POC II
Melhorando a Estimativa do PREDEP: Uma Alternativa Baseada em Regressão

Francisco Neves Tinoco Junior
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
francisconeves@dcc.ufmg.br
Orientado

Flavio Vinicius Diniz de Figueiredo
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
flaviovdf@dcc.ufmg.br
Orientador

Renato Martins Assunção
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
assuncao@dcc.ufmg.br
Colaborador do trabalho

Vinicius Fernandes dos Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
viniciussantos@dcc.ufmg.br
Colaborador do trabalho

***Abstract*—A métrica PREDEP foi recentemente proposta como uma ferramenta para quantificar a capacidade preditiva de uma variável aleatória X sobre outra variável Y, fornecendo uma medida baseada na redução da dispersão condicional. Contudo, os métodos originais para estimar o PREDEP apresentaram inconsistências, especialmente na estimativa da componente condicional $S_{Y|X}$. Este trabalho propõe uma abordagem alternativa para essa estimativa, baseada em regressão, que visa superar as limitações dos métodos baseados em particionamento do espaço. A eficácia do método proposto foi avaliada em diversos conjuntos de dados sintéticos, nos quais o valor teórico de $S_{Y|X}$ é conhecido. Os resultados mostram que a abordagem baseada em regressão apresenta maior estabilidade e precisão, especialmente em cenários com relações complexas entre X e Y ou quando se tem poucos dados disponiveis. Concluímos que o método proposto é uma alternativa robusta e prática para a estimativa do PREDEP.**

I. Introdução

A métrica PREDEP, estabelecida em [1], busca quantificar a redução da dispersão de uma variável aleatória dado o conhecimento de outra. Essa métrica funciona como uma medida da capacidade preditiva de uma variável com base em outra, fornecendo uma abordagem sistemática para avaliar essa relação.

No entanto, as formas estabelecidas no trabalho original para estimar a métrica PREDEP apresentam limitações, especialmente na estimativa de $S_{Y|X}$, conforme definido na Equação 2. Essas limitações geram inconsistências na aplicação prática, comprometendo a robustez dos resultados obtidos com o uso da métrica.

O presente trabalho tem como objetivo propor uma abordagem mais estável para a estimativa da condicional $S_{Y|X}$, utilizando um método baseado em regressão de X para Y. A proposta é explorar o uso da regressão como uma alternativa às metodologias baseadas em particionamento do espaço, buscando maior precisão e estabilidade na estimativa.

Para avaliar a eficácia do método proposto, serão realizados testes em diversos conjuntos de dados sintéticos, nos quais o valor teórico esperado da condicional é previamente conhecido. Esses testes permitirão comparar o desempenho do método baseado em regressão com outras formas de estimar a condicional, fundamentadas no particionamento do espaço.

II. Métrica

A métrica PREDEP é inspirada na medida proposta por Goodman e Kruskal, mas adaptada para variáveis contínuas. Neste trabalho, utilizamos a versão simplificada da métrica, considerando apenas duas variáveis aleatórias, X e Y. O cálculo do PREDEP depende de duas estimativas principais: uma relacionada à dispersão marginal (Equação 1) e outra à dispersão condicionada (Equação 2).

$$S_Y = \mathbb{E}\left[f_Y(Y)\right] = \int f_Y^2(y)\, dy, \tag{1}$$

$$\begin{aligned} S_{Y|X} &= \int \left[\int f^2_{Y|X=x}(Y)\, dy \right] f_X(x)\, dx \\ &= \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[f_{Y|X}(Y|X)\right]\right]. \end{aligned} \tag{2}$$

Os valores obtidos nas Equações 1 e 2 não possuem uma interpretação direta, pois dependem de fatores como a escala das variáveis, estando ambos os valores definidos no intervalo $(0, \infty)$. Para superar essa limitação, o PREDEP ($\alpha_{Y|X}$) é definido como uma métrica que combina essas estimativas em

um valor normalizado e interpretável, conforme apresentado na Equação 3.

$$\alpha_{Y|X} = \frac{S_{Y|X} - S_Y}{S_{Y|X}}. \tag{3}$$

Essa definição permite avaliar a capacidade preditiva de X sobre Y de forma intuitiva, com valores normalizados entre 0 e 1.

