Pontifícia Universidade Católica
do Rio de Janeiro

João Marco Braga da Cunha

Estimação de Redes Neurais Artificiais Através do Método Generalizado dos Momentos

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC–Rio como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica

Orientador: Prof. Alexandre Street de Aguiar

Rio de Janeiro
Dezembro de 2015

Pontifícia Universidade Católica
do Rio de Janeiro

João Marco Braga da Cunha

Estimação de Redes Neurais Artificiais Através do Método Generalizado dos Momentos

Tese apresentada ao Programa de Pós–graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Alexandre Street de Aguiar
Orientador
Departamento de Engenharia Elétrica — PUC–Rio

Prof. Marley Maria Bernardes Rebuzzi Vellasco
Departamento de Engenharia Elétrica — PUC-Rio

Prof. Cristiano Augusto Coelho Fernandes
Departamento de Engenharia Elétrica — PUC-Rio

Prof. Carlos Eduardo Cardoso Galhardo
IMETRO

Prof. Jessica Quintanilha Kubrusly
UFF

Prof. Renato Galvão Flôres Junior
FGV–RJ

Prof. Márcio da Silveira Carvalho
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 9 de Dezembro de 2015

João Marco Braga da Cunha

Graduou–se em Economia na PUC–RIO, em 2005. Concluiu os mestrados em Economia na FGV–EPGE, em 2008, e em Engenharia Elétrica na PUC–RIO, em 2009. Atualmente, trabalha como Economista na Área de Gestão de Riscos do BNDES.

Ficha Catalográfica

Cunha, João Marco Braga da

Estimação de redes neurais artificiais através do método generalizado dos momentos / João Marco Braga da Cunha; orientador: Alexandre Street de Aguiar — 2015.

90 f: il. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2015.

Inclui bibliografia.

1. Engenharia elétrica – Teses. 2. Redes Neurais Artificiais;. 3. *Perceptron* de Mútiplas Camadas;. 4. Sobreajuste;. 5. Não Linearidade Negligenciada;. 6. *Quasi* Máxima Verossimilhança;. 7. Método Generalizado dos Momentos;. 8. Teste J;. 9. Otimização Global.. I. Aguiar, Alexandre Street de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD:621.3

Dedico esta tese aos meus pais, especialmente ao meu pai, Marco Antonio, que acompanhou com muito entusiasmo o meu desenvolvimento acadêmico até os últimos momentos de sua vida.

Agradecimentos

Àqueles que me ajudaram, direta ou indiretamente, na proporção de suas ajudas, com uma menção especial à minha mãe, Glena Luiza, pelas inúmeras revisões deste e de outros textos produzidos ao longo do doutorado.

Resumo

Cunha, João Marco Braga da; Aguiar, Alexandre Street de. **Estimação de Redes Neurais Artificiais Através do Método Generalizado dos Momentos**. Rio de Janeiro, 2015. 90p. Tese de Doutorado — Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

As Redes Neurais Artificiais (RNAs) começaram a ser desenvolvidas nos anos 1940. Porém, foi a partir dos anos 1980, com a popularização e o aumento de capacidade dos computadores, que as RNAs passaram a ter grande relevância. Também nos anos 1980, houve dois outros acontecimentos acadêmicos relacionados ao presente trabalho: (i) um grande crescimento do interesse de econometristas por modelos não lineares, que culminou nas abordagens econométricas para RNAs, no final desta década; e (ii) a introdução do Método Generalizado dos Momentos (MGM) para estimação de parâmetros, em 1982. Nas abordagens econométricas de RNAs, sempre predominou a estimação por *Quasi* Máxima Verossimilhança (QMV). Apesar de possuir boas propriedades assintóticas, a QMV é muito suscetível a um problema nas estimações em amostra finita, conhecido como sobreajuste. O presente trabalho estende o estado da arte em abordagens econométricas de RNAs, apresentando uma proposta alternativa à estimação por QMV que preserva as suas boas propriedades assintóticas e é menos suscetível ao sobreajuste. A proposta utiliza a estimação pelo MGM. Como subproduto, a estimação pelo MGM possibilita a utilização do chamado Teste J para verificar a existência de não linearidade negligenciada. Os estudos de Monte Carlo realizados indicaram que as estimações pelo MGM são mais precisas que as geradas pela QMV em situações com alto ruído, especialmente em pequenas amostras. Este resultado é compatível com a hipótese de que o MGM é menos suscetível ao sobreajuste. Experimentos de previsão de taxas de câmbio reforçaram estes resultados. Um segundo estudo de Monte Carlo apontou boas propriedades em amostra finita para o Teste J aplicado à não linearidade negligenciada, comparado a um teste de referência amplamente conhecido e utilizado. No geral, os resultados apontaram que a estimação pelo MGM é uma alternativa recomendável, em especial no caso de dados com alto nível de ruído.

Palavras–chave

Redes Neurais Artificiais; *Perceptron* de Mútiplas Camadas; Sobreajuste; Não Linearidade Negligenciada; *Quasi* Máxima Verossimilhança; Método Generalizado dos Momentos; Teste J; Otimização Global.

Abstract

Cunha, João Marco Braga da; Aguiar, Alexandre Street de (Advisor). **Estimating Artificial Neural Networks with Generalized Method of Moments**. Rio de Janeiro, 2015. 90p. PhD Thesis — Departmento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Artificial Neural Networks (ANN) started being developed in the decade of 1940. However, it was during the 1980's that the ANNs became relevant, pushed by the popularization and increasing power of computers. Also in the 1980's, there were two other two other academic events closely related to the present work: (i) a large increase of interest in nonlinear models from econometricians, culminating in the econometric approaches for ANN by the end of that decade; and (ii) the introduction of the Generalized Method of Moments (GMM) for parameter estimation in 1982. In econometric approaches for ANNs, the estimation by *Quasi* Maximum Likelihood (QML) always prevailed. Despite its good asymptotic properties, QML is very prone to an issue in finite sample estimations, known as overfitting. This thesis expands the state of the art in econometric approaches for ANNs by presenting an alternative to QML estimation that keeps its good asymptotic properties and has reduced leaning to overfitting. The presented approach relies on GMM estimation. As a byproduct, GMM estimation allows the use of the so-called J Test to verify the existence of neglected nonlinearity. The performed Monte Carlo studies indicate that the estimates from GMM are more accurate than those generated by QML in situations with high noise, especially in small samples. This result supports the hypothesis that GMM is susceptible to overfitting. Exchange rate forecasting experiments reinforced these findings. A second Monte Carlo study revealed satisfactory finite sample properties of the J Test applied to the neglected nonlinearity, compared with a reference test widely known and used. Overall, the results indicated that the estimation by GMM is a better alternative, especially for data with high noise level.
*

Keywords

Artificial Neural Networks; Multilayer *Perceptron*; Overfitting; Neglected Nonlinearity; *Quasi* Maximum Likelihood; Generalized Method of Moments; J-Test; Global Optimization.

Sumário

Lista de figuras

Lista de tabelas

Acrônimos e Siglas

Tabela 1: Siglas e Acrônimos

Sigla:	Significado:
BFGS	(Algoritmo) Broyden-Fletcher-Goldfarb-Shanno
CSA	(Solução) candidata a sobreajuste
DP	Diferença percentual
LWG	Teste para Não Linearidade Negligenciada de Lee-White-Granger
MGM	Método Generalizado dos Momentos
MM	Método dos Momentos
MQNL	(Método dos) Mínimos Quadrados Não Lineares
MQO	(Método dos) Mínimos Quadrados Ordinários
MQP	(Método dos) Mínimos Quadrados Ponderados
MV	(Método da) Máxima Verossimilhança
NLN	Não Linearidade Negligenciada
PGD	Processo gerador de dados
PMC	*Perceptron* de Múltiplas Camadas
PQIM	Programação Quadrática Inteira-Mista
QMV	(Método da) *Quasi* Máxima Verossimilhança
QMVP	(Método da) *Quasi* Máxima Verossimilhança Ponderada
REQM	Raiz do erro quadrático médio
RNA	Rede Neural Artificial
SQM	Soma dos quadrados dos resíduos
TML	Teste do Multiplicadores de Lagrange
TNLN	Teste para não linearidade negligenciada
Acrônimo:	**Significado:**
SNARC	*Stochastic Neural-Analog Reinforcement Calculator*

I can't live the buttoned down life like you. I want it all: the terrifying lows, the dizzying highs, the creamy middles! Sure, I might offend a few of the blue-noses with my cocky stride and musky odor - oh, I'll never be the darling of the so–called 'City Fathers' who cluck their tongues, stroke their beards, and talk about "What's to be done with this Homer Simpson?".

Homer J. Simpson, *no episódio Lisa's Rival, a propósito do sobreajuste e do que fazer a respeito.*

1
Introdução

O termo Redes Neurais Artificiais (RNAs), ou simplesmente Redes Neurais, está associado a uma vasta gama de modelos. Como característica comum, pode-se dizer que RNAs são modelos computacionais desenvolvidos com o objetivo de replicar características do cérebro humano, reconhecendo, associando e generalizando padrões. Do ponto de vista matemático, as RNAs são modelos capazes de aproximar funções genéricas e potencialmente desconhecidas [1].

Além da inspiração para o desenvolvimento destes modelos, outra analogia com o cérebro humano está na capacidade que estes modelos possuem de armazenar informações na sua estrutura em um processo semelhante à aprendizagem. Há também paralelos entre a estrutura de nós interconectados da rede e o sistema nervoso central. Não por acaso, estes nós são chamados de neurônios.

Os primeiros desenvolvimentos teóricos remontam aos anos 1940, como o modelo de neurônio artificial *Psychon* [2]. A primeira implementação apareceu na década seguinte. O *Stochastic Neural-Analog Reinforcement Calculator*, SNARC [3], operava em um computador valvulado e é considerada a primeira máquina capaz de aprender.

Um caso relevante dentro da cronologia das RNAs é o chamado *Perceptron*. Apresentado em 1958 [4], este classificador binário é o arquétipo de uma classe ampla de RNAs, chamadas RNAs Alimentadas Adiante (ou *feedfoward*). Porém, uma publicação de 1969 [5] mostrou a impossibilidade do *Perceptron* simples de resolver o problema do "ou-exclusivo", que vem a ser, para um conjunto de pares de entradas binárias, classificar em um grupo os casos nos quais as entradas são iguais e noutro os casos com entradas diferentes. Esta constatação colocou os *Perceptrons* em relativo ostracismo por mais de uma década. Uma análise detalhada deste caso pode ser encontrada em [6].

Nos anos 1980, porém, com a popularização dos computadores e seu significativo incremento de capacidade de processamento, as RNAs passaram a ter uma grande relevância nos meios acadêmicos. Novos tipos de RNAs foram desenvolvidos, como, por exemplo, as redes recorrentes Hopfield [7], bem como novos algoritmos de aprendizagem (ou treinamento). Ao final desta década, já havia congressos e publicações exclusivamente dedicados às RNAs.

Neste contexto, os *Perceptrons* reapareceram na versão com camadas escondidas entre as entradas e a saída [8]. Os chamados *Perceptrons* de Múltiplas

Camadas (PMCs) podiam resolver o "ou-exclusivo", porém seu poder ia muito além. Como foi demonstrado em [9], um PMCs com uma única camada escondida, sob certas condições, é capaz de aproximar qualquer função Borel-mensurável. Este resultado ficou conhecido como Teorema da Aproximação Universal e foi um grande impulsionador do uso dos PMCs nos anos posteriores. Outro fator que contribuiu decisivamente para a difusão dos PMCs foi a criação do método de treinamento conhecido como retropropagação. Este método calcula o gradiente da função-perda (geralmente o desvio absolto ou quadrático entre a saída da rede e o valor almejado) aplicando a regra da cadeia iterativamente ao longo das camadas escondidas dos PMCs.

Uma série de outros tipos de RNAs foram desenvolvidos desde então. Muitas delas possuem uma estrutura desenhada para um propósito específico e, consequentemente, são treinadas através de algoritmos próprios. Apenas como ordem de grandeza, o verbete da Wikipédia dedicado aos tipos de RNAs (em inglês) apresentava, em janeiro de 2015, sete tipos principais de RNAs (muitos dos quais com subtipos), além de uma categoria denominada "*Other Types*', com mais oito tipos aparentemente menos relevantes. Detalhes sobre a maioria destes tipos de RNAs podem ser encontrados no compêndio [1].

Apesar da diversidade de algoritmos de aprendizagem desenvolvidos existentes, todos podem ser classificados dentro de três paradigmas básicos. Na chamada aprendizagem supervisionada, há um conjunto estático de dados e, para cada vetor de entradas, existe um valor de referência para as saídas da rede, denominado alvo. A função-perda é definida como função do desvio entre a saída da rede e o respectivo alvo. A retropropagação enquadra-se neste paradigma. No caso do treinamento não-supervisionado, há também um conjunto estático de dados, porém não há um valor de referência e a função-perda tem como argumentos as entradas e a saída da RNA, como ocorre, por exemplo, em problemas de *clustering*. Existe, ainda, a aprendizagem por reforço, utilizada para a tomada de decisões em ambientes incertos no contexto de sistemas de controle. Neste paradigma, o conjunto de dados é dinâmico, atualizado de acordo com as interações entre a saída da RNA e o ambiente.

Um dos principais desafios encontrados na aplicação das RNAs é o chamado sobreajuste, especialmente nos casos em que a aprendizagem supervisionada é utilizada e há ruído nos valores dos alvos. Diz-se que uma RNA está sobreajustada quando ela ajusta uma quantidade relevante do ruído. O sintoma deste problema é uma grande discrepância entre a qualidade do ajuste dentro da amostra utilizada no treinamento vis-à-vis o ajuste em outras amostras. O ajuste dentro da amostra tende a ser excelente, enquanto o ajuste fora da amostra, também chamado de capacidade de generalização, costuma ser

pobre.

No caso dos PMCs, destacam-se duas classes de abordagens para lidar com o sobreajuste diretamente no algoritmo de treinamento, ambas descritas em [1]. A primeira delas é a combinação da validação cruzada com a parada prematura. Em linhas gerais, esta técnica consiste em separar a amostra em duas partes e aplicar o método da retropropagação em uma das amostras, monitorando o ajuste da rede à outra amostra (validação cruzada) e interrompendo o treinamento quando este ajuste começa a piorar (parada prematura). A segunda técnica é conhecida como regularização e consiste em modificar a função-perda da retropropagação, incluindo um termo que pune a complexidade da RNA, em geral associado à magnitude dos parâmetros. Há ainda as abordagens indiretas que buscam evitar o sobreajuste através de uma escolha parcimoniosa das entradas e da arquitetura do modelo. Um estudo comparativo entre representantes destas três classes de abordagens pode ser encontrado em [10], que realizou testes com dados simulados. Os resultados não permitem afirmar que alguma das abordagens é sistematicamente superior às demais.

O sobreajuste e outras dificuldades práticas, porém, não impediram a grande proliferação de aplicações de RNAs. As RNAs obtiveram resultados satisfatórios em tarefas das mais diversas, que vão desde o reconhecimento de escrita manual até aferição de risco de crédito, passando por previsões meteorológicas ou avaliação de imóveis. Há, na literatura, artigos que compilam aplicações de RNAs em áreas de conhecimento específicas, como, por exemplo, sistemas de controle [11], negócios [12, 13], processamento de imagens [14], finanças [15], medicina [16], economia [17] e sistemas elétricos [18].

Paralelamente à emergência das RNAs, e igualmente impulsionado pelo desenvolvimento e popularização dos computadores, os modelos de regressão não lineares ganharam espaço entre os econometristas. As técnicas econométricas utilizadas para lidar com este tipo de modelagem podem ser divididas em dois grupos principais.

No primeiro grupo, temos as abordagens não paramétricas e semiparamétricas, com especial destaque para as regressões com função-núcleo, como as propostas em [19, 20]. O outro grupo de técnicas utiliza-se de funções paramétricas flexíveis, a fim de aproximar as verdadeiras funções que, em geral, são as esperanças condicionais do problema de regressão. Como exemplos deste grupo, pode-se elencar as regressões polinomiais, os polinômios ortogonais (como os de Legendre e Chebyshev), os splines, as *wavelets* e as RNAs.

A utilização de RNAs sob uma perspectiva econométrica foi inaugurada no final dos anos 1980, com especial destaque para os PMCs com uma única camada escondida. Neste ponto, convém, para fins de demarcação, denominar

como RNAs tradicionais a literatura e as técnicas desenvolvidas fora do contexto econométrico. A tabela 1.1 apresenta as diferenças de nomenclatura utilizadas nestas duas abordagens de RNAs[1]. No presente trabalho, exceto quando explicitamente mencionado em contrário, o termo RNA, no contexto de modelos de regressão não linear, será equivalente a PMC com uma única camada escondida, exceto quando explicitamente mencionado em contrário.

