\ X_L'X_J & X_L'X_L + \gamma k \mathbb{I}_{|L|} \end{pmatrix}^{-1} e_j.$$

Using the standard formula for block inverse of a matrix we can write this as

$$= \frac{k}{|J|}\sum_{j\in J} e_j \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}^{-1} e_j.$$

where $A_1 = (X_J'X_J + \gamma k \mathbb{I}_{|J|} - X_J'X_L(X_L'X_L + \gamma k \mathbb{I}_{|L|})^{-1}X_L'X_J)^{-1}$. Substituting that in we have

$$= \frac{k}{|J|}\mathrm{Tr}(X_J'X_J + \gamma k \mathbb{I}_{|J|} - X_J'X_L(X_L'X_L + \gamma k \mathbb{I}_{|L|})^{-1}X_L'X_J)^{-1}.$$

Therefore, taking $\frac{k}{|J|}$ to the other side, we are left to show that

$$\mathrm{Tr}(X_J'X_J + \gamma k \mathbb{I}_{|J|} - X_J'X_L(X_L'X_L + \gamma k \mathbb{I}_{|L|})^{-1}X_L'X_J)^{-1} \geq \mathrm{Tr}(\frac{k}{|J|}X_J'X_J + \gamma k \mathbb{I}_{|J|})^{-1} \tag{30}$$

Next, given 4 matrices A,B,C, and D where A and C are invertible, it is easy to show that

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1}.$$

Suppose we define

$$\begin{aligned} A &= X_J'X_J + \gamma k \mathbb{I}_{|J|}, \\ B &= -X_J'X_L, \\ C &= (X_L'X_L + \gamma k \mathbb{I}_{|L|})^{-1}, \\ D &= X_L'X_J. \end{aligned}$$

Note that in this case, A and C are invertible by observation. In light of this, and

the linearity of the Trace operator, we can rewrite the left hand side of (30) as

$$\begin{aligned}
&\mathrm{Tr}(A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1}) \\
=&\mathrm{Tr}A^{-1} - \mathrm{Tr}(A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1}) \\
=&\mathrm{Tr}(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1} - \mathrm{Tr}(A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1})
\end{aligned}$$

where A, B, C and D are as defined above. So (30) can be written as:

$$\mathrm{Tr}(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1} - \mathrm{Tr}(A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1}) \geq \mathrm{Tr}(\frac{k}{|J|}X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1}$$

To show this inequality it is therefore sufficient to show that

$$\mathrm{Tr}(A^{-1}B(C^{-1} + DA^{-1}B)DA^{-1}) \leq 0, \tag{31}$$

$$\mathrm{Tr}(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1} \geq \mathrm{Tr}(\frac{k}{|J|}X_J'X_J + \gamma k\mathbb{I}_k)^{-1}. \tag{32}$$

We now show each of these in turn. Let us start with the first. Note that $B = -D'$ we have:

$$(31) \iff \mathrm{Tr}(A^{-1}D'(C^{-1} - DA^{-1}D')DA^{-1}) \geq 0.$$

In turn, since A is symmetric, so is A^{-1}, so defining $Q \equiv A^{-1}D'$

$$\iff \mathrm{Tr}(Q(C^{-1} - DA^{-1}D')Q') \geq 0.$$

Since QMQ' is a positive semidefinite matrix if M is a positive semidefinite matrix (see e.g. Observation 7.1.8 of Horn and Johnson (1990)), it is sufficient to show that $(C^{-1} - DA^{-1}D')$ is a positive semidefinite matrix (the trace of a matrix equals the sum of all its eigenvalues, and the eigenvalues of a positive semidefinite matrix are all non-negative). So to show (31), it is sufficient to show that $(C^{-1} - DA^{-1}D')$ is positive semidefinite. To see this, observe that:

$$\begin{aligned}
&(C^{-1} - DA^{-1}D') \\
=&X_L'X_L + \gamma k\mathbb{I}_{|L|} - X_L'X_J(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1}X_J'X_L \\
=&X_L'\left(\mathbb{I}_N - X_J(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1}X_J'\right)X_L + \gamma k\mathbb{I}_{|L|}
\end{aligned}$$

It is therefore sufficient to show that each of these two matrices are positive semidefinite. The latter is positive definite by observation. To show that the former is positive semidefinite, by another appeal to Observation 7.1.8 of Horn and Johnson (1990), it is sufficient to show that $\left(\mathbb{I}_k - X_J(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1}X_J'\right)$ is positive semidefinite. But observe that:

$$\begin{aligned}&\mathbb{I}_k - X_J(X_J'X_J + \gamma k\mathbb{I}_{|J|})^{-1}X_J' \\ =&\mathbb{I}_k - \frac{1}{\gamma k}X_J(\frac{1}{\gamma k}X_J'X_J + \mathbb{I}_{|J|})^{-1}X_J' \end{aligned} \tag{33}$$