III. Metodo Proposto

Intuição: Assuma que a relação entre as variáveis X e Y pode ser descrita como $Y = f(X) + \epsilon$, onde o termo de erro ϵ é completamente independente de X. Assim, condicionar Y por X equivale a conhecer o valor de $f(x)$ somado ao erro ϵ. Isso implica que a única fonte de incerteza na distribuição condicional $Y|X$ advém de ϵ.

Formalização: Considere duas variáveis X e Y relacionadas pela equação $Y = g(X) + \epsilon$, onde ϵ é uma variável aleatória independente de X. A seguir, demonstramos que a variância condicional $S_{Y|X}$ depende exclusivamente de ϵ.

1. A distribuição condicional de Y dado $X = x_i$ é dada por:

$$Y|X = x_i = g(x_i) + \epsilon,$$

onde $g(x_i) \in \mathbb{R}$.

2. Utilizando a definição de $S_{Y|X}$, temos:

$$S_{Y|X} = \int f_X(x) \int f_{Y|X=x_i}(y)^2 \, dy \, dx.$$

Substituindo $Y|X = g(X) + \epsilon$, obtemos:

$$S_{Y|X} = \int f_X(x) \int f_{g(x_i)+\epsilon}(y)^2 \, dy \, dx.$$

3. Para duas variáveis aleatórias A e $B = A + c$, com $c \in \mathbb{R}$, é sabido que as funções de densidade $f_A(x)$ e $f_B(x)$ estão relacionadas por:

$$f_B(x + c) = f_A(x).$$

Assim, para $g(x_i) + \epsilon$, temos:

$$f_{g(x_i)+\epsilon}(y) = f_\epsilon(y - g(x_i)).$$

4. Substituindo essa relação na expressão de $S_{Y|X}$, obtemos:

$$S_{Y|X} = \int f_X(x) \int f_\epsilon(y - g(x_i))^2 \, dy \, dx.$$

5. Utilizando a propriedade das integrais:

$$\int_a^b h(x) \, dx = \int_{a-c}^{b-c} h(x + c) \, dx,$$

para $c \in \mathbb{R}$ e limites $a = -\infty$, $b = \infty$, segue que:

$$\int h(x) \, dx = \int h(x + c) \, dx.$$

Aplicando essa propriedade:

$$S_{Y|X} = \int f_X(x) \int f_\epsilon(y)^2 \, dy \, dx.$$

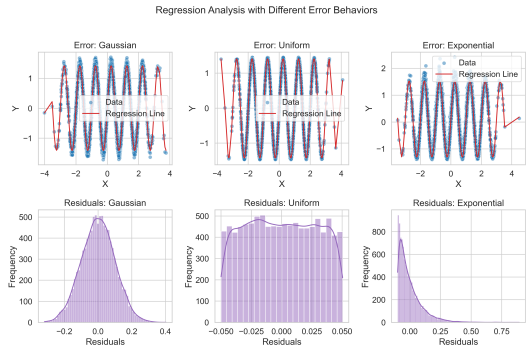

Fig. 1. Exemplo de ϵ adquiridos por regressão

6. Definindo $S_\epsilon = \int f_\epsilon(y)^2 \, dy$, temos:

$$S_{Y|X} = \int f_X(x) S_\epsilon \, dx.$$

7. Como ϵ é independente de X, $\int f_X(x) \, dx = 1$. Assim:

$$S_{Y|X} = S_\epsilon.$$

Concluímos, portanto, que $S_{Y|X}$ dependen exclusivamente de ϵ.

Utilizando o método descrito anteriormente, observa-se que, ao obter o termo de erro ϵ, é possível estimar diretamente o valor da condicional $Y|X$ sem a necessidade de métodos complexos de particionamento. Além disso, essa abordagem permite utilizar todos os dados disponíveis para realizar a estimativa, maximizando a eficiência do processo.

A obtenção do termo ϵ requer apenas a escolha de uma regressão adequada para modelar a relação entre X e Y. Uma vez definida a regressão, o erro restante pode ser estimado diretamente, representando o termo ϵ. A imagem 1 demonstra o processo de obtenção de ϵ.