Tabela 1.1: Diferenças de Nomenclatura

RNAs Tradicionas	RNAs em Econometria
Entradas	Regressores ou variáveis explicativas
Vieses e pesos sinápticos	Parâmetros
Neurônio	Unidades escondida
Saída	Esperança condicional estimada
Alvo	Variável dependente
Treinamento ou aprendizagem	Estimação paramétrica

Os trabalhos pioneiros na utilização de RNAs para regressão não linear foram [21, 22], nos quais é proposta a utilização do Método dos Mínimos Quadrados Não lineares (MQNL) para estimação dos parâmetros. Este método equivale ao Método da Máxima Verossimilhança sob ruídos gaussianos, independentes e homocedásticos, e, em casos mais gerais, equivale ao Método da *Quasi* Máxima Verossimilhança (QMV). Além disso, nestes artigos, as condições para identificação do modelo foram estabelecidas e foram demonstradas a consistência e a normalidade assintótica dos parâmetros estimados. Em [23], foi apresentado um teste estatístico para não linearidade negligenciada.

De acordo com [24], o processo de modelagem estatística utilizando-se RNAs é composto por três etapas: (i) a definição da arquitetura da RNA, (ii) a escolha dos regressores e (iii) a estimação dos parâmetros. Apesar de (i) e (ii) serem problemas de naturezas distintas, em ambos os casos, subjaz o princípio da parcimônia.

No caso da definição da arquitetura, existe um *tradeoff* entre flexibilidade da RNA e risco de sobreajuste. Algumas estratégias incrementais de determinação da arquitetura, utilizando testes estatísticos para nortear a decisão de colocar uma unidade escondida adicional, foram propostas, por exemplo, em [24, 25]. Uma das vantagens destas estratégias está no fato de evitar, ao longo do processo, a estimação de modelos sobreparametrizados (com excesso de unidades escondidas) e, consequentemente, não identificados. Os testes es-

[1]Neurônios podem estar na camada escondida ou na camada de saída. O termo unidade escondida refere-se aos neurônios da camada escondida.

tatísticos aplicados neste contexto são os chamados testes para não linearidade negligenciada (TNLNs).

O TNLN introduzido por [23] e revisitado em [26], conhecido teste Lee-White-Granger (LWG), utiliza a estatística do Teste dos Multiplicadores de Langrange para a regressão dos resíduos gerados pela RNA em um subconjunto dos componente principais de um conjunto de unidades escondidas com parâmetros gerados aleatoriamente. O TNLN apresentado [27] também utilizaram a estatística dos Multiplicadores de Langrange, porém regredindo os resíduos da RNA no polinômio de terceiro grau das variáveis explicativas. Testes utilizando regressões não paramétricas com funções-núcleo são apresentados em [28, 29]. Comparações entre alguns destes métodos podem ser encontradas em [26, 30].

Outra possibilidade presente na literatura é escolha da arquitetura baseada em critérios de informação, como o Critério de Informação de Schwartz [31], aplicado, por exemplo, em [10]. Há, ainda, propostas de utilização da validação cruzada como critério [32].

Para a seleção dos regressores, a mesma tríade de abordagens é aplicável: testes estatísticos [24], critérios de informação [33] e validação cruzada [34]. Uma comparação entre estratégias baseadas nestes três critérios, tanto para definição de arquitetura como para escolha de variáveis explicativas, pode ser encontrada em [35].

Por fim, em relação à estimação dos parâmetros, a QMV predomina fortemente. Em todos os artigos supracitados que utilizam RNAs dentro de um arcabouço econométrico, as estimações foram conduzidas por QMV. Como pontos favoráveis à utilização deste método, pode-se citar a consistência e a normalidade assintótica, demonstradas em [21, 22]. A consistência garante a convergência dos parâmetros para o verdadeiro valor, à medida que o tamanho de amostra vai para infinito, enquanto a normalidade assintótica permite a realização de testes de hipótese sobre os parâmetros estimados baseados na distribuição assintótica.

A propósito da estimação de parâmetros, para além das aplicações em RNAs, também nos anos 1980, mais precisamente em 1982, Lars Peter Hansen introduziu o Método Generalizado dos Momentos (MGM) [36]. Este método, baseado no Método dos Momentos (MM) desenvolvido por Pearson no final do século XIX, teve influência enorme em economia e finanças, a ponto de render ao seu criador o Prêmio Sveriges Riksbank de Ciências Econômicas em Memória de Alfred Nobel, em 2013.

Além de ser consistente e assintoticamente normal sob condições razoáveis, o MGM é um estimador muito versátil, que pode ser aplicado (com

os devidos ajustes) em todas as circunstâncias nas quais a QMV se aplica, além de outras particularmente relevantes. A enorme popularidade deste método entre os econometristas deve-se, em parte, a esta característica.

Apenas para dar alguma materialidade a popularidade do MGM, o *Google Scholar* (ferramenta de busca digital de publicações acadêmicas do *Google*) registrava, em julho de 2015, 9.667 citações do artigo original [36], enquanto a base bibliográfica dedicada à economia IDEIAS o colocava entre os vinte mais citados, na mesma época.

1.1 Motivação

Apesar da ampla utilização da estimação por QMV no contexto de RNAs como modelos econométricos e das suas boas propriedades assintóticas, o método possui um ponto fraco em amostras finitas. Trata-se da sua suscetibilidade à ocorrência de sobreajuste.

À primeira vista, pode-se pensar que este problema poderia ser resolvido utilizando-se uma arquitetura parcimoniosa. Porém, os testes realizados em [10] mostraram que, em alguns casos, as RNAs estimadas por QMV podem ter ajustes fora da amostra significativamente piores que outra RNAs, estimadas por outros métodos, mesmo quando o números de unidades escondidas na RNA estimada por QMV é determinado pelo método que o autor chamou "*divine guidance*" (orientação divina, numa tradução livre), ou seja, o número correto de unidades. A análise dos resultados mostra que nem mesmo este método de determinação de arquitetura, não factível em aplicações práticas, foi capaz de conter completamente o sobreajuste. No caso do método factível testado no artigo, que foi a seleção da arquitetura baseada no Critério de Informação de Schwartz, também com estimação por QMV, os sinais de sobreajuste são ainda mais evidentes e a performance ainda mais deteriorada. Um método de determinação da arquitetura da RNA não pode, por si só, evitar o sobreajuste[2], de tal maneira que é desejável que o método de estimação dos parâmetros colabore neste sentido.

Uma alternativa óbvia seria a utilização de uma das técnicas desenvolvidas e amplamente utilizadas na literatura de RNAs tradicionais: a combinação da validação cruzada com a parada prematura e a regularização. No caso da primeira técnica, pode-se dizer que, apesar do seu inegável valor prático, esta

[2]Não há contradição entre o *tradeoff* entre flexibilidade da RNA e risco de sobreajuste e a impossibilidade de a escolha arquitetura, por si só, evitar o sobreajuste. A primeira afirmativa estabelece que, basicamente, que o incremento de arquitetura da RNA tem como ponto positivo a maior flexibilidade e com efeito colateral o maior risco de sobreajuste, enquanto a segunda afirmativa explicita, simplesmente, que a arquitetura adequada não é capaz de oferecer garantias contra o sobreajuste sozinha.

possui pouco formalismo estatístico. Não há resultados que garantam a consistência e a normalidade assintótica dos parâmetros estimados, o que praticamente descarta sua utilização para fins econométricos.

Já o caso da regularização é diferente. Além de apresentar resultados bons resultados práticos na contenção do sobreajuste, esta técnica pode ser enquadrada em uma classe de estimadores chamados Estimadores de Máxima Verossimilhança Penalizada ou, no caso mais geral, na classe da *Quasi* Máxima Verossimilhança Penalizada (QMVP). Nesta classe, a função-objetivo é dada pela soma de uma parcela equivalente à QMV e outra que é uma penalização pela complexidade do modelo. Para esta classe de estimadores, há resultados de consistência e normalidade assintótica, sob certas condições de regularidade. Porém, a sua utilização apresenta algumas questões a serem consideradas.

A primeira delas é que a estimação por QMVP não é compatível com alguns dos métodos utilizados na definição da arquitetura e na escolha das variáveis, como, por exemplo, os critérios de informação. Portanto, dependendo da estratégia escolhida pelo econometrista para estas finalidades, a estimação por QMVP deixa de ser uma opção.

Outra questão fundamental é relativa à forma funcional da penalização a ser utilizada. É comum a utilização do produto entre uma constante, chamada de parâmetro de regularização, e alguma norma aplicada ao vetor de parâmetros do modelo. Tanto a escolha do parâmetro de regularização como a escolha da norma podem afetar significativamente os resultados da estimação. No caso da utilização de normas canônicas L^p com $0 < p < 1$, a função de penalização não é convexa. Como consequência, o procedimento de estimação passa a interferir na definição da arquitetura e escolha dos parâmetros, podendo forçar a zero os parâmetros de alguns neurônios ou variáveis [37]. Esta interferência pode ser indesejável ao econometrista.

A escolha do parâmetro de regularização é igualmente crítica. Um parâmetro alto demais introduz um viés que pode ser além do tolerável para a estimação. Segundo [10], é comum que esta escolha seja baseada em validação cruzada ou *bootstrapping*. Estes procedimentos podem ser extremamente custosos do ponto de vista computacional. Nos casos em que $p = 1$ ou $p = 2$, é possível interpretar o termo de penalização como uma *prior* Bayesiana, com distribuições de Laplace e Gaussiana, respectivamente. Nestes casos, é possível realizar uma estimação conjunta do parâmetro de regularização e dos parâmetros do modelo, usando o método conhecido como Máxima Probabilidade à Posteriori, que não é trivial e também pode envolver um alto custo computacional.

A escolha da função de penalização não é uma mera questão de pre-

ferências do econometrista, uma vez que possui um papel fundamental nas propriedades assintóticas do estimador. Via de regra, as condições de regularidade impostas sobre a função de penalização são que a penalização cresça ilimitadamente com o tamanho de amostra, porém com peso na estimação vis-à-vis a parcela equivalente à QMV tendendo a zero. Com isso, garante-se que, assintoticamente, o estimador de QMVP convergirá para o de QMV. Não é de se estranhar que nem todos os econometristas sintam-se confortáveis em utilizar um estimador com penalização baseando-se em propriedades assintóticas obtidas sob condições nas quais a penalização tem peso nulo. Adicionalmente, [37] cita o fato de que as limitações impostas pelas condições de regularidade podem dificultar a escolha do parâmetro de regularização na prática, enquanto [38] argumenta que, apesar de serem típicas e plausíveis, as condições de regularidade costumam ser difíceis de se verificar a partir de condições mais primitivas.

Em resumo, a estimação por QMVP, ou a regularização, pode, por conta destas limitações, não ser uma alternativa adequada à QMV no caso da estimação dos parâmetros de RNAs.

1.2 Objetivo

O objetivo do presente trabalho é propor um método para a estimação dos parâmetros de RNAs que conjugue três características principais:

- mais robustez ao sobreajuste que a QMV;
- boas propriedades assintóticas (consistência e normalidade); e
- ausência das limitações como as dos estimadores de QMVP (por exemplo, a dificuldades supracitadas referentes à escolha da função de penalização).

A proposta aqui apresentada é a utilização do MGM. Além da sua versatilidade e das boas propriedades assintóticas, acredita-se que o MGM seria menos propenso a gerar sobreajuste. Uma explicação intuitiva para esta crença está no fato de o MGM romper com o paradigma de minimização dos erros quadráticos, fortemente associado ao sobreajuste. Argumentos mais técnicos serão desenvolvidos mais adiante.

1.3 Contribuições

A principal contribuição do presente trabalho é a **apresentação de um arcabouço geral para estimação dos parâmetros de RNAs utilizando o MGM**, combinando condições de momento derivadas: (i) do problema de otimização da estimação por QMV e (ii) da hipótese de exogeneidade estrita (comumente utilizada em regressão linear) aplicada a funções não lineares dos regressores.

Como contribuições secundárias, pode-se mencionar:

1. **apresentação e implementação de um caso particular do arcabouço proposto**, utilizando como funções não lineares dos regressores combinações lineares de saídas de unidades escondidas com parâmetros aleatórios;

2. **realização de experimentos de Monte Carlo a fim de comparar a qualidade das estimações realizadas pelo MGM vis-à-vis a QMV**, sob diferentes níveis de ruído e com diferentes tamanhos de amostra;

3. **realização de experimentos de previsão de taxas de câmbio e comparação da precisão das previsões geradas por modelos RNAs estimadas por QMV e MGM**, utilizando dez séries mensais com mais de vinte anos de duração;

4. **apresentação e implantação de um novo TNLN para RNAs estimadas por MGM** e a realização de experimentos de Monte Carlo para compará-lo com um *benchmark* tradicional da literatura, o teste LWG;

5. **derivação de um limite inferior aproximado para a probabilidade de ocorrência de sobreajuste** em RNAs estimadas por QMV, como função do tamanho da amostra e do nível de ruído;

6. **apresentação de uma formulação do problema de otimização da estimação por QMV com garantia de otimalidade global** para RNAs com um tipo específico de função de ativação.

As abordagens econométricas para RNAs [21, 22] não dispõem de mecanismos diretos para controlar o sobreajuste. As alternativas desenvolvidas para a contenção do sobreajuste na literatura de RNAs tradicionais [1] sofrem de falta de formalismo estatístico ou de dificuldades práticas na obtenção de

garantias em relação às suas propriedades assintóticas. A contribuição do presente trabalho ao estado da arte está em fechar estas lacunas, apresentando um método de estimação de RNAs robusto ao sobreajuste e que preserva as propriedades estatísticas desejáveis presentes em [21, 22].

1.4 Organização do Trabalho

Por tratar de um tema situado na fronteira entre RNAs e estatística, os dois próximos capítulos da presente tese trazem um apanhado dos conceitos fundamentais para que leitores com formação em qualquer um dos dois campos tenham um entendimento pleno das contribuições realizadas. O segundo capítulo apresenta os principais conceitos de RNAs relacionados ao presente trabalho. No terceiro capítulo são revistos os fundamentos estatísticos relevantes a este neste trabalho, com foco nos métodos de estimação QMV e MGM. O quarto traz a estimação de RNAs por QMV e questões relacionadas, o arcabouço geral para a estimação de RNAs pelo MGM, o caso particular proposto e o TNLN para RNAs estimadas pelo MGM. No quinto capítulo estão os estudos de caso: (i) os experimentos de Monte Carlo para estimação, (ii) os experimentos de previsão de taxas de câmbio, e (iii) os experimentos de Monte Carlo para os TNLNs. As conclusões e possíveis extensões do trabalho estão no sexto e último capítulo.

2
Redes Neurais Artificiais

Neste capítulo, serão apresentados com maior profundidade técnica os conceitos de RNAs tradicionais que possuem grande interseção com o presente trabalho.

2.1
Perceptrons

Os *Perceptrons* originais [4] eram simples classificadores binários, que mapeavam um vetor de entradas $\vec{x} \in \mathbb{R}^j$ em um saída escalar $y \in \{0, 1\}$. Sua parametrização é dada por um vetor de pesos sinápticos, $\vec{\omega} \in \mathbb{R}^j$, e um viés, $b \in \mathbb{R}$. Sua fórmula é dada por:

$$f(\vec{x}) = \begin{cases} 1, & \text{se } \vec{\omega}' \cdot \vec{x} + b \geq 0, \\ 0, & \text{caso contrário.} \end{cases} \tag{2-1}$$

Este classificador, desenvolvido para reconhecimento de imagens com fins militares, causou grande expectativa em relação às suas potencialidades. Porém, suas limitações se tornaram claras quando ficou comprovada sua impossibilidade de resolver o problema do "ou-exclusivo". Neste problema, há duas entradas binárias ($\vec{x} \in \{0, 1\}^2$) e a classificação esperada é:

$$f(\vec{x}) = \begin{cases} 1, & \text{se } x_1 \neq x_2, \\ 0, & \text{se } x_1 = x_2. \end{cases} \tag{2-2}$$

Para se obter a classificação correta, seria necessário encontrar parâmetros $\vec{\omega} \in \mathbb{R}^2$ e $b \in \mathbb{R}$ que satisfaçam as seguinte desigualdades:

1. $\omega_1 \cdot 0 + \omega_2 \cdot 0 + b < 0$;
2. $\omega_1 \cdot 1 + \omega_2 \cdot 0 + b \geq 0$;
3. $\omega_1 \cdot 0 + \omega_2 \cdot 1 + b \geq 0$; e
4. $\omega_1 \cdot 1 + \omega_2 \cdot 1 + b < 0$.

