# Обзор моделей

В случае обычной линейной регрессии, мы используем метод наименьших квадратов для получения оценок неизвестных коэффициентов, тем самым минимизируя сумму квадратов отклонений истинных значений объясняемой переменной от значений, спрогнозированных нашей моделью. То есть наша задача состоит в том, чтобы подобрать такие коэффициенты, при которых функция потерь (1) достигает минимума:

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2, \tag{1}$$

где $n$ – число наблюдений, $p$ - число регрессоров.

Одним из свойств ошибки прогноза линейной регресии является её разложение на смещение и разброс. Пусть $y = f(\vec{x}) + \epsilon$, где $f(\vec{x})$ - детерминированная функция, причём истинное значение переменной распределено нормально $y \sim \mathcal{N}\left(f\left(\vec{x}\right), \sigma^2\right)$ и случайная ошибка $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ также нормально распределённая. Мы пытаемся приблизить неизвестную функцию $f(\vec{x})$ линейной функцией от регрессоров $\hat{f}\left(\vec{x}\right)$. Таким образом, ошибка раскладывается как

$$E[(y - \hat{f})^2] = E[y^2] + E[\hat{f}^2] - 2E[y\hat{f}] \tag{2}$$

Заметим, что

$$E[y^2] = Var(y) + E[y]^2 = \sigma^2 + f^2$$

$$E[\hat{f}^2] = Var(\hat{f}) + E[\hat{f}]^2 \tag{3}$$

$$E[y\hat{f}] = E[(f + \epsilon)\hat{f}] = E[f\hat{f}] + E[\epsilon\hat{f}] = fE[\hat{f}] + E[\epsilon]E[\hat{f}] = fE[\hat{f}]$$

Объединяя всё полученное, имеем следующий результат:

$$\begin{aligned} E[(y - \hat{f})^2] &= \sigma^2 + f^2 + Var(\hat{f}) + E[\hat{f}]^2 - 2fE[\hat{f}] \\ &= (f - E[\hat{f}])^2 + Var(\hat{f}) + \sigma^2 \\ &= Bias(\hat{f})^2 + Var(\hat{f}) + \sigma^2 \end{aligned} \tag{4}$$

Математическое ожидание квадрата ошибки прогноза можно представить в необходимом для нас виде:

$$E(\epsilon^2) = (E(X\hat{\beta}) - X\beta)^2 + E(X\hat{\beta} - E(X\hat{\beta}))^2 + \sigma^2 = Bias^2 + Variance + \sigma^2, \tag{5}$$

где $X$ – матрица наблюдений объясняющих переменных размерности $n \times p$; $\hat{\beta}$ – вектор полученных оценок коэффициентов при объясняющих переменных размерности $p \times 1$.

Таким образом, ошибка прогноза складывается из квадрата смещения (средняя ошибка по всем наборам данных), диспресии ошибки (насколько сильно ошибка отличается, если рассмотривать разные наборы данных) и неустранимой ошибки. В идеале, нам бы хотелось иметь и низкое смещение и низкую дисперсию, однако на практике часто возникает дилемма (bias–variance tradeoff) и приходится находить баланс. Как правило, при увеличении сложности модели (например, при увеличении количества свободных параметров), уменьшается смещение оценки, что не даёт нам упустить связь между признаками и объясняемой переменной, но увеличивается дисперсия (разброс) оценки, то есть повышается чувствительность к малым отклонениям и случайному шуму. В ситуации чрезмерной сложности модели можно говорить о таком понятии, как переобучение, излишнее подстраивание под конкретные данные, а не вычлелнение главных взаимосвязей. Хоть решение МНК и даёт несмещённую оценку регрессии, иногда стоит намеренно увеличить смещенность модели ради ее стабильности, т.е. ради уменьшения дисперсии. Это помогают сделать методы регуляризации, которые вводят большее смещение в решение регрессии, зато дисперсия становится меньше, что обеспечивает в итоге меньшую суммарную среднеквадратичную ошибку.