Now, we know that for any square matrix P,

$$\begin{aligned}(\mathbb{I}+P)^{-1} &= I - (\mathbb{I}+P)^{-1}P, \\ &= I - P + (\mathbb{I}+P)^{-1}P^2, \\ &= I + \sum_{j=1}^{\infty}(-1)^j P^j.\end{aligned}$$

Substituting in $P = \frac{1}{\gamma k}X_J'X_J$, we have that

$$\begin{aligned}X_J(\frac{1}{\gamma k}X_J'X_J + \mathbb{I}_{|J|})^{-1}X_J' &= X_J\left(\mathbb{I}_{|J|} - \sum_{j=1}^{\infty}(-\frac{1}{\gamma k})^j(X_J'X_J)^j\right)X_J' \\ &= X_JX_J' - \sum_{j=1}^{\infty}(-\frac{1}{\gamma k})^j(X_JX_J')^{j+1} \\ &= (X_JX_J')(\mathbb{I}_{|J|} - \sum_{j=1}^{\infty}(-\frac{1}{\gamma k})^j(X_JX_J')^j) \\ &= (X_JX_J')(\mathbb{I}_{|J|} + \frac{1}{\gamma k}X_JX_J')^{-1}\end{aligned}$$

Therefore we have that

$$\begin{aligned}(33) &= \mathbb{I}_k - \frac{1}{\gamma k}(X_JX_J')(\mathbb{I}_{|J|} + \frac{1}{\gamma k}X_JX_J')^{-1} \\ &= \mathbb{I}_k - (X_JX_J')(\gamma k\mathbb{I}_{|J|} + X_JX_J')^{-1} \\ &= \gamma k(\gamma k\mathbb{I}_{|J|} + X_JX_J')^{-1}\end{aligned}$$

which is positive definite by observation.

We are left, then, to show (32), i.e. that:

$$\begin{aligned}&\mathrm{Tr}(X_J'X_J+\gamma k\mathbb{I}_{|J|})^{-1}\geq \mathrm{Tr}(\frac{k}{|J|}X_J'X_J+\gamma k\mathbb{I}_k)^{-1},\\ \iff &\mathrm{Tr}((X_J'X_J+\gamma k\mathbb{I}_{|J|})^{-1}-(\frac{k}{|J|}X_J'X_J+\gamma k\mathbb{I}_k)^{-1})\geq 0.\end{aligned}$$

Algebra shows

$$\begin{aligned}&\mathrm{Tr}((X_J'X_J+\gamma k\mathbb{I}_{|J|})^{-1}-(\frac{k}{|J|}X_J'X_J+\gamma k\mathbb{I}_k)^{-1})\\ =&\mathrm{Tr}((X_J'X_J+\gamma k\mathbb{I}_{|J|})^{-1}\left(\frac{k-|J|}{|J|}X_J'X_J\right)(\frac{k}{|J|}X_J'X_J+\gamma k\mathbb{I}_k)^{-1})\\ =&\mathrm{Tr}(\frac{k-|J|}{|J|}X_J(X_J'X_J+\gamma k\mathbb{I}_{|J|})^{-1}(\frac{k}{|J|}X_J'X_J+\gamma k\mathbb{I}_k)^{-1}X_J').\end{aligned}$$

The final matrix into the trace operator is positive semidefinite by Observation 7.1.8 of Horn and Johnson (1990). □

B.4 Proof of Proposition 3

Proof. For an agent with prior π the agent's ex-ante expected loss on seeing a dataset of size n is

$$\mathbb{E}_{m(\pi)}[L^*(\pi,D_n)]=\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}\int_{y,x}(y-x'\widehat{\beta}(D_n))^2dQ_\theta(y,x)dQ_\theta(D_n)d\pi(\theta).$$