Ocorre que, sempre que três destas desigualdades forem satisfeitas, a outra, necessariamente, não será. Uma forma simples de provar esta afirmativa é mostrar que a soma dos lados esquerdos das desigualdades 1 e 4 é igual à soma dos lados esquerdos das desigualdades 2 e 3. Logo, pelo menos uma das

desigualdades não é válida uma vez que a soma de dois números não positivos não pode ser igual à soma de dois números positivos.

Esta simples demonstração [5] expôs as limitações dos *Perceptrons*. De forma mais geral, demonstrou-se que os *Perceptrons* simples são classificadores lineares e, portanto, só conseguem classificar corretamente se as entradas das diferentes classes forem linearmente separáveis, ou seja, se houver um hiperplano capaz de separar as entradas das duas classes.

2.2 Perceptrons de Múltiplas Camadas

A fim de possibilitar classificações não lineares, foram propostos os PMCs [8]. O número de camadas é maior ou igual a três, sendo uma camada de entradas, uma ou mais camadas escondidas e uma camada de saídas, nesta ordem, cada qual com uma determinada quantidade de unidades que, exceto na camada de entradas, são denominadas neurônios.

A ideia básica por trás dos PMCs é que cada camada (exceto a primeira) transforma os valores da camada imediatamente anterior. O valor do j-ésimo neurônio da k-ésima camada, $v_{j,k}$ para $k \geq 2$, pode ser obtido iterativamente pela fórmula:

$$v_{j,k} = h_k\left(b_{j,k} + \sum_{m=1}^{M} \omega_{j,m,k} \cdot v_{m,k-1}\right), \tag{2-3}$$

onde $b_{j,k}$ é o viés do j-ésimo neurônio da k-ésima camada, $\omega_{j,m,k}$ é o peso sináptico da ligação entre o j-ésimo neurônio da k-ésima camada e o m-ésimo neurônio da camada anterior, M é o número de neurônios da $(k-1)$-ésima camada, e $h_k(\cdot)$ é a chamada função de ativação utilizada na k-ésima camada.

As funções de ativação costumam ser não lineares, especialmente nas camadas escondidas. As funções mais utilizadas em PMCs são:

- logística: $h(x) = (1 - \exp^{-x})^{-1}$;
- tangente hiperbólica: $h(x) = \frac{\exp^{x} - \exp^{-x}}{\exp^{x} + \exp^{-x}}$;
- rampa ou linear por partes: $h(x) = \max(-1, \min(1, x))$;
- degrau: $h(x) = \begin{cases} 1, & \text{se } x \geq 0, \\ 0, & \text{caso contrário} \end{cases}$;
- identidade: $h(x) = x$.

Representações gráficas destas funções de ativação encontram-se na figura 2.1.

De volta ao problema do "ou-exclusivo", tomemos um PMC com a seguinte estrutura: (i) uma única camada escondida com dois neurônios; (ii) todos os pesos sinápticos entre as entradas e a camada escondida iguais a um;

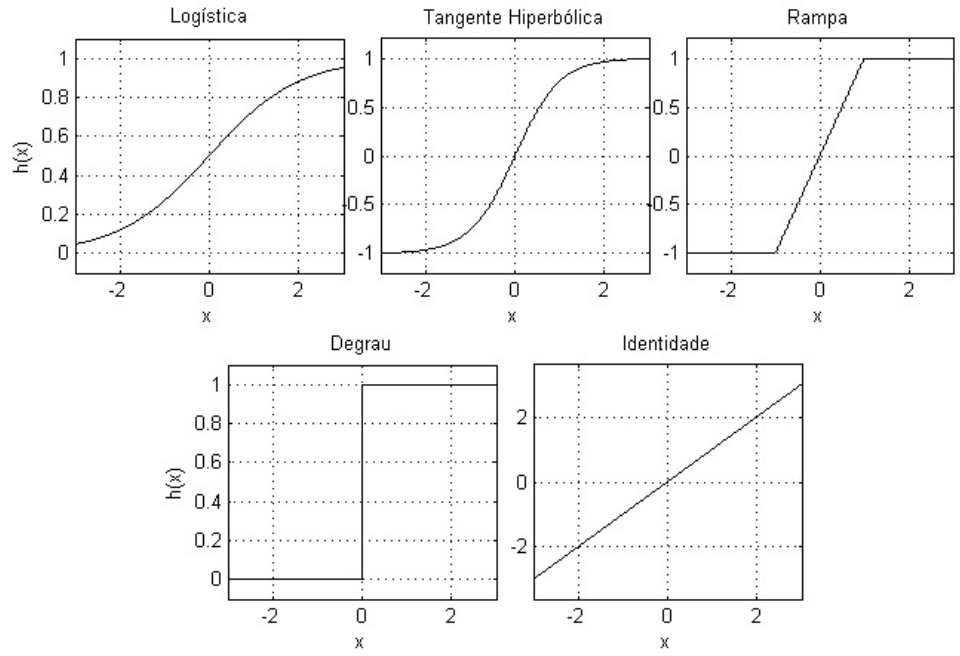

Figura 2.1: Funções de Ativação.

(iii) os vieses $b_{1,2} = -1/2$, $b_{2,2} = -3/2$ e $b_{1,3} = -1/2$; (iv) os demais pesos sinápticos $\omega_{1,1,3} = 1$ e $\omega_{1,2,3} = -1$; e (v) todas as funções de ativação do tipo degrau. A tabela 2.1 descreve como este PMC classifica corretamente o "ou-exclusivo".

Tabela 2.1: PMC para o "Ou-exclusivo"

Entradas		Camada Escondida		Saída
x_1	x_2	$v_{1,2}$	$v_{2,2}$	$v_{1,3}$
0	0	0	0	0
1	0	1	0	1
0	1	1	0	1
1	1	1	1	0

Os PMCs, com uma única camada escondida, são capazes de realizar classificações corretas mesmo nos casos em que as classes não são linearmente separáveis. Porém, sua capacidade vai além dos problemas de classificação.

2.3 Teorema da Aproximação Universal

Em 1989, [9] demonstrou que um PMC, com uma única camada escondida utilizando neurônios com funções de ativação do tipo logística e com função

de ativação identidade na única saída, é capaz de aproximar qualquer função contínua definida em um domínio compacto, com um nível de precisão arbitrário. Este resultado ficou conhecido na literatura como Teorema da Aproximação Universal:

Teorema 2.1 *Sejam $I_J \in [0,1]^J$ um hipercubo J-dimensional, $C(I_J)$ o espaço das funções contínuas em I_J, $\|\cdot\|$ a norma do supremo e $h(\cdot)$ a função de ativação logística. Para qualquer função $f(x) \in C(I_J)$ e $\varepsilon > 0$, existem M e $\{\omega_{m,1,3}, \{\omega_{j,m,2}\}_{j=1}^{J}, b_m\}_{m=1}^{M}$, que definem uma função*

$$G(\vec{x}) = \sum_{m=1}^{M} \omega_{m,1,3} \cdot h\left(b_m + \sum_{j=1}^{J} \omega_{j,m,2} \cdot x_j\right),$$

tal que $\|f(\vec{x}) - G(\vec{x})\| < \varepsilon$.

Ainda em 1989, [39] apresentaram resultados mais gerais, expandido o resultado para aproximação de uma classe mais abrangente de funções, as funções Borel-mensuráveis, utilizando qualquer norma canônica, e demonstrando que basta a função de ativação da camada escondida ser limitada e não constante. Ou seja, a capacidade de aproximação dos PMCs está relacionada à sua arquitetura e não ao tipo de função de ativação utilizado.

Estes importantes resultados contribuíram fortemente para a popularização dos PMCs, especialmente com uma única camada escondida. Porém, eles não trazem informações sobre qual seria no valor de M necessário para atingir o nível de precisão desejado, nem, tampouco, sobre como se obtém $\{\omega_{m,1,3}, \{\omega_{i,m,2}\}_{j=1}^{J}, b_m\}_{m=1}^{M}$.

2.4 Retropropagação

Aprendizagem ou treinamento são os nomes atribuídos ao processo de determinação dos pesos sinápticos e vieses em uma RNAs. No paradigma de aprendizagem supervisionada, este processo consiste na redução iterativa de uma função-perda definida sobre os desvios (ou erros) das saídas da RNA em relação aos respectivos alvos. Normalmente, as funções perda utilizadas são o erro quadrático médio (EQM) e o erro absoluto médio (EAM). Estas iterações são repetidas até que algum critério de parada seja atingido. Exemplos de critérios de parada são: (i) numero de iterações; (ii) valor da função-perda; e (iii) módulo do vetor gradiente da função-perda.

No caso dos PMCs com funções de ativação diferenciáveis, pode-se utilizar o algoritmo da retropropagação [8]. A ideia por trás deste algoritmo é atualizar os vieses e pesos sinápticos na direção oposta à do vetor gradiente da

função-perda, dado pelas derivadas parciais da função-perda em relação aos vieses e pesos sinápticos.

Supondo um PMCs com K camadas escondidas e uma única saída, a derivada da função-perda l em relação a um peso sináptico $\omega_{j,m,k}$, que liga o j-ésimo neurônio da k-ésima camada e o m-ésimo neurônio da camada anterior que liga, para $K < k < 1$, é dada por:

$$\begin{aligned}\frac{\partial l}{\partial \omega_{j,m,k}} &= (\nabla_{\vec{v}_{j,k}} l)' \cdot \frac{\partial \vec{v}_{j,k}}{\partial \omega_{j,m,k}} = \\ &= (\nabla_{\vec{v}_{j,k}} l)' \cdot diag\left(h'_k\left(b_{j,k} + \sum_{m=1}^{M} \omega_{j,m,k} \cdot \vec{v}_{m,k-1}\right)\right) \cdot \vec{v}_{m,k-1}, \quad (2\text{-}4)\end{aligned}$$

onde a seta sobre o valor dos neurônios denota um vetor coluna obtido pelo empilhamento dos valores do neurônio para cada um dos vetores de entradas, ∇ o vetor gradiente em relação ao subíndice, h'_k a derivada da função de ativação da k-ésima camada e $diag(\cdot)$ a matriz diagonal com os elementos dados pelo argumento.

Ou seja, a derivada parcial em relação a um peso sináptico depende de três elementos: (i) a derivada parcial de l em relação ao valor do neurônio ao qual ele chega; (ii) a derivada da função de ativação; e (iii) o valor do neurônio do qual ele parte (no caso dos vieses, este valor é constante e igual a um). O primeiro elemento é dado por:

$$\nabla_{\vec{v}_{j,k}} l = \begin{cases} diag(l'(\vec{y}_j - \vec{v}_{j,k})) \cdot \vec{v}_{1,k}, & \text{se } k = K, \\ \sum_{j=1}^{M_{k+1}} diag(\nabla_{\vec{v}_{j,k+1}} l) \cdot \left(h'_{k+1}\left(c_{j,k+1}\right) \cdot \omega_{j,m,k+1}\right), & \text{se } K > k > 1. \end{cases} \quad (2\text{-}5)$$

onde $l'(\cdot)$ denota a derivada da função-perda, $\vec{y}_j$ é o vetor coluna com os valores dos alvos para a j-ésima saída, M_{k+1} é o número de neurônios da $(k+1)$-ésima camada e $c_{j,k+1} \equiv b_{j,k+1} + \sum_{m=1}^{M_k} \omega_{j,m,k+1} \cdot \vec{v}_{m,k}$.

Portanto, é possível, partindo da última camada e retroagindo iterativamente, obter todas as derivadas parciais da função-perda em relação aos pesos sinápticos e vieses. Este processo é o cerne da retropropagação, que pode ser sumarizada nos seguintes passos:

1. Inicializa $\{\{\omega_{j,m,k}\}_{j=1}^{M_k}, b_m\}_{m=1}^{M_{k-1}}\}_{k=2}^{K}$ com valores aleatórios (preferencialmente próximos de zero);

2. Calcula $\{\{\vec{v}_{j,k}\}_{j=1}^{M_k}\}_{k=2}^{K}$;

3. Calcula iterativamente as derivadas parciais $\frac{\partial l}{\partial \omega_{j,m,k}}$;

4. Atualiza os pesos sinápticos (ou vieses) através da regra:

$$\omega_{j,m,k} \leftarrow \omega_{j,m,k} - q \cdot \frac{\partial l}{\partial \omega_{j,m,k}},$$

   onde $0 < q < 1$ é o a chamada taxa de aprendizagem;

5. Calcula o(s) critério(s) de parada;

6. Repete os passos 2-5 até o atingimento do(s) critério(s) de parada.

Este algoritmo explica em parte a popularidade dos PMCs e até mesmo o ressurgimento do interesse pelas RNAs. Segundo [1], um dos motivos pelos quais os PMCs são o tipo de RNAs mais utilizado é a sua facilidade de implementação, o que está diretamente ligado à retropropagação. Não obstante, o algoritmo possui deficiências, como a possibilidade de convergência para um ponto de mínimo local indesejável e a inicialização randômica, que adiciona um grau de aleatoriedade aos resultados finais obtidos. [1], novamente, apresenta alguns avanços no sentido de minimizar estes problemas.

2.5 Sobreajuste

Em muitas situações nas quais se utilizam RNAs, os valores dos alvos possuem algum nível de ruído. Em outras palavras, é comum que os valores dos alvos sejam compostos por uma função determinística das entradas e um ruído aleatório. Idealmente, nestes casos, esperar-se-ia que a RNA fosse capaz de aproximar somente a parte determinística e desprezar o ruído.

Na prática, porém, pode ocorrer de a RNA, por conta da sua flexibilidade, encontrar uma forma de ajustar parte do ruído, de modo a minimizar o desvio entre as saídas e os alvos. Este fenômeno é denominado sobreajuste e possui dois sintomas: (i) alto nível de ajuste na amostra utilizada no treinamento; e (ii) baixo poder de generalização, ou seja, baixa capacidade de predição em amostras diferentes da utilizada no treinamento.

Um simples exemplo, utilizando outra forma funcional flexível, os polinômios, pode ilustrar este fenômeno. Foram gerados dados a partir do seguinte processo:

$$y_n = \cos(x_n) + \epsilon_n,\ n \in 1, 2, \ldots, 100$$

onde $\cos(\cdot)$ é a função cosseno, $x_n \equiv \frac{n}{10}$ e ϵ_n é sorteado de uma normal com média nula e desvio padrão igual a dois. Em seguida, foram rodadas as regressões polinomiais:

$$\hat{y}_n = \sum_{j=0}^{J} \beta_j x_n^j,$$

para $J = 0, 1, 2, \ldots, 10$. A amostra obtida (asteriscos), a parte determinística (linha sólida) e alguns das funções estimadas (linhas tracejadas) estão na figura 2.2.

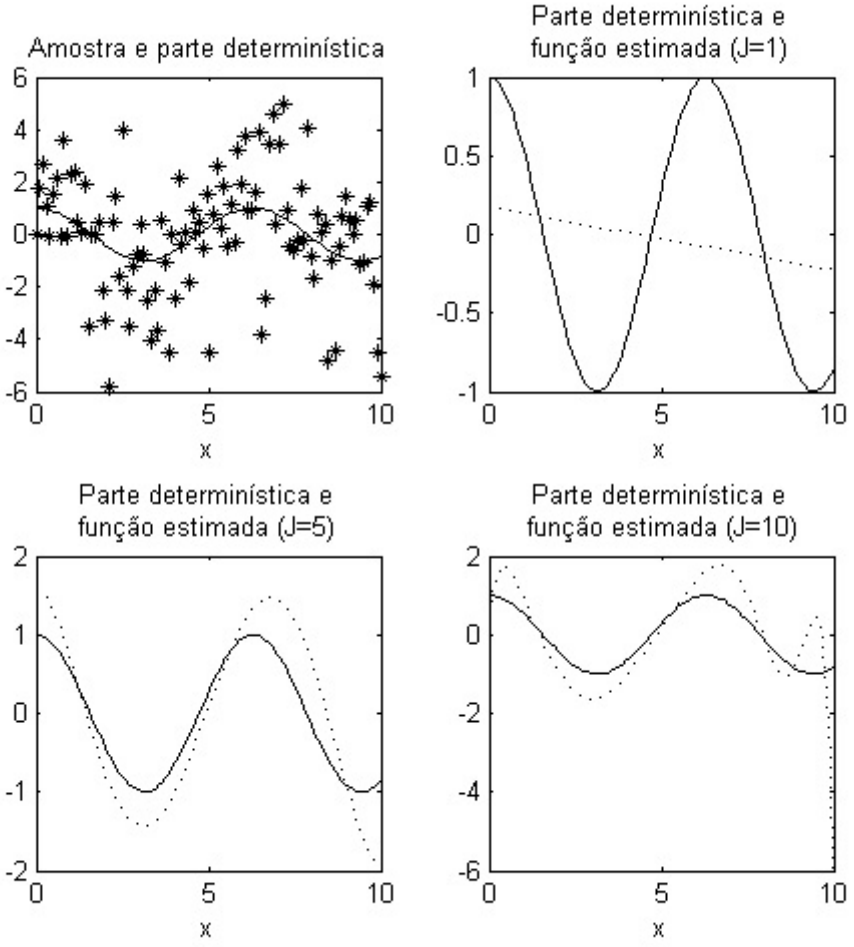

Figura 2.2: Sobreajuste em Regressões Polinomiais.