Рассморим более детельно технику подхода регуляризации и используемые для этого модели: Ridge, Lasso, ElasticNet. Основной идеей, объединяющей эти модели, явлется то, что они предотвращают сильный рост коэффициентов, возникший в результате переобучения или неккоретно поставленной задачи (например, когда между перемеными наблюдается мультиколинеарность) добавляя новый член в функцию, минимизирующую сумму квадратов остатков модели.

Модель Ridge использует обычный МНК, только со штрафным членом (L2–regularization).

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{i=1}^{p} \beta_j^2 \tag{6}$$

То есть новый член штрафует коэффициенты за их размер, причём параметр $\lambda$ отвечает за то, с какой силой мы будем стягивать коэффициенты. Видно, что при $\lambda$ = 0, полученные оценки будут совпадать с оценками МНК. А при $\lambda \to +\infty$ все коэффициенты будут приближаться к нулю. С увеличением $\lambda$ снижается сложность модели. Из этого следует, что выбор $\lambda$ имеет решающее значение. Описание методов подбора оптимального параметра будет описано далее.

Коэффициенты, полученные обычным МНК являются масштабно-эквивалентными. Если мы умножим каждый предиктор на некую константу $c$, то соответствующие им коэффициенты отмасштабируются умножением $\frac{1}{c}$. Однако, как при построении как Ridge, так и рассматриваемых далее моделей, необходимой процедурой является стандартизация, то есть центрирование и нормирование, объясняющих переменных:

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{7}$$

Так как мы штрафуем все коэффициенты за их абсолютную величину с одинаковым $\lambda$, необходимо предотвратить ситуацию, когда переменные с отличающимися масштабами будут по-разному влиять на штраф. Иначе может возникнуть ситуация, в которой зануляются существенные регрессоры.

Концепция модели Lasso (Least Absolute Shrinkage and Selection Operator) схожа с Ridge. Их отличие состоит в том, что член регуляризации Lasso (L1–regularization) есть сумма абсолютных значений коэффициентов модели, а не их квадратов, как в Ridge.

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{i=1}^{p}|\beta_j| \tag{8}$$

Впервые данный метод был описан в статье [1, Tibshirani, 1996]. Авторы показали, что такое ограничение на параметры может приводить к полному обнулению некоторых переменных, объясняющая сила которых несущественна. То есть метод может использоваться не только для регуляризации, но и для задачи выбора предикторов, так как на выход мы получаем разреженный вектор оценок коэффициентов. Если коэффициенту присваивается 0, то это фактически означает признание переменной нерелевантной и её удаление из нашей модели. Такой приём может обеспечить хорошую точность, ведь без существенного увеличения смещения, мы уменьшим дисперсию оценок. Как и у Ridge, коэффициент $\lambda$ в Lasso характеризует силу, с которой мы штрафуем коэффициенты. Если мы установим $\lambda$ раным нулю, получим обычные МНК оценки. При $\lambda \to +\infty$ все коэффициенты станут равны 0.

Lasso помагает увеличить интерпритируемость модели, определяя подмножество переменных с самым сильным воздействием, снижая перенасыщение лишними предикторами. Данное свойство модели помогает при оценивании регрессии с большим количеством объясняемых переменных, в том числе в условиях проклятия размерности, когда используется очень много предикторов, намного превосходящих число наблюдений. Такая проблема

является актуальной из-за возросших объемов доступных данных и требует решения, ведь высока вероятность, что с ростом признаков, некоторые из них будут коррелировать друг с другом.

Одним из отличий моделей является то, как они решают проблему мультиколлинеарности признаков. В данной ситуации Lasso произвольно выбирает одну переменную среди коррелированных, а коэффициенты остальных сводит к нулю. Ridge же ставит коррелированным предикторам похожие коэффициенты. Из-за этого считается, что Ridge будет работать лучше, когда большинство предикторов действительно влияют на объясняющую переменную. С другой стороны ожидается, что Lasso будет показывать себя лучше, когда имеется небольшое количество значимых параметров, а влияние большинства нерелевантно.