É importante ressaltar, no entanto, que a escolha da regressão pode introduzir desafios, como problemas de overfitting ou underfitting. Esses problemas podem comprometer a precisão da estimativa da condicional, especialmente se o modelo de regressão escolhido não capturar adequadamente a relação entre X e Y. Contudo, dada a ampla gama de métodos de regressão disponíveis e os avanços significativos na área, é bastante plausível e prático construir regressões robustas que minimizem esses problemas e forneçam estimativas confiáveis.

IV. Metodologia

Serão testadas quatro abordagens distintas para estimar a distribuição condicional $Y|X$:

1) Equal: Este método utiliza um particionamento igualitário de todo o espaço amostral. Isso significa que cada partição conterá o mesmo número de pontos, independentemente das características do espaço, só da distribuição dos dados.
2) Agglomerative: Baseado no particionamento hierárquico aglomerativo, este método segue a abordagem apresentada no artigo original. Nele, as partições são formadas

por meio de um processo iterativo que agrupa pontos com base em critérios de similaridade.

3) POC I: Trata-se de um particionamento desenvolvido na primeira etapa. Ele utiliza uma versão modificada do particionamento hierárquico, onde a métrica de junção é baseada diretamente na estimativa da condicional. Essa abordagem busca otimizar o particionamento para refletir melhor a dependência entre X e Y
4) Linear: Este método emprega uma regressão linear sobre $g(x)$. A escolha da regressão linear foi motivada por sua simplicidade, evitando problemas de overfitting. No entanto, na prática, esta abordagem apresenta limitações, pois é altamente improvável que sejam conhecidos, a priori, os fatores específicos de X necessários para prever Y. Apesar disso, ela serve como um ponto de partida e referência por sua simplicidade metodológica.

Para avaliar a qualidade dos métodos de estimação da condicional, foi definida uma série de conjuntos de dados sintéticos nos quais o valor teórico da condicional $Y|X$ é conhecido. Esses dados foram gerados com base na relação $Y = g(X) + \epsilon$, onde $g(X)$ representa a função determinística e ϵ é um termo de erro aditivo.

As funções $g(X)$ utilizadas para gerar os dados são as seguintes:

- Linear: $Y = X + \epsilon$,
- Quadrática: $Y = X^2 + \epsilon$,
- Sinusoidal: $Y = \sin(X) + \epsilon$.

O termo de erro ϵ foi modelado utilizando diferentes distribuições, a saber:

- Distribuição Gaussiana,
- Distribuição Uniforme,
- Distribuição Exponencial.

Adicionalmente, outros fatores foram variados para testar o desempenho dos métodos de estimação:

- Quantidade de pontos: O tamanho dos conjuntos de dados foi ajustado para avaliar como a disponibilidade de dados influencia os resultados.
- Escala dos erros: A magnitude do termo ϵ foi escalada para analisar o impacto de diferentes níveis de ruído nas estimativas.

Esses experimentos permitem uma análise abrangente da capacidade dos métodos de estimação em diferentes cenários, com funções $g(X)$ variadas, distribuições de erro distintas e condições controladas de tamanho amostral e intensidade do ruído.

V. Resultados

Os resultados obtidos para os diferentes métodos de estimação apresentam comportamentos distintos, dependendo da função utilizada, do número de pontos, figura 2, e da magnitude do erro, figura 3.

Em relação ao número de pontos, os resultados foram os esperados para as funções linear e quadrática. À medida que o número de pontos aumenta, os métodos tendem a apresentar resultados similares. Contudo, para conjuntos de dados com poucos pontos, o método baseado em regressão mostrou-se superior, possuindo uma vantagem significativa sobre os métodos de particionamento.