Como se pode notar, no caso em que $J = 1$, a função resultante é afim e não possui flexibilidade o suficiente para ajustar-se às curvas da parte determinística de y. Este caso é denominado subajuste. Já para a $J = 5$, a função estimada segue adequadamente a parte determinística de y, com algum nível de erro que é inerente à presença de ruído. Já no caso em que $J = 10$, o excesso de graus de liberdade permite à função estimada ajustar-se parcialmente ao ruído. Este fenômeno é bem visível na queda abrupta que a função estimada tem nos valores de x próximos de 10. Este é um exemplo de sobreajuste.

No gráfico da figura 2.3, estão representados os erros quadráticos médios (EQMs) dos polinômios estimados em relação à amostra (linha sólida) e à sua parte determinística (linha tracejada). O ajuste em relação à amostra só

melhora à medida que o grau do polinômio é incrementado (EQM diminui). Por outro lado, o EQM em relação à parte determinística atinge um mínimo quando $J = 5$ e cresce vertiginosamente em seguida.

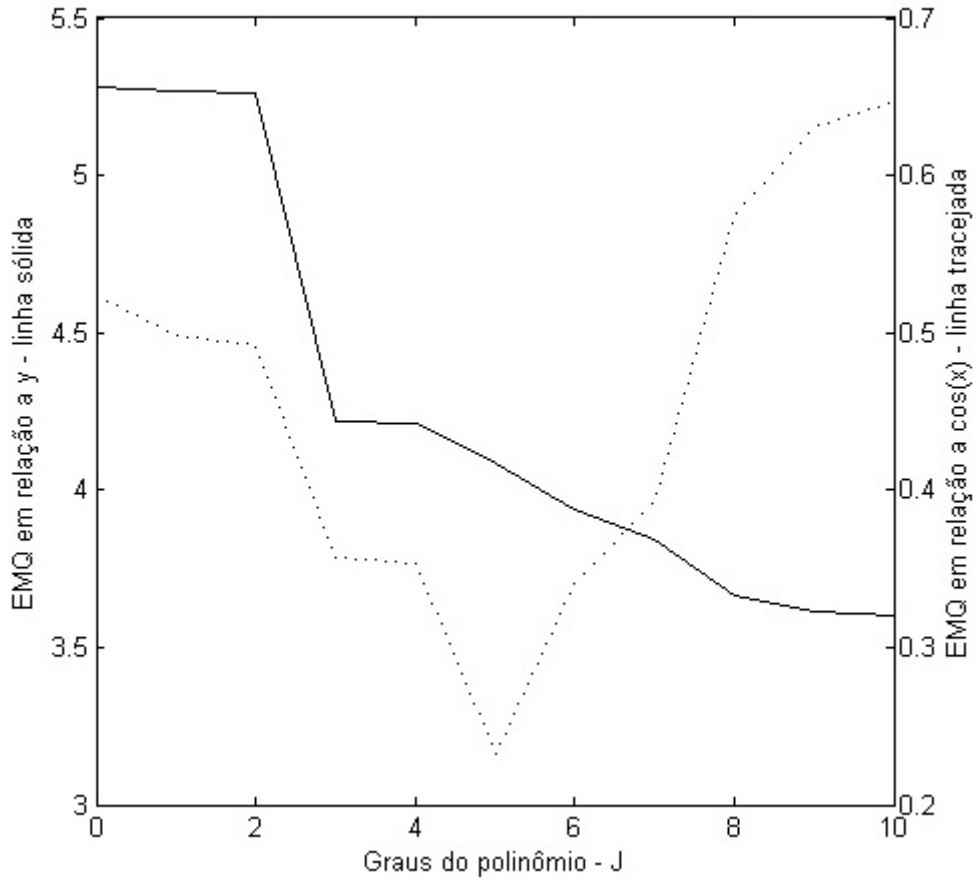

Figura 2.3: Ajustes das Regressões Polinomiais.

De forma análoga ao que se observa no exemplo das regressões polinomiais, as RNAs e, em particular, os PMCs também se tornam mais propensos à ocorrência de sobreajuste quanto mais complexas forem suas estruturas. No caso PMCs, a regularização e combinação da validação cruzada com parada prematura são as técnicas mais comumente aplicadas a fim de conter o sobreajuste.

3
Métodos de Estimação

Neste capítulo, serão apresentados os dois principais métodos de estimação utilizados no presente trabalho, que são a *Quasi* Máxima Verossimilhaça (QMV) e o Método Generalizado dos Momentos (MGM). Além disso, serão também expostos, em linhas gerais, os dois métodos dos quais estes se originaram: a Máxima Verossimilhança (MV) e o Método dos Momentos (MM), respectivamente.

Para fins didáticos, um mesmo exemplo simples de problema de estimação será abordado por estes quatro métodos. Trata-se de uma amostra com 10 observações $X = \{3, 2, 2, 1, 2, 4, 5, 2, 5, 5\}$, a qual se acredita ter sido gerada independentemente a partir de uma Distribuição de Poisson, definida por um único parâmetro, λ, desconhecido. Os dados, de fato, foram gerados por uma Distribuição de Poisson com $\lambda = 3$. É importante ressaltar que os exemplos apresentados são meramente ilustrativos, expostos a fim de demonstrar os procedimentos inerentes a cada método de estimação, e os resultados obtidos não devem servir de base para comparações entre os métodos.

3.1
Máxima Verossimilhança

Desenvolvido no início do século XX e apresentado nos trabalhos [40,41], a MV tornou-se um dos mais importantes métodos para estimação de parâmetros e de inferência para estatísticos e econometristas.

Informalmente, o princípio da estimação por MV é que, dadas uma amostra e uma família de distribuições conjuntas, definida por um conjunto de parâmetros, a distribuição conjunta que gerou a amostra deve ser aquela cujos parâmetros tornam mais provável a ocorrência desta amostra em um sorteio aleatório.

Formalmente, seja uma amostra $\{x_n\}_{n=1}^{N}$ de tamanho N, a família de distribuições de probabilidade conjuntas $p(\cdot, \psi)$ definida pelos parâmetros $\psi \in \Psi$ e a função de verossimilhança $L(\psi) = p(\{x_n\}_{n=1}^{N}, \psi)$. O estimador de MV dos parâmetros ψ é dado por:

$$\hat{\psi}_{MV} = \underset{\psi \in \Psi}{\operatorname{argmax}}\, L(\psi). \tag{3-1}$$

Aplicando-se a MV ao exemplo deste capítulo, temos que $\psi = \{\lambda\}$, $\Psi = \mathbb{R}^+$ e a função de verossimilhança $L(\lambda) = \prod_{n=1}^{N} \lambda^{x_n} \cdot^{-\lambda} \cdot (x_n!)^{-1}$. Inserindo-

se estes elementos na equação (3-1), obtemos $\hat{\lambda}_{MV} = 3,1$. A figura 3.1 apresenta os valores da função de verossimilhança para um intervalo escolhido de λ.

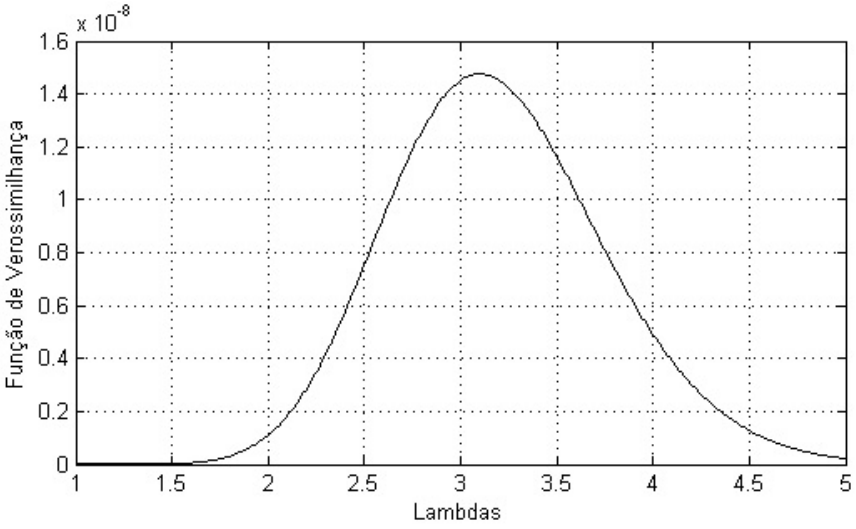

Figura 3.1: Exemplo de Estimação por MV.

Os estimadores de MV, sob certas condições razoáveis, possuem as seguintes propriedades assintóticas: (i) consistência; (ii) normalidade; e (iii) eficiência[1]. O detalhamento destas condições de regularidade e as demonstrações das propriedades fogem ao escopo do presente trabalho, mas podem ser encontradas, por exemplo, em [42].

Em termos de testes de hipóteses sobre os parâmetros, há uma trinca clássica que acompanha os estimadores de MV: (i) teste de Wald; (ii) teste dos Multiplicadores de Lagrange (ou Score); e (iii) teste da Razão de Verossimilhança. Sejam as hipóteses do teste:

- Hipótese Nula: $\psi \in \Psi'$ onde $\Psi' \subset \Psi$, contra;
- Hipótese Alternativa: $\psi \in (\Psi \setminus \Psi')$.

No caso do Teste de Wald, somente o modelo irrestrito ($\psi \in \Psi$) é estimado, enquanto o Teste dos Multiplicadores de Lagrange baseia-se na estimação do modelo restrito ($\psi \in \Psi'$) e o Teste da Razão de Verossimilhança requer ambas as estimações. Assintoticamente, os três testes são equivalentes e seguem, sob a hipótese nula, distribuições $\mathcal{X}^2$, quando válidas as condições de normalidade assintótica do estimador de MV.

A premissa fundamental da estimação por MV é que a família de distribuições conjuntas contém a verdadeira distribuição conjunta da qual a amostra foi obtida. Diz-se, nestes casos, que o modelo está corretamente

[1]Eficiência no sentido que a matriz de covariância do estimador converge para o limite inferior de Cramér-Rao. Detalhes apresentados no apêndice B.

especificado. Em muitos casos práticos, porém, não há como garantir a validade desta premissa.

3.2 Quasi Máxima Verossimilhança

O que acontece com os estimadores de MV quando o modelo não está corretamente especificado? Eles possuem algum significado? Quais são suas propriedades? Pode-se dizer algo a respeito de sua convergência? Os testes de hipótese permanecem válidos? Estas perguntas motivaram o desenvolvimento dos estimadores de *Quasi* (ou Pseudo) Máxima Verossimilhança (QMV).

Basicamente, os estimadores de QMV são obtidos pela maximização de uma função de verossimilhança que não necessariamente está especificada corretamente. Neste sentido, a MV é um caso particular da QMV no qual o modelo está corretamente especificado. Por exemplo, no caso de uma regressão linear com ruídos gaussianos heteroscedásticos, uma estimação por MV assumindo-se um ruído homocedásticos poderia ser classificada como uma estimação (consistente) por QMV.

Apesar de importantes desenvolvimentos anteriores, [43] destaca-se entre os trabalhos pioneiros nesta área[2]. Neste artigo, são apresentadas condições suficientes para garantir a consistência e a normalidade assintótica desta classe de estimadores, além de apresentar versões robustas a falhas na especificação do modelo para os testes de Wald e dos Multiplicadores de Lagrange.

Um caso de particular interesse para o presente trabalho é quando os parâmetros de interesse são os que determinam a esperança condicional e assume-se que as distribuições conjuntas são gaussianas independentes com uma variância constante σ^2 pré-determinada (não necessariamente verdadeira). Os resultados apresentados em [47] garantem a consistência e a normalidade assintótica destes estimadores:

$$\hat{\psi}_{QMV} = \underset{\psi \in \Psi}{\operatorname{argmax}} \prod_{n=1}^{N} (\sigma \cdot \sqrt{2\pi})^{-1} . ^{-\frac{(x_n - \mu(\psi))^2}{2 \cdot \sigma^2}} . \quad (3\text{-}2)$$

Aplicando-se a função logaritmo (transformação monotônica) à função objetivo de (3-2), eliminando-se os seus termos constantes (todos positivos) e multiplicando-a por menos um (trocando a maximização por minimização), obtém-se o problema equivalente:

[2]Outros trabalhos relacionados ao que viria a ser conhecido como o Método da *Quasi* (ou Pseudo) Máxima Verossimilhança são [44–47]. A controvérsia a respeito da autoria do método foge ao escopo do presente trabalho.

$$\hat{\psi}_{QMV} = \underset{\psi \in \Psi}{\operatorname{argmin}} \sum_{n=1}^{N} (x_n - \mu(\psi))^2. \tag{3-3}$$

Ou seja, este caso particular da QMV é equivalente a um problema de minimização da soma (ou da média) dos erros quadráticos.

Retomando o exemplo deste capítulo, a figura 3.2 apresenta uma comparação entre os valores da funções de verossimilhança e de *quasi* verossimilhança (função objetivo da equação (3-2)) com $\sigma^2 = 3$. Apesar de possuírem formas distintas, ambas as funções atingem o máximo no mesmo ponto $\hat{\lambda} = 3,1$.

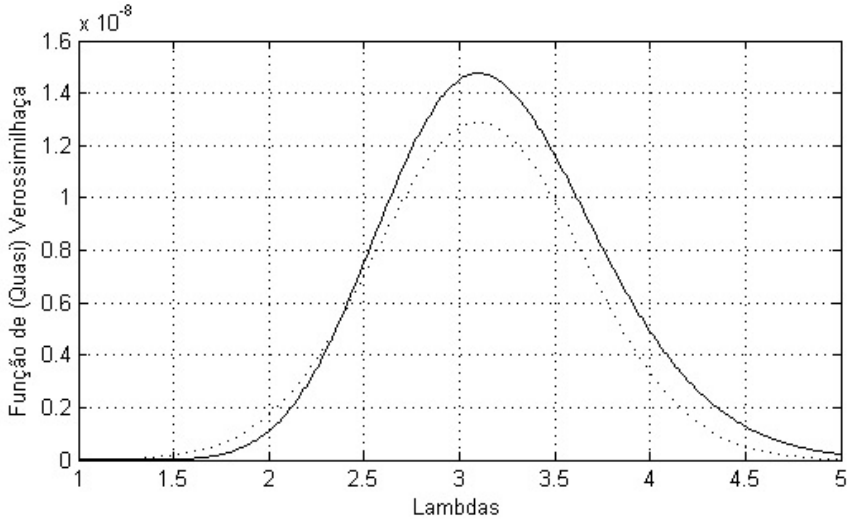

Figura 3.2: Comparação entre MV e QMV.

No escopo do presente trabalho, o termo QMV é equivalente ao caso particular representado pela equação (3-2) embora, em muitos casos, seja conveniente utilizar-se a formulação equivalente apresentada na equação (3-3).

3.3 Método dos Momentos

Assim como diversos outros métodos estatísticos amplamente difundidos, o Método dos Momentos (MM) deve-se a Karl Pearson[3]. Sua motivação para o desenvolvimento deste método, que remonta ao final do século XIX, era a estimação dos parâmetros das distribuições paramétricas que ele havia desenvolvido, posteriormente nomeadas Distribuições de Pearson tipos I a VII.

Estas distribuições possuem momentos populacionais descritos analiticamente como função dos seus parâmetros. Os momentos amostrais, por sua vez, são grandezas facilmente calculáveis. O MM propõe a utilização de uma quan-

[3] Apenas para mencionar alguns métodos desenvolvidos por Pearson, há o Teste $\mathcal{X}^2$, o p-valor e Análise de Componentes Principais

tidade de momentos (normalmente os de ordem mais baixa) igual ao número de parâmetros a serem estimados, ajustando-os de tal forma que os momentos amostrais se igualem às suas contrapartes populacionais, em um sistema perfeitamente identificado.

De volta ao exemplo deste capítulo, como há apenas um parâmetro a ser estimado, o MM utiliza apenas o primeiro momento, a média, na estimação. No caso das Distribuições de Poisson, a média populacional é igual ao parâmetro λ. Portanto, a estimação é dada por:

$$\hat{\lambda}_{MM} = \frac{1}{N}\sum_{n=1}^{N} x_n = 3,1. \tag{3-4}$$

Neste caso, em particular, o estimador de MM coincide com os de MV e QMV.

Os estimadores de MM são consistentes sob condições pouco restritivas. Em pequenas amostras, porém, o MM não garante que os parâmetros estimados estarão dentro de uma região viável[4]. Apesar de ser mais facilmente computado, o MM perdeu espaço para MV ao longo do século XX[48]. Muitos pacotes estatísticos e econométricos utilizam os parâmetros estimados pelo MM como ponto inicial do processo iterativo de otimização da verossimilhança.