В статье [2, Belloni, Chernozhukov, 2013] были изучены свойства пост-модельных оценок Lasso. Авторы пришли к выводу, что если оценить классическим методом наименьших квадратов значения коэффициентов, которые до этого Lasso отобрала как релевантные, то можно получить менее смещённые относительно Lasso оценки.

Сравнивая две модели: Ridge и Lasso, мы приходим к выводу, что, несмотря на схожесть концепций, между ними всё же есть существенные различия. Как было отмечно выше, Lasso может занулять некоторые коэффициенты, полностью убирать из модели некоторые переменные. Ridge же, в свою очередь, всегда будет включать абсолютно все признаки, лишь заставляя быть коэффициентам при них ниже. То есть он не избавляется от несущественных признаков, как Lasso, а сводит к минимуму их влияние. Это очень важное отличие, которое означает неспособность модели Ridge выполнять такую процедуру, как выбор переменных, которая становится все более важной в современном анализе данных. Выясним, почему так происходит и чем обосновывается данный факт.

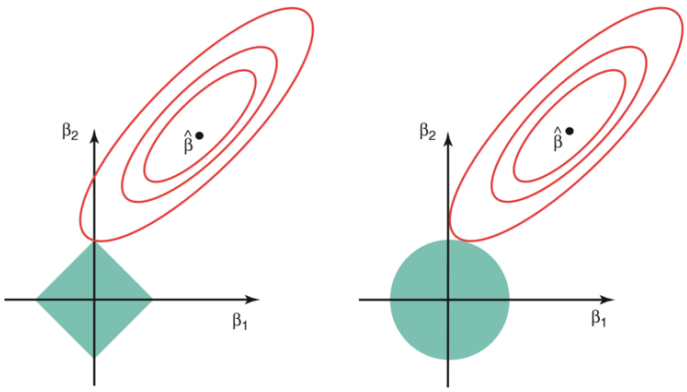


Рис. 1: Области ограничения Lasso (слева) и Ridge (справа)

На рис. 1 (источник: https://towardsdatascience.com) изображено главное отличие Lasso и Ridge — формы областей ограничения. Рассмотрен случай с двумя оцениваемыми параметрами. Эллиптическим контуром предсталена сумма квадратов ошибок, которую мы минимизируем, при ограничениях на параметры. Сами ограничения в данном случае можно записать в виде уравнений и получить: окружность ${\beta_1}^2 + {\beta_2}^2 \leq s$ для Ridge и ромб $|\beta_1| + |\beta_2| \leq s$ для Lasso, где $s$ — это некая константа, определённая для любого $\lambda$. При большом значении $s$ (при малом $\lambda$), области ограничения будут содержать центр эллипса. Это соотвествует случаю, когда оценки нагих моделей совпадают с оценками МНК. В остальных случаях, оценки коэффициентов Lasso и Ridge задаются точкой касания эллипса и области ограничения. Поскольку у Ridge она имеет круговую форму, это касание никогда не будет находится на оси, и поэтому оценки коэффициентов будут ненулевыми. Однако ограничение Lasso имеет углы на каждой из осей, и поэтому эллипс будет часто пересекать область ограничения именно на оси. Когда это произойдет, один из коэффициентов будет равен нулю. При рассмотрении случая с оцениванием параметров болшим, чем два, многие из оценок могут оказаться нулевыми.

В статье [3, Zou, Hastie, 2005] был предствлен ещё один способ регуляризации – Elastic net. По мнению авторов, в Lasso есть несколько существенных недостатков. Во-перых, в случае $p > n$ (число предикторов превышает число наблюдений) Lasso выбирает не более n переменных с ненулевыми коэффициентами из-за характера задачи выпуклой оптимизации. Во-вторых, как уже отмечалось выше, если существует группа переменных, среди которых корреляции между парами очень высоки, то Lasso выберет только одну переменную из группы, а коэффициенты при остальных занулит. Для обычных ситуаций, когда $n > p$, если есть высокие корреляции между предикторами, эмпирически наблюдалось, что в эффективности прогнозирования Ridge проявляет себя лучше, чем Lasso [1, Tibshirani, 1996]. Также Lasso ингода обвиняют в чрезмерной регуляризации (over-regularize) и компактности, а в следствие этого недостаточной предсказательности. Всю вышеупомяную критику, по-мнению авторов, учитывает и исправляет Elastic Net - комбинация двух моделей Ridge и Lasso:

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda_1 \sum_{i=1}^{p}\beta_j^2 + \lambda_2 \sum_{i=1}^{p}|\beta_j| \tag{9}$$

Если положить $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, то формулу (9) можно представить в следующем виде:

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \alpha \sum_{i=1}^{p}\beta_j^2 + (1-\alpha) \sum_{i=1}^{p}|\beta_j| \tag{10}$$

Исходя из формулы, видно, что при $\alpha = 1$, мы получаем модель Ridge, а при $\alpha = 0$ – Lasso. На практике встаёт вопрос о том, какое $\alpha$ является отпитмальным, с каким весом включать каждую из моделей, чтобы получить наилучший результат (как отмчалось выше, аналогичная проблема встречается для параметра $\lambda$ в Lasso и Ridge).

Параметр можно выбрать, используя инофрмационные критерии Акаике (AIC) и BIC (Bayes Information criterion). С помощью них выбор происходит очень быстро, но они опирается на правильную оценку степеней свободы, минимизируют лишь внутривыборочную дисперсию ошибки прогноза с поправкой на число параметров, а также выводится для больших выборок (асимптотические результаты). Они также имеют тенденцию давать неадекватный результат, когда число предикторов превосходит число наблюдений. Поэтому на практике наиблее используемым методом определния наилучшего значения параметра является кросс-валидация. В случае временных рядов можно использовать скользящую кросс-валидацию из-за отсутствия возможности перемешивания данных (как это требуется при K-fold cross-validation).

Elastic net объединяет основные преимущства двух моделей, комбинацией которых является. Так, Elastic net, как и Ridge, имеет способность формировать группу переменных, которые являются коррелированными и назначить им близкие коэффициенты. Как и Lasso, Elastic net может выступать не только как регуляризатор, но и выбирать множество наиболее релевантных предикторов, назначая нулевые коэффициенты перед незначимыми переменными. Эмпирические исследования показали, что Elastic net может превзойти Lasso, если в данных имеются высоко коррелированными предикторы. Это особо актуально для решения такой проблемы, как проклятие размерности. Когда у нас большое число регресоров, сильно превышающее число наблюдений, высока вероятность, что некоторые из них будут коррелировать друг с другом.

Существует еще один фундаментальный класс моделей, применяемых для регуляризации. Рассмотрим векторную авторегрессию (VAR) – это модель динамики, используемая для выявления взаимозависимости нескольких временных рядов между собой, в которой текущие значения этих рядов зависят от прошлых значений этих же временных рядов. Модель была предложена Кристофером Симсом как альтернатива системам одновременных уравнений и оказалась особенно полезной для описания динамического поведения экономических и финансовых временных рядов и для прогнозирования [4, Sims, 1980].

Все переменные в VAR входят в модель одинаковым образом: каждая переменная имеет

уравнение, основанное на ее собственных запаздывающих значениях и запаздывающих значениях других переменных модели и ошибки. Неограниченный (unrestricted) VAR включает в себя все переменные в каждом уравнении. Ограниченный (restricted) VAR может включать некоторые переменные в одном уравнении, а другие в другом. Поскольку большинство экономических рядов являются низкочастотными (ежемесячно, ежеквартально или ежегодно) имеется достаточно данных для точного прогноза с использованием больших неограниченных VAR. Проблема VAR-моделей заключается в резком росте количества параметров с увеличением количества анализируемых временных рядов и количества лагов. Следовательно, в таких приложениях пространство параметров VAR должно быть уменьшено.

