

## 6 Conclusão

Neste trabalho, foram fornecidos métodos eficientes para a aplicação da regressão por mínimos quadrados parciais em diversas áreas além da originária, a quimiometria. Para tanto, foram incorporados ao algoritmo de regressão PLS1 paralelismo e métodos diretos para o caso PLS2, procurando o tratamento de grandes massas de dados. Para o aumento do poder preditivo, foi acoplada também não-linearidade de forma eficiente via funções de núcleo.

### 6.1 Principais resultados

Num primeiro passo, a arquitetura de modelos competidores MEM [33, 32] demonstra a pertinência da regressão PLS pela simplicidade e qualidade apresentadas. Além disto, o uso da transformada Haar de wavelets, por fornecer uma representação do perfil do padrão, permitiu uma melhor partição do espaço de amostras.

Para o tratamento de grandes conjuntos de dados, foram desenvolvidos dois algoritmos. O primeiro é o PPLS [37, 41, 40], uma versão paralela para o caso PLS1 de uma variável dependente. Dentre suas principais características destacamos:

1. fácil implementação com baixo custo;
2. bom desempenho, com um aproveitamento de 75% de cada processador ao serem usadas 4 máquinas.

O segundo algoritmo apresentado é o DPLS [37, 41, 36], uma versão aproximada para o caso PLS2 de mais de uma variável dependente, possuindo os seguintes atributos:

1. curva PRESS semelhante à do método exato, mostrando qualidade de predição competitiva;

2. ganho acima de 40% no desempenho computacional.

O aumento significativo de desempenho conseguido faz dos dois algoritmos candidatos para o tratamento de grandes conjuntos de dados. DPLS fornece um método direto para a construção de modelos preditivos, enquanto PPLS explora a natureza distribuída da entrada, para paralelizar de forma eficiente o cálculo dos fatores PLS.

O uso de funções de núcleo permitiu a incorporação de não-linearidade. Isto foi realizado de forma eficiente, pois características não lineares das variáveis independentes foram adicionadas à entrada de forma implícita, não aumentando a complexidade computacional.

A primeira contribuição é LPLS, uma versão do algoritmo PLS1, que em relação às versões existentes [46] apresenta

1. maior eficiência computacional, pois explora a particularidade de apenas uma variável dependente;
2. um algoritmo iterativo para a predição, não apresentando problemas de estabilidade numérica;
3. representação clássica do PLS1, com os escores não normalizados.

Seguindo a mesma linha da versão linear, é fornecido KDPLS, uma versão *kernel* para o algoritmo PLS2 aproximado DPLS. Nos experimentos, as mesmas características foram observadas: qualidade competitiva e melhor desempenho computacional.

A terceira contribuição utilizando funções de núcleo, é o MKPLS, que estende os algoritmos *kernel* existentes, empregando mais de um núcleo para a construção do modelo não-linear. Isto é realizado empregando uma função de núcleo para os primeiros fatores, e uma segunda para os demais. Dentre os principais atributos, são destacados:

1. modelagem mais compacta, pois possui mesma taxa de aprendizado do PLS para os primeiros fatores;
2. poder preditivo competitivo quando comparado com o LPLS;
3. demonstra, em certos casos, que a modelagem linear para os primeiros fatores fornece melhor predição para os demais não-lineares;
4. possui desempenho no mínimo equivalente ao de outros modelos.

Com isto, o MKPLS pode ser considerado uma alternativa aos algoritmos para regressão PLS baseados em funções de núcleo.

## 6.2

### Trabalhos Futuros

No decorrer do desenvolvimento dos algoritmos aqui apresentados, alguns pontos levantados merecem maior atenção, em trabalho ainda a ser realizado.

Com relação ao PPLS, uma implementação em arquitetura de memória compartilhada deve fornecer melhor desempenho.

Quanto ao DPLS, a mesma abordagem aproximada, de decomposição em blocos, pode ser usada para a análise em componentes principais. Seria interessante analisar a estabilidade dos resultados, juntamente com o ganho em desempenho computacional.

Para o LPLS, a implementação de uma versão PLS2 resolvendo o caso geral da equação 5-11, fornecendo uma versão não-linear do algoritmo clássico, para comparação com [46]. Além disto, uma heurística para a escolha do núcleo com seus ajustes, forneceria um método eficiente para a calibração do modelo.

A elaboração da versão multi-núcleos MKPLS levou ao aumento do número de parâmetros no treinamento, acarretando nas seguintes considerações:

1. observar desempenho com dois núcleos não aditivos, ou seja, sem que  $K_2 = K_1 + K$ ;
2. explorar possível benefício ao se usar mais de dois núcleos, já que isto pode ser realizado com os algoritmos para desconto apresentados;
3. medir o desempenho em conjuntos provenientes de outras áreas além da química [24];
4. a estratégia multi-núcleos desenvolvida para o MKPLS, deve se aplicar com o mesmo sucesso aos outros métodos de regressão parcial, como o PCR [45].

Dado que o objetivo é mostrar a viabilidade da estratégia multi-núcleos, não foi realizada nenhuma filtragem adicional nos dados para remoção de possíveis dados espúrios. Além disto, dada a origem do PLS, o conjunto NIR possui dados com população reduzida, sendo de no máximo 282 amostras. A aplicação de técnicas de filtragem, além do uso de conjuntos de maior tamanho, devem fornecer predições melhores e mais estáveis.