3.4 Método Generalizado dos Momentos

O Método Generalizado dos Momentos (MGM) foi introduzido em 1982 [36]. A generalização em relação ao MM se deve ao fato de possibilitar a utilização de mais condições de momentos do que o número de parâmetros a serem estimados. Não obstante, o MGM também generaliza diversos outros estimadores como, por exemplo, o de Mínimos Quadrados em Dois Estágios e MV.

Em muitas situações, a teoria econômica gera condições de momento e, em particular, condições de ortogonalidade. Essa foi a principal motivação apresentada no artigo original para a utilização do MGM, método que possibilita a utilização destas condições de momento diretamente na função de estimação.

Nos casos em que há mais condições de momento do que variáveis a serem estimadas, diz-se que há sobreidentificação do modelo. Este é o caso, por exemplo, quando há mais variáveis instrumentais do que regressores em um modelo com endogeneidade. Há, nestes casos, uma potencial impossibilidade de igualar os momentos populacionais implicados pelos parâmetros e os momentos

[4]Uma vez que o sistema de equações é perfeitamente identificado, a imposição de restrições é inócua ou inviabiliza a solução.

amostrais observados. O MGM lida com esta questão minimizando uma norma quadrática da distância entre essas duas grandezas.

Por se tratar de um pilar fundamental do presente trabalho, a definição e as propriedades do MGM serão apresentadas em mais detalhes e com mais formalismo estatístico em comparação aos demais métodos de estimação.

Definição 3.1 *Sejam Ω o conjunto de pontos amostrais em um espaço de probabilidades subjacente, e $E[\cdot]$ o operador esperança e $\{x_n\}_{n\geq 1}$ um processo estocático J-variado definido neste mesmo espaço de probabilidades, e $\{x_n(\omega_0)\}_{n=1}^{N}$ uma amostra de tamanho N obtida deste processo, para algum $\omega_0 \in \Omega$. Sejam, também, $\psi^* \in \Psi$ um vetor de parâmetros, onde $\Psi \subset \mathbb{R}^L$, e $f(\cdot)$ uma função definida em $(\mathbb{R}^J \times \Psi) \to \mathbb{R}^K$ tais que:*

$$E[f(x, \psi^*)] = \vec{0}. \tag{3-5}$$

Definindo-se a função $c(\psi) = N^{-1} \cdot \sum_{n=1}^{N} f(x_n(\omega_0), \psi)$ e a matriz de ponderação $W_N \in \mathbb{R}^{K\times K}$, a estimação pelo MGM é dada por:

$$\hat{\psi}_{MGM} = \underset{\psi\in\Psi}{\operatorname{argmin}}\, c(\psi)' \cdot W_N \cdot c(\psi). \tag{3-6}$$

Retomando o exemplo deste capítulo, temos, no caso da Distribuição de Poisson, que a média e a variância são iguais ao parâmetro λ, de onde conclui-se que a média do quadrados é dada por $\lambda^2+\lambda$. Pode-se usar estas duas condições de momento (média e média dos quadrados), cujos valores amostrais são $3,1$ e $11,7$, respectivamente, na estimação pelo MGM. Neste caso, tem-se que:

$$f(x, \lambda) = \left[\ x - \lambda, \quad x^2 - (\lambda^2 + \lambda)\ \right]. \tag{3-7}$$

Considerando-se W_N uma matriz identidade 2×2 e aplicando-se à equação (3-6), chega-se à formulação:

$$\hat{\lambda}_{MGM} = \arg\min_{\lambda\in\mathbb{R}^+} (3,1 \cdot \lambda)^2 + (11,7\lambda^2 - \lambda)^2 = 2,96. \tag{3-8}$$

Neste exemplo em particular, a estimação pelo MGM chegou mais próxima ao verdadeiro valor do parâmetro em comparação às estimações anteriormente apresentadas. Exemplos mais sofisticados da aplicação do MGM, contemplando dados transversais, séries temporais e dados em painel, pode ser encontrada em [49].

O MGM possui uma tríade de testes de hipótese sobre os parâmetros análoga à da MV: um teste baseado apenas no modelo irrestrito; um outro apenas com o modelo restrito; e um terceiro, que utiliza as duas estimações.

Estes testes, cujas estatísticas seguem, assintoticamente e sob a hipótese nula, uma distribuição $\mathcal{X}^2$, são detalhadamente descritos em [50, 51].

Uma característica amplamente reportada nos casos em que o MGM é aplicado em pequenas amostras é a existência de viés e o seu agravamento a medida em que o número de condições de momento aumenta. Este fenômeno é especialmente notório quando o MGM é aplicado no contexto de regressores endógenos com variáveis intrumentais [52]. Um exemplo de viés na estimação de um modelo CAPM por se encontrado em [53]. Não há demostrações formais gerais a respeito deste fenômeno, até porque cada problema de estimação tem características particulares, bem como as condições de momento utilizadas, de tal forma que não é correto afirmar que a inclusão de condições de momento acarreta, necessariamente, em um aumento do viés na estimação. Não obstante, o apêndice C traz um exemplo simples no qual o fenômeno pode ser observado.

A seguir, serão apresentadas as propriedades assintóticas dos estimadores de MGM.

3.4.1 Consistência

A consistência é uma das propriedades mais importantes de um estimador, uma vez que esta garante que, à medida que o tamanho da amostra tende a infinito, os parâmetros estimados convergem em probabilidade para os verdadeiros parâmetros. As condições suficientes para a consistência dos estimadores de MGM apresentadas em [54] são:

1. $W_N \xrightarrow{p} W$, onde W é positiva semi-definida;
2. $W \cdot E[f(x, \psi)] = \vec{0}$ somente se $\psi = \psi^*$;
3. Ψ é compacto;
4. $f(x, \psi)$ é contínuo para cada $\psi \in \Psi$ com probabilidade um;
5. $E\left[\sup_{\psi \in \Psi} \|f(x, \psi)\|\right] < \infty$.

Satisfeitas estas condições, garante-se que $\hat{\psi}_{MGM} \xrightarrow{p} \psi^*$. A segunda condição garante a identificabilidade global do modelo e, consequentemente, a existência de um único ponto de mínimo no problema de otimização. Porém, como é salientado em [54], em muitos casos, é difícil garantir esta condição com base em premissas mais primitivas do que a própria assunção de sua validade.

3.4.2 Normalidade Assintótica

A normalidade assintótica é uma propriedade desejável, uma vez que possibilita a realização de testes de hipótese baseados na distribuição assintótica dos estimadores. Novamente, [54] apresenta condições suficientes para esta propriedade:

1. o estimador é consistente;
2. ψ^* está no interior de Ψ;
3. $f(x,\psi)$ é continuamente diferenciável em uma vizinhança ϑ de ψ^* com probabilidade um;
4. $E[\|f(x,\psi^*)\|^2] < \infty$;
5. $E\left[\sup_{\psi\in\vartheta}\|\nabla_\psi f(x,\psi)\|\right] < \infty$;
6. $G \cdot W \cdot G'$ é não singular para $G = E\left[\nabla_\psi f(x,\psi)\right]$.

Satisfeitas estas condições, garante-se que:

$$\sqrt{N}(\hat{\psi}_{MGM} - \psi^*) \xrightarrow{d} \mathcal{N}[0, (G'WG)G'WVWG(G'WG)^{-1}]. \qquad (3\text{-}9)$$

onde $V = E[f(x,\psi^*)' \cdot f(x,\psi^*)]$.

3.4.3 Eficiência

Como foi visto na propriedade anterior, a variância assintótica do estimador de MGM depende da matriz W. Não é difícil demonstrar que a variância assintótica do estimador de MGM atinge o mínimo valor possível quando W é a inversa da matriz de covariância assintótica das condições de momento:

$$W^* = \sum_{k=-\infty}^{\infty} E[f(\vec{z},\psi^*).f(L^k(\vec{z}),\psi^*)']\,\big)^{-1},$$

onde $L(\cdot)$ denota o operador diferença[5]. Neste sentido, este é um estimador eficiente, não só entre os MGMs possíveis, mas em toda a classe de estimadores assintoticamente normais estimados com base nas mesmas condições de momento (sem uso de informações adicionais).

[5]O operador diferença é utilizado em séries temporais para mapear uma variável aleatória de um determinado período de tempo para o período imediatamente anterior. Para o caso de dados transversais, basta convencionar que o operador defasagem retorna zero.

Na prática, porém, para que $W_N \xrightarrow{p} W^*$ é preciso que W_N seja obtido com base em uma estimativa consistente de ψ^*. Há três abordagens na literatura para este problema. Uma comparação entre esses métodos baseada em experimentos de Monte Carlo pode ser encontrada em [53].

MGM Eficiente em Dois Estágios

A primeira abordagem, mais simples, consiste, simplesmente, em usar uma matriz W_N positiva semidefinida arbitrária (geralmente uma identidade), realizar a estimação do MGM, utilizar os parâmetros estimados para obter uma segunda matriz W'_N e, de posse desta nova matriz, proceder com a estimação definitiva dos parâmetros.

Aplicando esta metodologia ao exemplo, a partir da primeira estimação, onde $\hat{\lambda}_{MGM} = 2,96$, chega-se a:

$$W'_N = \begin{bmatrix} 15,45 & -2.32 \\ -2.32 & 0,36 \end{bmatrix}. \tag{3-10}$$

Inserindo W'_N na equação (3-6), obtém-se $\hat{\lambda}_{MGM} = 2,89$. É interessante notar que, neste caso, a utilização de uma matriz de ponderação obtida com um estimador consistente de λ afastou o valor estimado do verdadeiro valor.

MGM Eficiente Iterado

Nesta abordagem, o processo de obtenção da matriz de ponderação e reestimação dos parâmetros do MGM Eficiente em Dois Estágios é repetido iterativamente, até a convergência do valor do parâmetro ou o atingimento de outro critério de parada, como o número de iterações. No exemplo deste capítulo, a iteração converge na sétima reestimação (considerando a sexta casa decimal) para $\hat{\lambda}_{MGM} = 2,87$ (ainda mais afastada do verdadeiro valor).

MGM Eficiente Continuamente Atualizado

Nesta última abordagem, a obtenção do ponderador é inserida diretamente no problema de otimização[6]:

$$\hat{\psi}_{MGM} = \underset{\psi \in \Psi}{\operatorname{argmin}}\, c(\psi)' \cdot \left(N^{-1} \cdot \sum_{n=1}^{N} f(x_n, \psi)' \cdot f(x_n, \psi) \right) \cdot c(\psi). \tag{3-11}$$

No caso do exemplo explorado neste capítulo, este método chega exatamente ao mesmo valor que MGM Eficiente Iterado. Apesar de, no caso do

[6] A fórmula (3-11) utiliza uma estimativa da matriz de covariância para dados transversais ou série temporais independentes e homocedásticas. Para casos mais gerais, pode-se utilizar, por exemplo, o estimador proposto em [55].

exemplo apresentado, as iterações adicionais da matriz de ponderação terem levado a um pior resultado da estimação, trata-se de um caso particular que não pode servir de base para conclusões. Estudos mais aprofundados apresentados em [53] indicam que o MGM Eficiente Continuamente Atualizado possui melhores propriedades em amostras finitas.

3.4.4 Teste J

Nos casos em que o modelo é sobreidentificado, é possível (ou mesmo quase certo) que não haja nenhum vetor de parâmetros no espaço paramétrico para os quais os momentos populacionais se igualem exatamente às suas contrapartes obtidas com uma dada amostra. Porém, caso o modelo seja corretamente especificado, espera-se que a distância entre essas grandezas seja relativamente pequena.

A ideia por trás do Teste J, também proposto em [36], é justamente checar, dados os parâmetros estimados, o quão perto de zero está a distância entre as condições de momento populacionais e as contrapartes amostrais, e, consequentemente, testar se as premissas usadas na geração das condições de momento são válidas.

Formalmente, as hipóteses testadas são:

- Hipótese nula: $E[f(x, \psi^*)] = 0$, contra
- Hipótese alternativa: $E[f(x, \psi^*)] \neq 0$, $\forall\ \psi \in \Psi$.

A estatística do teste é dada por:

$$J \equiv \frac{1}{N} \cdot \left(\sum_{n=1}^{N} f(x_n, \hat{\psi}_{MGM}) \right)' \cdot W_N \cdot \left(\sum_{n=1}^{N} f(x_n, \hat{\psi}_{MGM}) \right). \tag{3-12}$$

Se W_N converge em probabilidade para W^*, temos que, sob a hipótese nula:

$$J \xrightarrow{d} \mathcal{X}^2_{K-L}. \tag{3-13}$$

Uma rejeição no Teste J é um indício de que o modelo não está corretamente especificado. Esta interpretação do teste é particularmente interessante nos casos em que as condições de momento derivam de algum modelo teórico, como é comum em economia e finanças, por exemplo. Nesses casos, a rejeição no Teste J indica que o modelo teórico não encontra suporte nos dados.

Parte da popularidade do MGM se deve ao Teste J, uma vez que não há testes similares igualmente versáteis para outros métodos de estimação. Não obstante, a baixa potência do Teste J em diversos tipos de má especificação do modelo foi reporda, por exemplo, em [56, 57].

4
Estimação de RNAs

Neste capítulo, será apresentada a aplicação dos métodos de estimação (Capítulo 3) às RNAs (Capítulo 2). Os métodos aqui apresentados podem ser aplicados a uma vasta gama de RNAs que utilizam o paradigma de aprendizagem supervisionada. Sem perda de generalidade, seguindo [24], será explorado o PMC com uma única camada escondida e ligação direta entre os regressores e a esperança condicional, dado por:

$$y_n = G(x_n;\psi) + \varepsilon_n = \vec{\alpha}' \cdot \vec{x}_n + \sum_{m=1}^{M} \{\lambda_m \cdot h(\vec{\omega}'_m \cdot \vec{x}_n)\} + \varepsilon_n, \text{ para } n = 1, \ldots, N. \tag{4-1}$$

onde y_n é a variável dependente, $G(x_n;\psi)$ é uma RNA parametrizada por $\psi = [\vec{\alpha}', \lambda_1, ..., \lambda_M, \vec{\omega}'_1, ..., \vec{\omega}'_M]$, $\vec{x}_n$ é um vetor de regressores (I regressores, incluindo uma constante igual a um), $h(\cdot)$ é a função de ativação logística e ε_n é um ruído, com média nula, exógeno em relação aos regressores. Com M neurônios na camada escondida, o número de parâmetros a serem estimados é $L = (M+1) \cdot I + M$. As variáveis y, $\vec{x}$ e ε são estacionárias e ergóticas.

A seguir, serão apresentadas a estimação por QMV, que é o método mais utilizado dentro das abordagens econométricas de RNAs, e a proposta alternativa introduzida no presente trabalho, a estimação pelo MGM.

4.1
Estimação por QMV

Os trabalhos pioneiros de Halbert White [21, 22] propuseram a estimação dos parâmetros utilizando o QMV, método para o qual ele fez contribuições muito relevantes [43].

Um ponto de atenção no caso da estimação de RNAs é a identificação do modelo. Segundo [24], o modelo (4-1) (irrestrito) não é local ou globalmente identificável, devido a três características: (i) a possibilidade de permutações entre as unidades escondidas mantendo o modelo inalterado; (ii) o fato de que $h(x) = 1 - h(-x)$, o que permite 2^M parametrizações equivalentes do mesmo modelo; e (iii) a possibilidade de haver unidades escondidas desnecessárias, uma vez que, dado que $\lambda_m = 0$, o resultado do modelo é o mesmo para qualquer vetor $\vec{\omega}_m$. Os dois primeiros problemas podem ser superados impondo-se as restrições $\lambda_m \leq \lambda_{m+1}$ para $m = 1, 2, ..., M-1$ e $\omega_{m,1} > 0$ para $m = 1, 2, ..., M$. Já o terceiro problema requer que a RNA não tenha unidades escondidas

redundantes e [24] apresenta uma estratégia de modelagem que evita este problema.

A estimação propriamente dita é realizada, simplesmente, pela aplicação do modelo (4-1) à equação (3-3):

$$\hat{\psi}_{QMV} = \underset{\psi\in\Psi}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - G(x_n;\psi))^2. \tag{4-2}$$

As condições de indentificabilidade determinam o conjunto de parâmetros possíveis Ψ.

4.1.1 Otimização

O problema de otimização de (4-2) pode ser dividido em dois subproblemas. Para um dado conjunto $\{\vec{\omega}\}_{m=1}^{M}$, os demais parâmetros podem ser estimados por Mínimos Quadrados Ordinários (MQO), com custo computacional baixíssimo.