Запишем формулу VARX (VAR с экзогенными переменными) для $t = 1, ..., T$ с $k$-мерным эндогеннным временным рядом $\{y_t\}_{t=1}^{T}$ и $m$-мерным экзогенными временным рядом $\{x_t\}_{t=1}^{T}$:

$$y_t = v + \sum_{l=1}^{p} \Phi^{(l)} y_{t-l} + \sum_{j=1}^{s} \beta^{(j)} x_{t-j} + u_t, \tag{11}$$

где $v$ – $k$-мерный вектор констант; $\Phi^{(l)}$ – матрица коэффициентов эндогенных переменных размера $k \times k$ при лаге $l = 1, ..., p$; $\beta^{(j)}$ – матрица коэффициентов экзогенных переменных размера $k \times m$ при лаге $j = 1, ..., s$; $u_t$ – независимый и одинаково распределенный $k$-мерный белый шум с нулевым вектором математических ожиданий.

Заметим, что при $\beta^{(j)} = 0$ для $j = 1, ..., s$, имеем особый случай VARX - VAR. Когда количество включенных предикторов существенно меньше, чем длина ряда, модель VARX может быть оценена многомерным МНК. При отсутствии регуляризации, такая модель требует оценки $k(1 + kp + ms)$ параметров регрессии. Как мы уже знаем, можно наложить некоторые ”штрафы”на $\Phi$ и $\beta$, которые помогут уменьшить простарность параметров модели. В статье [5, Hsu et al, 2008] авторы сравнивают VAR с использованием Lasso с традиционными методами выбора параметров в модели VAR, информационным критериям AIC и BIC. В результате было получено, что VAR-Lasso работает лучше информационных критериев с точки зрения прогнозирования, а также более эффективен в вычислительном отношении. Так же эта модификация модели VARX была описана в [6, Nicholson, 2017] и носила название название VARX-L, потому что в ней используется L1-регуляризация для построения больших векторных авторегрессий. Задача минимизации для VAR-Lasso предствляется в следующем виде:

$$\underset{v,\Phi,\beta}{\operatorname{argmin}} \sum_{t=1}^{T} ||y_t - v - \sum_{l=1}^{p} \Phi^{(l)} y_{t-l} - \sum_{j=1}^{s} \beta^{(j)} x_{t-j}||^2 + \lambda(\mathrm{P}_y(\Phi) + (\mathrm{P}_x(\beta)) \tag{12}$$

Поскольку мы используем единый $\lambda$ - параметр штрафа для всех уранений, требуется, чтобы до оценки все ряды должны быть в одном масштабе, то есть стандартизированы. Что касается штрафов параметров $\mathrm{P}_y(\Phi)$ и $(\mathrm{P}_x(\beta)$, они выбираются в зависимости от спецификации модели. Базовой является следующая спецификация Basic Lasso:

$$\mathrm{P}_y(\Phi) = ||\Phi||_1, \quad \mathrm{P}_x(\beta) = ||\beta||_1 \tag{13}$$

Также была предложена спецификация Lag Group, делающая матрицу какого-либо лага энодогенных переменных либо целиком отличной от нуля, либо состоящую из одних нулей. Однако включение большого количества групп существенно увеличивает время вычислений. К тому же, во многих случаях может оказаться нецелесообразным уделять равное внимание каждой перемнной конкретного лага. Неэффективно включать всю группу, если, например, только один коэффициент действительно ненулевой. Спецификация Sparse Group учитывает внутригрупповую разреженность через выпуклую комбинацию штрафа Basic Lasso и Lag Group.

Диагональ матрицы лага эндогенной переменной представляет собой регрессию на собственные лаги и с вероятностью большей, чем недиагональные коэффициенты, будет иметь ненулевые значения. Данный факт лежит в основе спецификации Own/Other. Это разграничение между собственными и другими лагами часто используется в макроэкономическом прогнозировании в предположении, что большую часть вариации ряда составляют его прошлые значения, поэтому они берутся с меньшим коэффициентом в штрафе.