The agents' statistical model is $y=x'\beta+\epsilon$, $\epsilon\sim\mathcal{N}(0,\sigma^2)$

$$\begin{aligned}&=\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}\int_{y,x}(x'\beta+\epsilon-x'\widehat{\beta}(D_n))^2dQ_\theta(x,\epsilon)dQ_\theta(D_n)d\pi(\theta),\\ &=\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}\int_{x,\epsilon}((x'(\beta-\widehat{\beta}(D_n)))^2+\epsilon^2)dQ_\theta(x,\epsilon)dQ_\theta(D_n)d\pi(\theta)\\ &=\mathbb{E}_\pi[\sigma_\epsilon^2]+\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}\int_x(x'(\beta-\widehat{\beta}(D_n)))^2dQ_\theta(x)dQ_\theta(D_n)d\pi(\theta)\\ &=\mathbb{E}_\pi[\sigma_\epsilon^2]+\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}\left(\int_x((\beta-\widehat{\beta}(D_n))'xx'(\beta-\widehat{\beta}(D_n))dQ_\theta(x)\right)dQ_\theta(D_n)d\pi_J(\theta)\\ &=\mathbb{E}_{\pi_J}[\sigma_\epsilon^2]+\int_{\theta=(\beta,\sigma_\epsilon^2)}\int_{D_n}((\beta-\widehat{\beta}(D_n))'(\beta-\widehat{\beta}(D_n))dQ_\theta(D_n)d\pi(\theta)\end{aligned}$$

Where the last equality follows since $\mathbb{E}_P[xx'] = \mathbb{I}$ by assumption. Now, since $\gamma = 0$ by assumption, for dataset $D_n = (Y, X)$, we have that $\widehat{\beta}(D_n) = (X'_{J(\pi)}X_{J(\pi)})^{-1}X'_{J(\pi)}Y$. In a slight abuse of notation abbreviate $J(\pi)$ as J. Writing that $Y = X_J\beta + e$, where e is the $n \times 1$ vector collecting ϵ_i:

$$(\widehat{\beta}(D_n) - \beta) = (X'_J X_J)^{-1}X'_J e.$$

Substituting back in we have that:

$$\begin{aligned}
\mathbb{E}_{m(\pi)}[L^*(\pi, D_n)] &= \mathbb{E}_\pi[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \int_{D_n} ((\beta - \widehat{\beta}(D_n))'(\beta - \widehat{\beta}(D_n)) dQ_\theta(D_n) d\pi(\theta) \\
&= \mathbb{E}_{\pi_J}[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \int_{D_n} \left(e' X_J (X'_J X_J)^{-1} (X'_J X_J)^{-1} X'_J e\right) dQ_\theta(D_n) d\pi(\theta)
\end{aligned}$$

since $e' X_J (X'_J X_J)^{-1} (X'_J X_J)^{-1} X'_J e$ is a scalar

$$= \mathbb{E}_\pi[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \int_{D_n} \mathrm{Tr}(e' X_J (X'_J X_J)^{-1} (X'_J X_J)^{-1} X'_J e) dQ_\theta(D_n) d\pi_J(\theta)$$

Using the cyclic property of the trace operator,

$$= \mathbb{E}_\pi[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \int_{D_n} \mathrm{Tr}((X'_J X_J)^{-1} X'_J e e' X_J (X'_J X_J)^{-1}) dQ_\theta(D_n) d\pi_J(\theta)$$

by assumption, X_J and e are independent and $\mathbb{E}_{Q_\theta}[ee'] = \sigma_\epsilon^2 \mathbb{I}_n$. Thus,

$$\begin{aligned}
&= \mathbb{E}_\pi[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \sigma_\epsilon^2 \int_{X_J} \mathrm{Tr}((X'_J X_J)^{-1} X'_J X_J (X'_J X_J)^{-1}) dQ_\theta(X_J) d\pi_J(\theta) \\
&= \mathbb{E}_\pi[\sigma_\epsilon^2] + \int_{\theta=(\beta,\sigma_\epsilon^2)} \sigma_\epsilon^2 \int_{X_J} \mathrm{Tr}(X'_J X_J)^{-1} dQ_\theta(X_J) d\pi_J(\theta) \\
&= \mathbb{E}_\pi[\sigma_\epsilon^2] \left(1 + \frac{|J|}{n - |J| - 1}\right).
\end{aligned}$$

The last equation follows because when $x \sim \mathcal{N}(0, \mathbb{I}_k)$, $(X'_J X_J)$ is a Wishart distribution $\mathcal{W}(\mathbb{I}_J, n)$. Thus, $(X'_J X_J)^{-1}$ has an inverse wishart distribution and its expectation

equals $\mathbb{I}_J/(n-|J|-1)$, provided $n > |J|+1$. Finally

$$\frac{J'}{n-J'-1} < \frac{J}{n-J-1},$$

if and only if $n > 1$. Since $\mathbb{E}_\pi[\sigma_\epsilon^2]$ is common across all agents by assumption, the result follows. □