Consequentemente, é conveniente definir o problema de estimação somente em termos de $\{\vec{\omega}\}_{m=1}^{M}$, com os demais parâmetros sendo definidos implicitamente, por MQO.

Como o custo de estimação por MQO é muito baixo, pode-se definir um processo em duas etapas para a otimização. Primeiro, busca-se um ponto inicial, gerando-se um grande número de $\{\vec{\omega}\}_{m=1}^{M}$ aleatoriamente e estimando-se os demais parâmetros por MQO. Partindo-se do ponto com melhor ajuste obtido na etapa anterior, inicia-se um processo de otimização local utilizando-se algum algoritmo de gradiente. Em [24], onde utiliza-se este processo em duas etapas, há uma forte recomendação para o uso dos algoritmos Broyden-Fletcher-Goldfarb-Shanno ou Levenberg-Marquardt, ambos descritos em [58]. Estes métodos não garantem a otimalidade global, que seria necessária para a validade das propriedades do estimador de QMV.

Otimalidade Global

A questão da otimização global em RNAs é abordado em [59–61]. Nestas abordagens, porém, o que se propõe é a utilização de meta-heurísticas de otimização sem gradiente (como busca tabu ou algoritmos genéticos) para uma busca abrangente no espaço de parâmetros, combinado com um método de gradiente para a busca local do ponto ótimo. Nenhum destes métodos, porém, oferece garantias de otimalidade global de fato.

Uma primeira contribuição deste trabalho é mostrar que, no caso de RNAs com funções de ativação do tipo rampa (ou linear por partes), é possível

obter-se garantia de otimalidade global. Estas funções de ativação respeitam as condições necessárias para o Teorema da Aproximação Universal apresentadas por [39]. Heurísticas para sua estimação de RNAs com este tipo de função de ativação podem ser encontrada em [62, 63]. A principal limitação desta função de ativação é o fato de ela não ser diferenciável em todos os pontos, apresentando descontinuidades na derivada.

A solução proposta no presente trabalho consiste em reescrever o problema de QMV (equação (3-3)) de uma RNA com função de ativação rampa como um problema de Programação Quadrática Inteira-Mista (PQIM). O algoritmo *Branch-and-Bound* [64] garante a otimalidade global em problemas de PQIM[1].

A ideia é utilizar variáveis binárias para indicar quando se atinge uma parte plana da função de ativação. Somente nestes casos o valor da função de ativação pode ser diferente do seu argumento.

Definindo a RNA com função de ativação rampa a ser estimada:

$$y_n = G(x_n;\psi) + \varepsilon_n = \alpha' \cdot \hat{x}_n + \sum_{m=1}^{M} \{\lambda_m \cdot \min(1; \max(-1, \omega'_m \cdot \vec{x}_n))\} + \varepsilon_n. \quad (4\text{-}3)$$

Como, no caso das funções de ativação rampa, $h(x) = -h(-x)$, pode-se impor as restrições $\lambda_m \geq 0$ para $m = 1, 2, ..., M$, logo, pode-se reescrever:

$$y_n = \alpha' \cdot \vec{x}_n + \sum_{m=1}^{M} \min(\lambda_m; \max(-\lambda_m, (\lambda_m \cdot \omega'_m) \cdot \hat{x}_n)) + \varepsilon_n. \quad (4\text{-}4)$$

A representação gráfica da função de ativação rampa reparametrizada encontram-se na figura 4.1.

Aplicando-se esta formulação ao problema de otimização da equação (3-3), pode-se obter a formulação em PQIM.

[1]De fato, a otimalidade global só é garantida nos casos em que a matriz da parte quadrática da função objetivo é positiva definida, como será o caso na formulação aqui apresentada. Além disso, para ser mais preciso, esta garantia obtida é o fechamento do chamado *gap* de otimalidade, que é a diferença entre uma relaxação do problema original (por construção, melhor ou igual à melhor solução viável) e a melhor solução viável já encontrada durante as iterações do algoritmo. Apesar de ser um método ampla e exitosamente utilizado, os problemas para os quais o *Branch-and-Bound* foi desenvolvido (otimização discreta) são da classe dos NP-complexos. Isso significa que, nos piores casos, estes problemas não podem ser resolvidos em tempo polinomial ou, em outras palavras, o número (finito) de passos necessários para se chegar ao fim do algoritmo pode crescer mais do que polinomialmente com o tamanho do problema.

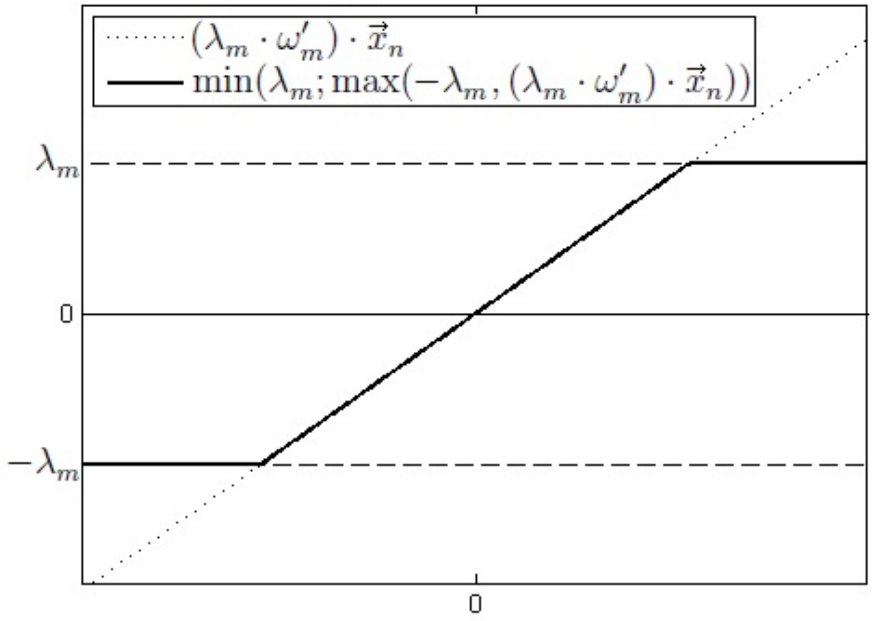

Figura 4.1: Função de Ativação Rampa Reparametrizada.

$$\min_{\substack{\{\alpha_i\}_{i=1}^{I},\{\lambda_m,\{\bar{\omega}_{i,m}\}_{i=0}^{I}\}_{m=1}^{M},\\ \{\varepsilon_n,\{p_{n,m},o_{n,m},b^+_{n,m},b^-_{n,m}\}_{m=1}^{M}\}_{n=1}^{N}}} \quad \sum_{n=1}^{N} \varepsilon_n^2 \tag{4-5}$$

Sujeito a:

$$y_n - \sum_{i=1}^{I} \alpha_i \cdot x_{i,n} - \sum_{m=1}^{M} o_{n,m} = \varepsilon_n \ \forall \ n = 1, ..., N; \tag{4-6}$$

$$p_{n,m} = \sum_{i=1}^{I} \bar{\omega}_{i,m} \cdot x_{i,n}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-7}$$

$$p_{n,m} + \lambda_m \leq \mu \cdot (1 - b^-_{n,m}), \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-8}$$

$$o_{n,m} + \lambda_m \leq \mu \cdot (1 - b^-_{n,m}), \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-9}$$

$$p_{n,m} + \lambda_m \geq -\mu \cdot b^-_{n,m}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-10}$$

$$p_{n,m} - \lambda_m \geq \mu \cdot (1 - b^+_{n,m}), \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-11}$$

$$o_{n,m} - \lambda_m \geq \mu \cdot (1 - b^+_{n,m}), \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-12}$$

$$p_{n,m} - \lambda_m \leq \mu \cdot b^+_{n,m}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-13}$$

$$p_{n,m} - o_{n,m} \geq -\mu \cdot b^-_{n,m}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-14}$$

$$p_{n,m} - o_{n,m} \leq \mu \cdot b^+_{n,m}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-15}$$

$$-\lambda_m \leq o_{n,m}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-16}$$

$$o_{n,m} \leq \lambda_m, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-17}$$

$$\lambda_1 \geq 0, \ \forall \ m = 1, ..., M; \tag{4-18}$$

$$\lambda_m \geq \lambda_{m+1}, \ \forall \ m = 1, ..., (M-1); \tag{4-19}$$

$$b^-_{n,m} \in \{0, 1\}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M; \tag{4-20}$$

$$b^+_{n,m} \in \{0, 1\}, \ \forall \ n = 1, ..., N, \ m = 1, ..., M. \tag{4-21}$$

onde $\{\lambda_m, \{\bar{\omega}_{i,m}\}_{i=0}^{I}\}_{m=1}^{M}$ são os parâmetros da RNA (reparametrizada com $\bar{\omega}_{i,m} \equiv \lambda_m \cdot \omega_{i,m}$), $\{\varepsilon_n\}_{n=1}^{N}$ são os resíduos, $\{\{p_{n,m}\}_{m=1}^{M}\}_{n=1}^{N}$ e $\{\{o_{n,m}\}_{m=1}^{M}\}_{n=1}^{N}$ são os argumentos e as saídas (respectivamente, a linha pontilha e a linha sólida grossa da figura 4.1) das M unidades escondidas para cada uma das N observações, enquanto $\{\{b_{n,m}^{+}, b_{n,m}^{-}\}_{m=1}^{M}\}_{n=1}^{N}$ são as variáveis binárias, um para para cada observação em cada unidade escondida, que controlam se as funções de ativação estão nas regiões planas, negativa ou positiva, respectivamente e, finalmente, μ é um valor constante arbitrariamente grande, conhecida na literatura de otimização como "*Big M*".

As restrições (4-6) geram os resíduos ε_n, enquanto (4-7) geram os argumentos $p_{n,m}$ da função de ativação. As restrições (4-8), (4-9) e (4-10) garantem que a variável binária $b_{n,m}^{-}$ será igual a um se e somente se o argumento da função de ativação ($p_{n,m}$) for menor ou igual a $-\lambda_m$ e que a saída da função de ativação ($o_{n,m}$) será maior que o argumento somente se $b_{n,m}^{-} = 1$. As restrições (4-11), (4-12) e (4-13) funcionam de forma análoga para as variáveis binárias $b_{n,m}^{+}$, que controlam a região plana da função de ativação quando o argumento é maior que λ_m. As restrições (4-14) e (4-15) garantem que a saída da função de ativação iguale o seu argumento sempre que $b_{n,m}^{-} = b_{n,m}^{+} = 0$. As restrições (4-16) e (4-17) restringem o valor da função de ativação ao intervalo $[-\lambda_m, \lambda_m]$. As condições de indentificabilidade estão em (4-18) e (4-19), enquanto (4-20) e (4-21) fazem com que as variáveis $\{\{b_{n,m}^{+}, b_{n,m}^{-}\}_{m=1}^{M}\}_{n=1}^{N}$ sejam binárias.

Para recuperar os parâmetros da formulação original (equação (4-3)), deve-se aplicar a transformação $\omega_{i,m} = \bar{\omega}_{i,m} \cdot (\lambda_m)^{-1}$. Além disso, com as devidas transformações nos parâmetros, este método pode ser utilizado para a geração do ponto inicial para RNAs com funções de ativação logística ou tangente hiperbólica, caso esta seja a opção feita pelo econometrista.

Os estudos realizados indicam que, utilizando *softwares* comerciais de otimização e computadores domésticos[2], a convergência do algoritmo *Branch-and-Bound* para instâncias pequenas ($N = 50$ e $M = 2$) ocorre em tempo razoável (menos de um minuto). Porém, o tempo de convergência cresce vertiginosamente com o aumento do tamanho do problema, o que torna o método pouco adequado a aplicações massivas e com instâncias maiores, como as realizadas no presente trabalho. Não obstante, este resultado é importante por ser uma primeira abordagem capaz de proporcionar garantias de otimalidade global no problema em questão, e pode servir de base para futuros desenvolvimentos, por exemplo, utilizando técnicas de decomposição ou incorporação de heurísticas dentro do *Branch-and-Bound*, que venham a

[2]Processador Intel Pentiun i5 com 8 Gb de memória RAM.

viabilizar a estimação de RNAs com mais parâmetros e observações.

4.1.2 Sobreajuste em RNAs Estimadas por QMV

Como foi explicado anteriormente, um dos principais desafios na utilização de RNAs é a ocorrência de sobreajuste. Esta questão, porém, não costuma ser abordada com o devido formalismo. Para possibilitar um entendimento mais profundo deste fenômeno, em particular para RNAs estimadas por QMV, será apresentado a seguir um limite inferior aproximado para a probabilidade de existência de uma solução com sobreajuste.

O primeiro desafio é a inexistência de uma definição rigorosa (do ponto de vista matemático) do que vem a ser o sobreajuste. Porém, partindo-se dos sintomas comumente mencionados (ajuste bom dentro da amostra e ruim fora dela), definiremos uma solução com sobreajuste como sendo $\psi^{SA} \in \Psi$ tal que:

1. $\sum_{n=1}^{N}(y_n - G(x_n;\psi^{SA}))^2 \leq \sum_{n=1}^{N}(y_n - G(x_n;\psi^{*}))^2$, ou seja, a solução com sobreajuste possui um ajuste dentro da amostra tão bom quanto ou superior ao obtido pelos verdadeiros parâmetros; e

2. $E[y \cdot G(x;\psi^{SA})] = E[y] \cdot E[G(x;\psi^{SA})]$, ou seja, a variável dependente é ortogonal a $G(x;\psi^{SA})$.

Esta definição é bastante restritiva, especialmente por conta da segunda restrição determina que o poder de explicação da RNA com sobreajuste sobre a variável dependente seja nulo, em linha com o sintoma de ajuste ruim fora da amostra. Pela primeira condição, referente ao sintoma de bom ajuste dentro da amostra, pode ser justificada de duas formas. Em primeiro lugar, como o próprio termo sobreajuste sugere, o fenômeno ocorre nos casos em que se obtém um nível de ajuste além do ideal, e o ajuste obtido pelos verdadeiros parâmetros é uma boa medida do que seria o ideal. Além disso, para que o sobreajuste aconteça de fato, é preciso que a solução com sobreajuste seja preferida à verdadeira solução pelo critério da QMV, caso contrário, o processo de estimação por QMV escolherá um vetor de parâmetros na região do verdadeiro.

A estratégia utilizada para derivar um limite inferior para a probabilidade de ocorrência de uma solução da QMV com sobreajuste terá três etapas: (i) definir o processo gerador de dados; (ii) definir uma solução candidata à sobreajuste que, por construção, atenda à segunda condição para tal; e (iii) calcular a probabilidade aproximada de a solução candidata atender à primeira condição.

Considere-se, sem perda de generalidade, o processo gerador de dados como sendo uma RNA $G(x, \psi^*)$, com $I = 2$ (uma constante igual a um e uma variável aleatória contínua), $M = 2$ e função de ativação logística $h(\cdot)$, com os verdadeiros parâmetros dados por $\psi^* \in \Psi$, tal qual está descrito na equação (4-1). Assumindo-se que as variáveis y e ε seguem distribuições normais e que foram reescaladas de tal forma que $y \sim \mathcal{N}_{i.i.d.}[\mu, 1]$ (incondicionalmente) e que $\varepsilon \sim \mathcal{N}_{i.i.d.}[0, q]$[3], onde $q < 1$ é denominada razão de ruído, quociente da variância do ruído pela variância da variável dependente. Por fim, assume-se que ε é ortogonal a $G(x, \psi^*)$.

Como candidata à solução com sobreajuste, definiremos um conjunto de parâmetros para os quais a RNA aproxima arbitrariamente bem uma função que seja constante em todo o domínio, exceto em um ponto, e que, portanto, atende à segunda condição da definição do sobreajuste. Sejam $\hat{\mu}$ constante e $\hat{n}$ o índice tal que $|y_{\hat{n}} - \hat{\mu}| \geq |y_n - \hat{\mu}| \forall n \neq \hat{n}$, definimos a função que será aproximada como:

$$H(x, \hat{\mu}) = \begin{cases} \hat{\mu}, & \text{quando x} \neq x_{\hat{n}}; \\ y_{\hat{n}}, & \text{quando x} = \text{x}_{\hat{n}}. \end{cases} \tag{4-22}$$

Ou seja, a função $H(x, \hat{\mu})$ assume o valor constante $\hat{\mu}$ sempre, exceto no ponto com maior distância entre $\hat{\mu}$ e y_n, ou seja, a função elimina o maior desvio da amostra em relação à constante.