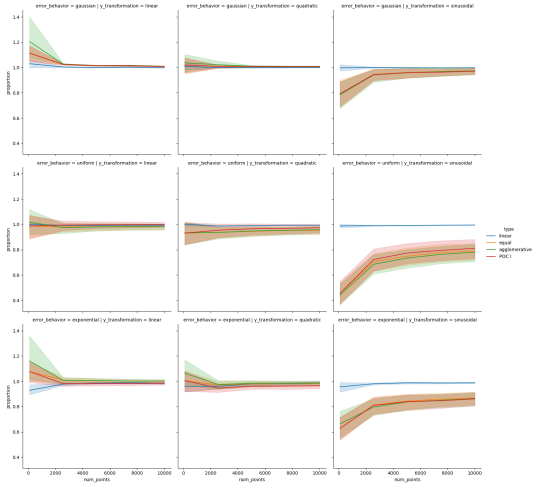

Fig. 2. Comportamento dos métodos de estimação em relação ao número de pontos.

A maior diferença foi observada na função sinusoidal. Essa função parece ser a mais sensível ao tipo de erro presente nos dados com os métodos de particionamento apresentaram desempenho bastante ruim quando o erro seguia uma distribuição uniforme, enquanto o método baseado em regressão não foi significativamente afetado.

De maneira geral, a maior complexidade do comportamento da função sinusoidal parece impactar negativamente os métodos de particionamento. Além disso, é importante destacar que, em todos os casos, a variação do método de regressão foi consistentemente menor do que a dos métodos de particionamento, indicando maior estabilidade.

O comportamento em relação à magnitude do erro apresentou maior complexidade. No caso mais simples, com erro gaussiano e relação linear, a quantidade de erro presente foi irrelevante para todos os métodos. Entretanto, para os outros cenários, um aumento na magnitude do erro resultou em estimativas proporcionalmente maiores para a condicional. Em relação à função sinusoidal, o comportamento foi semelhante ao observado para o número de pontos, com desempenho ruim dos métodos de particionamento.

Adicionalmente, foi possível observar que, no caso do método baseado em regressão, o erro com distribuição exponencial parece ser ligeiramente mais instável de estimar do que os outros tipos de erro. Contudo, essa variação não foi substancial. Por fim, o método de regressão mostrou-se novamente superior aos métodos de particionamento.

VI. Conclusão

Os experimentos realizados demonstraram a eficácia e as limitações dos diferentes métodos de estimação da condicional $Y|X$ em cenários variados. Os resultados evidenciam que a escolha do método de estimação é altamente dependente das

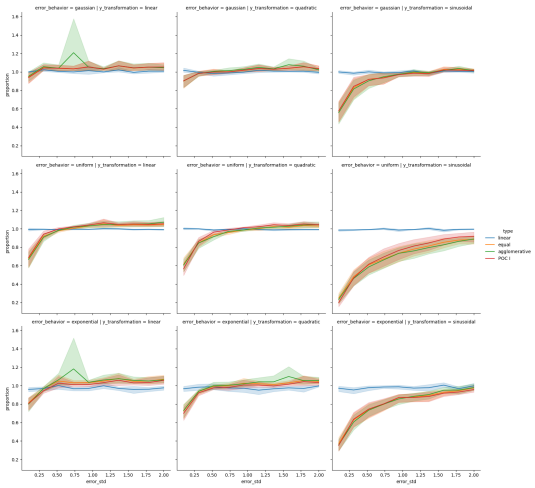

Fig. 3. Comportamento dos métodos de estimação em relação à magnitude do erro.

características dos dados, como a relação funcional entre X e Y, o tipo de distribuição do erro ϵ, o número de pontos disponíveis e a magnitude do erro.

O método baseado em regressão mostrou-se mais estável e robusto em todos os cenários, especialmente quando o número de pontos era pequeno ou a complexidade da relação entre X e Y era alta, como no caso da função sinusoidal. Além disso, a variação de desempenho desse método foi consistentemente menor em comparação aos métodos de particionamento. Contudo, a simplicidade intrínseca da regressão linear limita sua aplicabilidade em cenários mais realistas, onde os fatores necessários para prever Y podem não ser conhecidos.

REFERENCES

[1] Renato Assunção, Flávio Figueiredo, Francisco N. Tinoco Júnior, Léo M. de Sá-Freire, and Fábio Silva. An interpretable measure for quantifying predictive dependence between continuous random variables – extended version, 2025.