В частотной статистике для того, чтобы оценивать различные параметры, существует множетсво подходов. К основным относят метод максимального правдоподобия, метод моментов, метод наименьших квадратов. Мы предполагаем, что неизвестный оцениваемый параметр является детерминированным, то есть константой и, основываясь на имеющихся данных, с помошью вышеперечисленных методов пытаемся получить его оценку. Полученая оценка является слуйчайной величиной, так как это функция от наблюдений. Мы можем узнать её распределение, оценивать её описательные характеристики (среднее, дисперсия и т.п), строить доверительные инервалы и проверять гипотезы.

Однако, у данного подхода есть некоторые недостатки. Например, он не даёт возможно-

сти внести некоторую дополнительную информацию, которой мы обладаем до получения данных. Так же мы не можем работать с истинным значением параметра, оценивать его вероятность попадания на какой-либо интервал или отрезок, проверять для него гипотезы, так как мы предполагаем что он является постоянной величиной. Все эти действия мы могли делать лишь с оценкой параметра.

Существует иной метод, который не разделяет параметр и его оценку, а просто учитывает все наши знания о параметре в априорное распределение. Он носит назавание байесовский, так как основной формулой, на которой он построен, является формула Байеса:

$$P(\beta \mid Data) = \frac{P(Data \mid \beta) * P(\beta)}{P(Data)}, \tag{14}$$

где Data – это данные, которые нам удалось собрать, а $\beta$ – это параметр, который мы оцениваем.

Плотность $P(\beta)$ включает в себя информацию, которую мы знаем до проведения анализа, или априорная (prior) плостность. Если у нас есть некоторые предположения о том, какие должны быть параметры модели, мы можем включить их в нашу модель. Например, при оценивании параметров линейной регрессии, мы можем знать какой наклон будет у прямой или в какой области она пересекает ось Ox. Это отличается от частотного подхода, который предполагает, что все, что нужно знать о параметрах, происходит из данных. Если у нас нет предварительных оценок, мы можем использовать неинформативные априорные распределния, например, нормальное.

Функция $P(Data \mid \beta)$ назыввется правдопободием (likekihood) и отражает вероятность появления наших данных, при условии, что нам известны параметры модели.

$P(Data)$ – вероятность появления данных (evidence), выполняет роль нормировки или фактора пропорциональности, который гарантирует, что апостериорная плотность интегрируется в единицу. Заметим, что

$$P(Data) = \int_{\beta} P(Data \mid \beta) * P(\beta) d\beta \tag{15}$$

Мы пытаемся найти апостериорную (posterior) плотность $P(\beta \mid Data)$, которая даст нам распределение возможных параметров модели на основе функции правдоподобия и априорной информации. Так как знаменатель дроби является интегралом (или в случае, когда параметры могут принимать только дискретных значения, суммой) числителя по всем возможным значениям параметра, это означает, что знаменатель является константой. Тогда,

получаем слелующее:

$$p(\beta \mid Data) \propto p(Data \mid \beta) * p(\beta) \tag{16}$$

С помощью формулы Байеса мы видим, как наша априорная плотность превращается в апостериорную. Мы начинаем с первоначальной оценки, нашего априорного распределения, и по мере того, как мы собираем больше наблюдений, наша модель становится менее ошибочной. Формулу можно применять итерационно, после поступления каждой новой порции данных формула пересчитывается.

Таким образом, целью байесовского подхода является не нахождение единственного «наилучшего» значения параметров модели, а определение апостериорного распределения параметров модели, используя которое мы можем измерять вероятность, строить интервальные оценки (только теперь байесовские), проверять гипотезы. Так как мы точно не знаем, какая из гипотез подтвердится, поэтому проверяем сразу все гипотезы, сумма вероятностей которых равна 1, а не одну гипотезу, как в частотном подходе. Если какая-то гипотеза становится вероятнее, вероятность других снижается. Однако такой подход в связи со сложными расчётами, часто интегральными, является вычислительно затратным для человека, именно поэтому интерес к байесовскому подоходу вырос с развитием компьютеров.