Definindo-se, agora, as constantes $k_1 > 0$, $k_2 > 0$, os parâmetros candidatos à sobreajuste ψ^{CSA} são dados por:

$$\begin{aligned} \vec{\alpha}^{CSA} &= \left[\begin{array}{cc} \hat{\mu}, & 0 \end{array} \right]', \\ \vec{\lambda}^{CSA} &= \left[\begin{array}{cc} \dfrac{y_{\hat{n}} - \hat{\mu}}{h(k_1) - h(-k_1)}, & -\dfrac{y_{\hat{n}} - \hat{\mu}}{h(k_1) - h(-k_1)} \end{array} \right]', \\ \vec{\omega}_1^{CSA} &= \left[\begin{array}{cc} -k_2 \cdot x_{\hat{n},2} + k_1, & k_2 \end{array} \right]', \\ \vec{\omega}_2^{CSA} &= \left[\begin{array}{cc} k_2 \cdot x_{\hat{n},2} + k_1, & -k_2 \end{array} \right]'. \end{aligned}$$

Não é difícil demonstrar que, quando $k_1 \to 0$, $k_2 \to \infty$ e $k_1 \cdot k_2 \to \infty$ $G(x, \psi^{CSA}) \to H(x, \hat{\mu})$ (convergência uniforme). Consideraremos os casos limites das constantes k_1 e k_2 para os quais $G(x, \psi^{CSA}) = H(x, \hat{\mu})$. Os parâmetros ψ^{CSA} atendem à segunda condição da definição de solução com sobreajuste, uma vez que $E[y \cdot G(x; \psi^{CSA})] = E[y] \cdot \hat{\mu} = E[y] \cdot E[G(x; \psi^{CSA})]$.

Por fim, resta saber qual é a probabilidade P^{SA} de que ψ^{CSA} atenda, para algum valor de $\hat{\mu}$, à primeira condição da definição de sobreajuste, ou

[3]Implicitamente, determina-se que $G(x, \psi^*) \sim \mathcal{N}_{i.i.d.}[\mu, 1 - q]$.

seja:

$$P^{SA} \equiv P\left[\sum_{n=1}^{N}(y_n - G(x_n;\psi^{CSA}))^2 \leq \sum_{n=1}^{N}(y_n - G(x_n;\psi^*))^2\right]. \quad (4\text{-}23)$$

No caso da RNA com os verdadeiros parâmetros $\psi*$, tem-se que:

$$\sum_{n=1}^{N}(y_n - G(x_n;\psi^*))^2 = \sum_{n=1}^{N}\varepsilon_n^2 \equiv SQR^{\psi^*}. \quad (4\text{-}24)$$

Já para a RNA candidata a sobreajuste, escolhendo-se $\hat{\mu} = \mu$, que seria uma das escolhas sub-ótimas possíveis no processo de otimização da QMV, e definindo-se $\tilde{y}_n = G(x_n;\psi^*) - \hat{\mu}$, tem-se:

$$\sum_{n=1}^{N}(y_n - G(x_n;\psi^{CSA}))^2 = \sum_{n=1}^{N}(\tilde{y}_n + \varepsilon_n)^2 - \max\left\{(\tilde{y}_n + \varepsilon_n)^2\right\}_{n=1}^{N} \equiv SQR^{\psi^{CSA}}. \quad (4\text{-}25)$$

.

Abrindo-se $(\tilde{y}_n + \varepsilon_n)^2$, chega-se a:

$$SQR^{\psi^{CSA}} - SQR^{\psi^*} = \left[\sum_{n=1}^{N}\left\{\tilde{y}_n^2 + 2\cdot\tilde{y}_n\cdot\varepsilon_n\right\} - \max\left\{(\tilde{y}_n + \varepsilon_n)^2\right\}_{n=1}^{N}\right]. \quad (4\text{-}26)$$

.

Ignorando-se os termos $\tilde{y}_n \cdot \varepsilon_n$, que seguem uma distribuição com muita concentração de massa em zero[4], obtém-se a relação aproximada:

$$P^{SA} \approx P\left(\sum_{n=1}^{N}\tilde{y}_n^2 \leq \max\left\{(\tilde{y}_n + \varepsilon_n)^2\right\}_{n=1}^{N}\right). \quad (4\text{-}27)$$

Porém, a probabilidade de uma realização de uma variável aleatória ser menor ou igual ao máximo dentre um conjunto de realizações de outras variáveis aleatórias é igual a um menos a probabilidade de ser maior que todas as realizações do conjunto. Logo, obtém-se:

$$P^{SA} \approx 1 - P\left(\sum_{n=1}^{N}\tilde{y}_n^2 > (\tilde{y}_n + \varepsilon_n)^2, \forall n = 1, 2, ..., N\right). \quad (4\text{-}28)$$

Definido $p = 1 - q$, chega-se a:

$$P\left(\sum_{n=1}^{N}\tilde{y}_n^2 > (\tilde{y}_n + \varepsilon_n)^2\right) = P\left(p\cdot\sum_{n=1}^{N}\left(\frac{\tilde{y}_n}{p^{1/2}}\right)^2 > (\tilde{y}_n + \varepsilon_n)^2\right). \quad (4\text{-}29)$$

Valendo-se da independência de $\tilde{y}$ e ε (tanto entre as diferentes observações de cada uma das variáveis, quanto cruzada entre as duas variáveis), sabe-se que:

[4]Distrubuição Produto de Gaussianas, ambas com média nula.

$$\sum_{n=1}^{N}\left(\frac{\tilde{y}_n}{p^{1/2}}\right)^2 \sim \mathcal{X}_N^2. \tag{4-30}$$

Da mesma forma, tem-se que:

$$(\tilde{y}_n + \varepsilon_n)^2 \sim \mathcal{X}_1^2. \tag{4-31}$$

Logo, ignorando-se a correlação entre $\sum_{n=1}^{N}\tilde{y}_n^2$ e $(\tilde{y}_n + \varepsilon_n)^2$, chega-se ao resultado:

$$P^{SA} \approx 1 - \left(\int_0^{\infty} F_1(p\cdot z)\cdot f_N(z)\ dz\right)^N, \tag{4-32}$$

onde $F_1(\cdot)$ é a função densidade acumulada da distribuição $\mathcal{X}_1^2$ com um grau de liberdade, enquanto $f_N(\cdot)$ denota a função densidade da distribuição $\mathcal{X}_N^2$, com N graus de liberdade.

Percebe-se na equação (4-32) que P^{SA} é inversamente proporcional a p e, consequentemente, diretamente proporcional a q, a razão de ruído. Este resultado é bastante intuitivo, uma vez que quanto maior a magnitude dos ruídos, maior o ganho em ajustá-los. Nota-se também que, para um dado N, P^{SA} se aproxima de um quando p tende a zero. A tabela 4.1 apresenta os valores de P^{SA} para valores de N e q selecionados.

Tabela 4.1: Valores de P^{SA} (Probabilidade de Sobreajuste)

	N						
q	50	100	200	400	800	1600	3200
10%	0%	0%	0%	0%	0%	0%	0%
20%	0%	0%	0%	0%	0%	0%	0%
30%	0%	0%	0%	0%	0%	0%	0%
40%	0%	0%	0%	0%	0%	0%	0%
50%	0,02%	0%	0%	0%	0%	0%	0%
60%	0,16%	0%	0%	0%	0%	0%	0%
70%	1,18%	0%	0%	0%	0%	0%	0%
80%	10,34%	0,17%	0%	0%	0%	0%	0%
90%	74,18%	17,06%	0,23%	0%	0%	0%	0%
95%	99,75%	92,86%	29,07%	0,38%	0%	0%	0%
99%	100%	100%	100%	100%	97,77%	9,96%	0,01%
100%	100%	100%	100%	100%	100%	100%	100%

A tabela 4.1 mostra que, como seria de se esperar, um aumento no número de observações reduz P^{SA}. Outra constatação interessante é que pequenas variações na razão de ruído podem impactar significativamente o P^{SA}. Observa-se este efeito, por exemplo, na segunda coluna ($N = 50$) para $50\% \leq q \leq 90\%$. Para cada 10% de aumento em q, P^{SA} é aproximadamente multiplicada por 7.

Pode-se dizer que P^{SA} é um limite inferior para a probabilidade de existência de uma solução com sobreajuste porque: (i) considera apenas uma forma extremamente particular de sobreajuste; (ii) ignora a possibilidade de otimização de $\hat{\mu}$, que ocorreria na estimação por QMV, fixando-o de maneira *ad hoc* em μ; e (iii) utiliza uma definição bastante restritiva em relação ao quão baixo deve ser ajuste fora da amostra para caracterizar o sobreajuste, limitando-o a zero. Além disso, P^{SA} é uma aproximação, por conta das várias hipóteses simplificadoras e aproximações aplicadas ao longo da sua derivação, mas não deixa de ser um limite inferior válido nos casos em que $M > 2$ e/ou $I > 2$.

Apesar destas limitações, P^{SA} é útil na prevenção de estimações em situações nas quais há alta probabilidade de ocorrência de sobreajuste ou mesmo para definição de um tamanho de amostra mínimo para uma dada razão de ruído e máximo P^{SA} determinados *a priori*. Outro ponto interessante da derivação de P^{SA} é desvincular o conceito de sobreajuste da presença de unidades escondidas excessivas ou ociosas. Mesmo nos casos em que a RNA possui exatamente o número de unidades escondidas necessário para ajustar a verdadeira esperança condicional, a probabilidade de existência de uma solução com sobreajuste pode ser estritamente positiva e, até mesmo, próxima de 100%, para valores de N pequenos combinados a valores de q grandes. Isso significa que uma boa escolha de arquitetura da RNA pode ser insuficiente para evitar o sobreajuste em RNAs estimadas por QMV.

4.1.3 O Teste Lee-White-Granger para Não Linearidade Negligenciada

Os TNLNs, como sugere o nome, são utilizados para testar a hipótese de que a esperança condicional da variável dependente é uma função linear das variáveis explicativas ou se, alternativamente, há indícios de não linearidade. Consequentemente, os TNLN podem ser usados como critério para decisão entre uma modelagem linear ou não linear. De maneira geral, as hipóteses de um TNLN são:

- Hipótese nula: $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) = 1$ para algum $\vec{\alpha} \in \mathbb{R}^I$, contra
- Hipótese alternativa: $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) < 1$ para todo $\vec{\alpha} \in \mathbb{R}^I$.

Mais adiante, o presente trabalho apresentará um novo TNLN e, para fins de comparação, faz-se necessário apresentar um TNLN consagrado na literatura. Um dos TNLNs mais amplamente utilizados é o Teste Lee-White-Granger (LWG) [23, 26]. Em [26, 27, 30], o LWG foi comparado com diversos

outros TNLNs[5], para diversas formas de não linearidade, e obteve resultados bastante satisfatórios em termos de tamanho e potência. Em suma, o LWG presta-se perfeitamente às necessidades do presente trabalho[6].

Basicamente, o LWG testa se os resíduos do modelo linear não são explicados por uma conjunto de funções não lineares dos regressores. Estão funções são componentes principais (combinações lineares) de transformações não lineares dos regressores gerada aleatoriamente. A ideia por detrás deste teste é simples e pode ser melhor compreendida a partir do seu procedimento:

- estima-se o modelo linear e toma-se os resíduos $\{\varepsilon_n\}_{n=1}^N$
- gera-se, aleatoriamente, uma determinada quantidade D de vetores $\vec{\omega}_d$, de dimensão $J \times 1$;
- calcula-se $\{\{h(\vec{x}_n \cdot \vec{\omega}_d)\}_{n=1}^N\}_{d=1}^D$, onde $h(\cdot)$ é a função logística;
- toma-se um subconjunto $\{\vec{z}_n\}_{n=1}^N$ dos componentes principais de $\{\{h(\vec{x}_n \cdot \vec{\omega}_d\}_{n=1}^N\}_{d=1}^D$ (ou os q^* primeiros ou os q^* primeiros desconsiderando-se o primeiro);
- calcula-se:

$$LWG = \left(N^{\frac{1}{2}} \cdot \sum_{n=1}^{N} \vec{z}_n \cdot \varepsilon_n\right) \cdot V \cdot \left(N^{\frac{1}{2}} \cdot \sum_{n=1}^{N} \vec{z}_n \cdot \varepsilon_n\right)', \tag{4-33}$$

  onde V é a matriz de covariância assintótica de $\vec{z} \cdot \varepsilon$.

Sob a hipótese nula, LWG segue uma distribuição $\mathcal{X}^2_{q^*}$.

Para evitar a estimação de V, [26] propõe a utilização da estatística assintoticamente equivalente dada por:

$$LWG' = N \cdot \frac{\sum_{n=1}^{N} \bar{\varepsilon}_n^2}{\sum_{n=1}^{N} \varepsilon_n^2}, \tag{4-34}$$

onde $\{\bar{\varepsilon}_n\}_{n=1}^N$ são os resíduos da regressão linear dos resíduos da RNAs em $\{\vec{z}_n\}_{n=1}^N$.

Em [26], os vetores $\vec{\omega}_d$ são sorteados de distribuições uniformes entre $[-2, 2]$ e sugere q^* igual a dois ou três. [85] demonstra que o teste possui melhores propriedade quando a quantidade de vetores aleatórios gerados é grande, da ordem de mil.

[5]Em [27], o LWG foi testado contra o Teste Saikkonen-Luukkonen, em [26] contra os testes Bispectrum, RESET, de Keenan, de Tsay, de McLeod-Li, de Brock-Dechert-Scheinkman e de White (matriz de informação dinâmica). Já em [30], o LWG é testado contra dois testes que utilizam regressões não-paramétricas. Em todos os casos, o LWG obteve bons resultados nos experimentos de simulação.

[6]O LWG foi descrito nesta seção porque é compatível com a estimação por QMV.

4.2 Estimação pelo MGM

Nesta seção, serão apresentadas a principal contribuição do presente trabalho, que vem a ser o arcabouço geral para estimações dos parâmetros de RNAs utilizando o MGM, e duas contribuições secundárias, que são o caso particular do arcabouço geral e o novo TNLN, divididos nas subseções a seguir.

4.2.1 Arcabouço Geral

Pretende-se apresentar um arcabouço geral para a estimação de RNAs utilizando o MGM sobreidentificado. As mesmas questões relativas a múltiplas parametrizações e, consequentemente, a identificabilidade do modelo, apresentadas para o caso da estimação por QMV, precisam ser tratadas neste caso. Porém, para a estimação pelo MGM, imporemos uma restrição diferente aos parâmetros:

- $0 < \epsilon \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_M$, para um ϵ positivo e arbitrariamente pequeno;
- o primeiro elemento diferente de zero no vetor $(\vec{\omega}_m - \vec{\omega}_n)$ é menor ou igual a $-\epsilon^*$ para um ϵ^* positivo e arbitrariamente pequeno, sempre que $\lambda_m = \lambda_n$ e $1 \leq m < n \leq M$.

O motivo da preferência por estas condições de identificabilidade ficará claro mais adiante.

O primeiro passo na construção do estimador de MGM é a determinação das condições de momento utilizadas ou, na notação da definição 3.1, determinar a função $f(\cdot)$ de $(\mathbb{R}^J \times \Psi) \rightarrow \mathbb{R}^K$.

Um primeiro conjunto de L condições de momento pode ser obtido a partir do problema de otimização da QMV, apresentado na equação (3-3). As condições de primeira ordem deste problema são dadas por:

$$\nabla_\psi \sum_{n=1}^{N} (y_n - G(\vec{x}_n; \psi))^2 = \vec{0}, \tag{4-35}$$

Esta condição equivale a:

$$\sum_{n=1}^{N} (y_n - G(\vec{x}_n; \psi)) \cdot \nabla_\psi G(\vec{x}_n; \psi) = \vec{0}. \tag{4-36}$$

Isso significa que se espera que os resíduos sejam ortogonais ao gradiente de $G(\cdot; \psi)$ em relação a ψ. Esta condição de primeira ordem pode, portanto, ser interpretada como uma condição de momento L-dimensional da forma:

$$E[(y - G(\vec{x};\psi)) \cdot \nabla_{\psi} G(\vec{x};\psi)'] = \vec{0}. \tag{4-37}$$

Um estimador de MGM perfeitamente identificado, apenas com estas condições de momento, seria equivalente a um estimador de QMV.

Definindo agora a função $d(\cdot)$ como:

$$d(\vec{x};\{\vec{\omega}_m\}_{m=1}^{M}) = \begin{bmatrix} \nabla_{\vec{\alpha}} G(\vec{x};\psi) \\ \nabla_{\vec{\lambda}} G(\vec{x};\psi) \\ \lambda_1^{-1} \cdot \nabla_{\vec{\omega}_1} G(\vec{x};\psi) \\ \vdots \\ \lambda_M^{-1} \cdot \nabla_{\vec{\omega}_M} G(\vec{x};\psi) \end{bmatrix}, \tag{4-38}$$

e utilizando as condições de identificabilidade $0 < \epsilon \leq \lambda_1 \leq \lambda_1 \leq ... \leq \lambda_M$, conclui-se que a condição (4-37) será válida se e somente se

$$E[(y - G(\vec{x};\psi)) \cdot d(\vec{x};\{\vec{\omega}_m\}_{m=1}^{M})'] = \vec{0}. \tag{4-39}$$

Portanto, as condições (4-37) e (4-39) são equivalentes. Por ser mais conveniente para fins de otimização, como será visto mais a diante, serão utilizadas as condições (4-39).