Расмотрим баейсовскую (вероятностную) интерпритацию регуляризации, метода, используемого для предотвращения переобученияи или некорректно поставленых задач. Раньше мы могли оценить параметры линейной регресии методом максимального правдоподобия (MLE). Однако байесовский анализ основан на максимизации апостериорной плотности (MAP). Максимизируем (14), предполагая, что $P(Data)$ постоянно относительно параметров $\beta$:

$$\begin{aligned}
\hat{\beta}_{\text{MAP}} &= \arg\max_{\beta} P(\beta|Data) \\
&= \arg\max_{\beta} P(Data|\beta)P(\beta) \\
&= \arg\max_{\beta} \log(P(Data|\beta)P(\beta)) \\
&= \arg\max_{\beta} \log P(Data|\beta) + \log P(\beta)
\end{aligned} \tag{17}$$

Видно, что метод MAP максимизирует логарифм правдоподобия (как и MLE) плюс логарифм априорного распределения. Покажем, что второе слагаемое играет роль регуляриза-

тора.

Запишем функцию правдоподобия для обычной линейной регресии, предполагая, что $Data$ – это набор наблюдений $(y_1, ..., y_N)$, которые распределены остостельно линии регрессии с нормальной ошибкой $\epsilon \sim N(0, \sigma^2)$

$$\begin{aligned} P(Data|\beta) &= \prod_{i=1}^{n} P(y_i|\beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2}} \end{aligned} \tag{18}$$

До наблюдения выборки мы ничего не знаем о коэффициентах $\beta$ нашей модели, поэтому априорно будем считать, что коэффициенты линейной регрессии будут нормально распределены $\beta \sim N(0, \tau)$, тогда

$$P(\beta) = \prod_{j=0}^{p} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{\beta_j^2}{2\tau^2}} \tag{19}$$

Подставляем полученное априорное распределение и функцию правдоподобия в (17)

$$\begin{aligned} &\arg\max_{\beta} \Big[ -\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{\beta_j^2}{2\tau^2} \Big] \\ &= \arg\min_{\beta} \frac{1}{2\sigma^2} \Big[ \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2 + \frac{\sigma^2}{\tau^2} \sum_{j=0}^{p} \beta_j^2 \Big] \\ &= \arg\min_{\beta} \Big[ \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} \beta_j^2 \Big] \end{aligned} \tag{20}$$

Таким образом, мы получили формулу, используемую в Ridge регрессии, где перовое слагаемое - среднеквадратическая ошибка, а второе - регуляризатор с параметром штрафа $\lambda = \frac{\sigma^2}{\tau^2}$.

Теперь возьмём в качестве априорного распределения коэффициентов взяли бы не Гауссово распределение, а Лапласа

$$Laplace(\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \tag{21}$$

Предполагая, что среднее $\mu = 0$, проделываем ту же операцию, подставляем априорное распределение и функцию правдопободия в (17)

$$\begin{aligned}
&\arg\max_{\beta}\Big[-\sum_{i=1}^{n}\frac{(y_i-(\beta_0+\beta_1x_{i,1}+...+\beta_px_{i,p}))^2}{2\sigma^2}-\sum_{j=0}^{p}\frac{|\beta_j|}{2b}\Big]\\
&=\arg\min_{\beta}\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+...+\beta_px_{i,p}))^2+\frac{\sigma^2}{b}\sum_{j=0}^{p}|\beta_j|\Big] \qquad (22)\\
&=\arg\min_{\beta}\Big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+...+\beta_px_{i,p}))^2+\lambda\sum_{j=0}^{p}|\beta_j|\Big]
\end{aligned}$$

Видно, что полученное уравнение имеет тот же вид, что и формула, используемая в Lasso регресии с параметром штрафа $\lambda = \frac{\sigma^2}{b}$, где $b > 0$ – параметр машстаба в нашем априорном распределении Лапласа.

Как уже было отмечано выше, модель VAR имеет одну существенную проблему в реализции, которая заключается в резком росте количества параметров с увеличением количества анализируемых временных рядов и количества лагов. Одним из методов сокращения коэффициентов VAR является SSVS (Stochastic search variable selection) [7, Koop G. M, 2013]. Данный метод помагает в ситуации, когда имеется большое количество объясняющих переменных, а нам необходимо определить, какие из них могут быть важными. Чтобы это определить, априорное распределение каждого коэффициента регрессора определяется как взвешенное из нормальных с нулевыми средними, но с разными дисперсиями. Одно из них имеет очень малую дисперсию (что говорит о том, что коэффициент фактически равен нулю и данный фактор может быть исключен из модели), а другой имеет большую дисперсию (т.е. коэффициент, скорее всего, отличается от нуля, и, следовательно, переменная должна сохраняться в модели). Таким образом, для каждого коэффициента $\beta j$, априорное распределение определяется как:

$$P(\beta_j \mid \gamma_j) = (1-\gamma_j)N(0,\tau_{0j}^2)+\gamma_jN(0,\tau_{1j}^2), \qquad (23)$$

где $\gamma_j$ – это дамми-переменная, то есть принимает значения либо 0, либо 1, тем самым определеляя априорное распределение $\beta_j$. Причём, $\gamma_j$ рассматривается как неизвестный параметр и оценивается на основе данных. То есть, SSVS использует иерархическое априорное распределение, то есть выраженное через параметры, которые, в свою очередь, имеют собственное априорное распределение. SSVS предполагает, что каждый элемент $\gamma_j$ имеет априорное

распределение формы Бернулли:

$$P(\gamma_j = 1) = q_j,$$
$$P(\gamma_j = 1) = 1 - q_j \tag{24}$$

По умолчанию значение $qj$ выбирается равным 0.5 для всех $j$, то есть подразумевется, что каждый коэффициент априори с равной вероятностью будет включен как исключенный. Как было замечено ранее, диспресии берутся сильно отличными. Так, для каждого параметра в качестве $\tau_{0j}^2$ берётся дисперсия коэффициента парной регрессии на данную переменную, умноженная на 0.1, а в качестве $\tau_{1j}^2$ - дисперсия коэффициента парной регрессии на данную переменную, умноженная на 10.

В задачах выбора переменных рассматриваемый список моделей соответствует возможным подмножествам набора предикторов. Ясно, что число моделей быстро становится огромным по мере увеличения размера, поэтому существует необходимость в эффективных методах поиска моделей с высокой апостериорной вероятностью, а также при оценке вероятностей апостериорных моделей и апостериорных распределений для коэффициентов в каждой модели. В регрессионном анализе, при $p$ доступных предикторов, число потенциальных моедлей равно $2^p$. Чтобы не рассчитывать апостериорную вероятность для каждой из этих моделей, SSVS использует метод сэмплирования Гиббса, который быстро находит наиболее перпективные и высоковероятностные модели.

# Список литературы


[1] Tibshirani R. Regression shrinkage and selection via the lasso //Journal of the Royal Statistical Society: Series B (Methodological). – 1996. – Т. 58. – №. 1. – С. 267-288.

[2] Belloni A. et al. Least squares after model selection in high-dimensional sparse models //Bernoulli. – 2013. – Т. 19. – №. 2. – С. 521-547.

[3] Zou H., Hastie T. Regularization and variable selection via the elastic net //Journal of the royal statistical society: series B (statistical methodology). – 2005. – Т. 67. – №. 2. – С. 301-320.

[4] Sims C. A. Macroeconomics and reality //Econometrica: journal of the Econometric Society. – 1980. – С. 1-48.

[5] Hsu N. J., Hung H. L., Chang Y. M. Subset selection for vector autoregressive processes using lasso //Computational Statistics and Data Analysis. – 2008. – Т. 52. – №. 7. – С. 3645-3657.

[6] Nicholson W. B., Matteson D. S., Bien J. VARX-L: Structured regularization for large vector autoregressions with exogenous variables //International Journal of Forecasting. – 2017. – Т. 33. – №. 3. – С. 627-651.

[7] Koop G. M. Forecasting with medium and large Bayesian VARs //Journal of Applied Econometrics. – 2013. – Т. 28. – №. 2. – С. 177-203.