A fim de se obter a sobreidentificação do MGM, precisa-se de mais condições de momento válidas. Para tal, recorremos à hipótese de exogeneidade estrita dos resíduos em relação às variáveis explicativas:

$$E[\varepsilon \mid \vec{x}] = 0, \ \forall \vec{x} \in \mathbb{R}^I. \tag{4-40}$$

Esta condição implica que $E[\varepsilon \cdot l(\vec{x})] = \vec{0}, \ \forall \vec{x} \in \mathbb{R}^I$, para qualquer função Borel-mensurável $l(\cdot)$ de $\mathbb{R}^I \rightarrow \mathbb{R}^{K-L}$. Portanto, o único ingrediente necessário para se criar uma condição de momento válida é uma função Borel-mensurável $l(\cdot)$ de $\mathbb{R}^I\mathbb{R}$. Em particular, formas funcionais flexíveis, como os polinômios, polinônios ortogonais, *splines*, *wavelets* e RNAs são candidatos naturais para desempenhar este papel, bem como algumas formas funcionais utilizadas na literatura de Testes de Diferença Martingal, como exponencial de combinações lineares e funções indicador.

Definido agora, para uma função Borel-mensurável genérica $l(\cdot)$ de $\mathbb{R}^I \rightarrow \mathbb{R}^{K-L}$,

$$d^*(\vec{x};\{\vec{\omega}_m\}_{m=1}^{M}) = \begin{bmatrix} \nabla_{\vec{\alpha}} G(\vec{x};\psi) \\ \nabla_{\vec{\lambda}} G(\vec{x};\psi) \\ \lambda_1^{-1} \cdot \nabla_{\vec{\omega}_1} G(\vec{x};\psi) \\ \vdots \\ \lambda_M^{-1} \cdot \nabla_{\vec{\omega}_M} G(\vec{x};\psi) \\ l(\vec{x}) \end{bmatrix}, \tag{4-41}$$

podemo-se definir as condições de momento do MGM sobreidentificado como:

$$f(\vec{x}, \psi) = (y - G(\vec{x}; \psi)) \cdot d^*(\vec{x}_n; \{\vec{\omega}_m\}_{m=1}^M) = \vec{0}. \quad (4\text{-}42)$$

e aplicá-las na fórmula (3-6), utilizando um ponderador W_N eficiente estimado por um dos métodos apresentados na seção 3.4.

É importante notar que estas condições de momentos adicionais são assintoticamente redundantes, no sentido de não reduzirem a variância assintótica dos parâmetros estimados. Alguns estudos, como [66, 67], sugerem que a inclusão de condições de momentos assintoticamente redundantes tende a deteriorar a estimativas em amostra finita. Estes estudos se concentram em condições de ortogonalidade entre variáveis instrumentais (que não dependem dos parâmetros) e os resíduos. No nosso caso de interesse, $d(\vec{x}; \{\vec{\omega}_m\}_{m=1}^M)$ depende dos parâmetros, enquanto $l(\vec{x})$ não depende. A informação contida em $l(\vec{x})$ só é irrelevante quando $\{\vec{\omega}_m\}_{m=1}^M$ se está de $\{\vec{\omega}_m^*\}_{m=1}^M$. Em pequenas amostra ou com alto ruído, é possível que esta distância seja relativamente grande. Por conta disso, as conclusões sobre o efeito de condições de momentos redundantes pode ser inválido para este caso. Esta conjectura está em linha com os resultados das experiências de Monte Carlo apresentados na próxima seção.

Neste ponto, vale a pena expor as razões pelas quais se espera que este arcabouço seja capaz de reduzir a suscetibilidade das RNAs ao sobreajuste. O primeiro deles está no fato de existirem condições de momento adicionais que, indiretamente, reduzem a liberdade da RNA para ajustar parte do ruído. A RNA, ao buscar satisfazer um número d condições de momento maior do que o número de parâmetros teria uma margem menor para buscar ruídos distantes da esperança condicional. O segundo, o MGM rompe com o paradigma de minimização dos ruídos, estritamente relacionado com a ocorrência de sobreajuste, e o substitui pelo da maximização da ortogonalidade. A magnitude dos ruídos deixa de ser relevante. Os resultados apresentados na seção seguinte estão alinhados com a previsão decorrente destas conjecturas.

Consistência e Normalidade Assintótica

Utilizando as condições para consistência do MGM apresentadas na seção 3.4 para o caso das RNAs estimadas pelo MGM[7] segundo o arcabouça geral proposto, pode-se dizer que:

1. $W_N \xrightarrow{p} W$**, onde** W **é positiva semi-definida**:
   - Para que esta condição seja atendida, nos casos em que se utiliza alguma das versões do MGM Eficiente descritas na seção 3.4,

[7] A condições serão exploradas na notação da seção 3.4, onde o vetor x engloba a variável dependente e os regressores.

basta que não haja nenhuma condição de momento que seja uma combinação linear de duas ou mais outras. Com isso, garante-se que a estimação da matriz V será positiva definida, assim como sua inversa, para qualquer N;

2. $W \cdot E[f(x,\psi)] = \vec{0}$ **somente se** $\psi = \psi^*$:

   – As condições de identificabilidade impostas garantem que existe uma única parametrização que satizfaz as condições de momento derivadas da QMV. Valendo-se da positividade definida do ponderador (obtida na condição anterior) pode-se garantir que esta condição seja satisfeita;

3. Ψ **é compacto**:

   – A compacidade do espaço paramétrico pode ser obtida impondo-se um limite superior para o módulo dos parâmetros. Este limite pode ser tão grande quanto convier, de tal forma que pode ser sempre definido de modo a ter folga;

4. $f(x,\psi)$ **é contínuo para cada** $\psi \in \Psi$ **com probabilidade um**:

   – A utilização de funções de ativação logísticas, que são contínuas e com derivadas contínuas, garante que $f(x,\psi)$ é contínua para qualquer vetor de parâmetros ψ limitado;

5. $E\left[\sup_{\psi\in\Psi} \|f(x,\psi)\|\right] < \infty$:

   – Partindo-se da compacidade de Ψ e considerando-se que a função de ativação logística é limitada em um, basta que $\|E[x]\| < \infty$ para que esta condição seja assegurada.

Já a respeito das condições para normalidade assintótica, pode-se pontuar:

1. **o estimador é consistente**:

   – A consistência pode ser obtida pelas condições tratadas acima;

2. ψ^* **está no interior de** Ψ:

   – Pelo mesmo argumento utilizado para a terceira condição de consistência, pode-se definir o espaço paramétrico Ψ grande o bastante de tal forma que ψ^* não esteja na fronteira;

3. $f(x, \psi)$ **é continuamente diferenciável em uma vizinhança** ϑ **de** ψ^* **com probabilidade um**:

   – A utilização das funções de ativação logísticas garante que $f(x, \psi)$ tenha derivada contínua em todo ponto x, logo qualquer subconjunto convexo de Ψ que contenha ψ^* pode ser a vizinhança ϑ;

4. $E[\|f(x, \psi^*)\|^2] < \infty$:

   – Esta condição depende, somente, da finitude da variância dos componentes do vetor x. É necessário, porém não suficiente, que x seja estacionário;

5. $E\left[\sup_{\psi \in \vartheta} \|\nabla_\psi f(x, \psi)\|\right] < \infty$:

   – No cálculo das derivadas parciais de $f(x, \psi)$, aparecem termos de produtos cruzados entre os elementos do vetor x. Com Ψ compacto, novamente será suficiente que x possua variância finita;

6. $G \cdot W \cdot G'$ **é não singular para** $G = E\left[\nabla_\psi f(x, \psi)\right]$:

   – As condições de momento derivadas da QMV garantem que G possua posto cheio. Sendo W positiva definida, a forma quadrática $G \cdot W \cdot G'$ será positiva definida e, portanto, não singular.

Em suma, o atendimento das condições suficientes para garantia de consistência e normalidade assintótica dos parâmetros de uma RNA estimada pelo MGM no arcabouço proposto não depende de nenhuma hipótese forte ou demasiadamente restriva.

Otimização

Nos casos em que se utiliza o MGM Eficiente em Dois Estágios ou o MGM Eficiente Iterado, o ponderador W_N está fixo cada vez que se aplica a fórmula (3-6). Nestes casos, a exemplo do que é feito nas estimações por QMV, pode-se dividir o problema de otimização em dois subproblemas, conforme será explicitado a seguir.

Convencionando-se as matrizes Y $(N \times 1)$, X $(N \times I)$ e U $(N \times 1)$ com as realizações da variável dependente, das explicativas e dos resíduos, respectivamente, empilhadas nas linhas, Ω $(I \times M)$ a matriz composta pelo enfileiramento nas colunas dos vetores pré-definidos $\{\vec{\omega}_m\}_{m=1}^M$, e D^* $(N \times K)$ a matriz com os vetores $\{d^*(\vec{x}_n; \{\vec{\omega}_m\}_{m=1}^M)'\}_{n=1}^N$ empilhados nas colunas, o problema de estimação pode ser reescrito como:

$$\hat{\psi}_{MGM} = \underset{\psi \in \Psi}{\operatorname{argmin}}\, U' \cdot D^* \cdot W_N \cdot D^{*\prime} \cdot U. \tag{4-43}$$

Esta formulação, porém, é um problema de Mínimos Quadrados Ponderados (MQP), cuja solução analítica é dada por:

$$[\hat{\alpha}'_{MGM}, \hat{\lambda}'_{MGM}] = (Z' \cdot W^* \cdot Z)^{-1} \cdot Z' \cdot W^* \cdot Y, \tag{4-44}$$

onde $Z = [X, h(Z \cdot \Omega)]$ e $W^* = D^* \cdot W_N \cdot D^{*\prime}$.

Ou seja, a exemplo do que ocorre na estimação por QVM, para um dado conjunto de $\{\vec{\omega}_m\}_{m=1}^M$, os demais parâmetros podem ser estimados com fórmula fechada[8] e, portanto, convém definir o problema de otimização somente em termos de $\{\vec{\omega}_m\}_{m=1}^M$, considerando os demais parâmetros implícitos.

Baseado no processo de duas etapas para a estimação por QMV proposto por [24], pode-se propor um processo análogo, com três etapas, para a estimação via MGM:

1. obtenção do ponto inicial:
   - gera-se um grande número de $\{\vec{\omega}\}_{m=1}^M$ aleatórios;
   - para cada $\{\vec{\omega}\}_{m=1}^M$, estima-se os demais parâmetros, $\vec{\alpha}$ e $\vec{lambda}$, por MQO e toma-se os resíduos;
   - com os resíduos e $\{\vec{\omega}\}_{m=1}^M$, calcula-se um ponderador W_N;
   - de posse do ponderador, reestima-se os parâmetros $\vec{\alpha}$ e $\vec{\lambda}$ por MQP e calcula-se o valor da função objetivo de (3-6);
   - toma-se os parâmetros $\{\vec{\omega}\}_{m=1}^M$ para os quais a função objetivo atinge o menor valor e recalcula-se o ponderador W_N com base em $\{\vec{\omega}\}_{m=1}^M$ e nos resíduos da estimação por MQP.
2. otimização local:
   - com base em $\{\vec{\omega}\}_{m=1}^M$ e W_N obtidos na etapa anterior aplica-se um método de otimização local[9]para obter-se uma nova estimativa de $\{\vec{\omega}\}_{m=1}^M$;
   - recalcula-se o ponderador W_N com base nos novos $\{\vec{\omega}\}_{m=1}^M$ e nos respectivos resíduos.
3. reiterações:
   - repete-se a etapa de otimização local uma (MGM Eficiente em Dois Estágios) ou mais vezes (MGM Eficiente Iterado).

[8]Com baixíssimo custo computacional quando comparado a outros métodos que otimização que poderiam ser utilizados, como métodos de gradiente, por exemplo.

[9]Há uma recomendação em [24] para o uso de métodos quasi-newton, Broyden-Fletcher-Goldfarb-Shanno ou LevenbergMarquardt, descritos em [58]. Outros métodos, porém, podem ser aplicados neste caso como, por exemplo o método *Trust Region* ou o Simplex Nelder-Mead, descritos em [68].

Assim como ocorre no caso da QMV, este procedimento não garante a otimalidade global dos parâmetros estimados, que é condição necessária para a validade do estimador de MGM. Uma discussão mais profunda sobre a questão da otimalidade global em estimações pelo MGM pode ser encontrada em [69].

4.2.2 Caso Particular

Dentre as diversas possibilidades de funções Borel-mensuráveis $l(\cdot)$ possíveis para a geração das condições de momento adicionais, optou-se, nas implementações realizadas no escopo do presente trabalho, por utilizar combinações lineares de unidades escondidas com função de ativação logística (ou RNAS com $M = 1$, $\vec{\alpha} = \vec{0}$ e $\lambda = 1$) com parâmetros gerados aleatoriamente. A ideia de RNAs com parâmetros randomizados não é inédita [23, 26], porém, sua utilização no contexto de estimação, até onde se pode verificar, não encontra precedentes.

Esquematicamente, o processo de obtenção das funções $l(X)$ é dado por:

- sorteia-se 1000 vetores de parâmetros $\vec{\omega}$ de uma Distribuição Uniforme entre $[-2, 2]$ (seguindo [26]) e os enfilera na matriz Ω;
- calcula-se $Z = h(X \cdot \Omega) - X \cdot (X' \cdot X)^{-1} \cdot X' \cdot l(X \cdot \Omega)$;
- toma-se os $K - L$ primeiros componentes principais de Z como sendo as funções $l(X)$.

O primeiro passo gera diversos formatos de não linearidade. O segundo expurga toda a linearidade existente de forma a evitar redundância em relação às condições de momento derivadas da otimização da QMV. O último passo concentra ao máximo a informação gerada nos passos anteriores.

É possível justificar a escolha de unidades escondidas com função de ativação logística em detrimento de todas as outras possibilidades supracitadas pela existência de um resultado segundo o qual, sendo ε uma variável aleatória com variância finita e $\vec{x}$ um vetor aleatório limitado em $\mathbb{R}^I$ tais que $P[E(\varepsilon \mid \vec{x}) = 0] < 1$, então o conjunto $O = \{\vec{\omega} \in \mathbb{R}^I \mid E[\varepsilon \cdot h(\vec{x} \cdot \vec{\omega}')] = 0\}$ possui medida de Lesbegue nula. Isso garante que, mesmo com a escolha de um vetor $\vec{\omega}$ ao acaso, a chance de que ele não agregue informação à estimação é desprezível. Uma outra função que também possui essa propriedade, a exponencial de uma combinação linear dos regressores, pode gerar não-estacionariedade no caso de séries de tempo.

4.2.3
O Teste J para Não Linearidade Negligenciada

No arcabouço geral proposto para extimação de RNAs, os MGMs são sobre-identificados. Nesses casos, é possível realizar o Teste J apresentado no Capítulo 3, cujas hipóteses são:

- Hipótese nula: $E[f(x, \psi^*)] = 0$, contra
- Hipótese alternativa: $E[f(x, \psi)] \neq 0$, $\forall\ \psi \in \Psi$.

Sob a hipótese alternativa, nunhum vetor de parâmetros do espaço paramétrico é capaz de fazer com que todas as condições de momento sejam satisfeitas.

As condições de momento derivadas da QMV espelham uma propriedade da esperança condicional e, portanto, a não satisfação de uma destas condições só pode ocorrer no caso em que $G(\vec{x}; \psi^*) \neq E[y \mid \vec{x}]$. Da mesma forma, as condições adicionais, derivadas da hipótese de exogeneidade entre os regressores e o ruído, também só serão violadas nos casos em que $G(\vec{x}; \psi^*) \neq E[y \mid \vec{x}]$[10]. Portanto, $G(\vec{x}; \psi^*) = E[y \mid \vec{x}]$ é suficiente para que todas as condições de momento sejam satisfeitas e a rejeição da hipótese nula indica que nenhum vetor de parâmetros é capaz de aproximar a verdadeira esperança condicional.

O Teorema da Aproximação Universal [9, 39] garante que a RNA com uma quantidade suficiente de unidades escondidas é capaz de aproximar arbitrariamente bem qualquer função Borel-mensurável limitada. Dito de outra forma, um RNA não será capaz de aproximar uma determinada função Borel-mensurável somente se não tiver a quantidade suficiente de unidades na camada escondida.

Portanto, sabe-se que a hipótese nula no Teste J será rejeitada somente se RNA não for capaz de aproximar a esperança condicional, e que RNA não será capaz de aproximar a esperança condicional somente se não possuir uma quantidade suficiente de unidades na camada escondida. Logo, uma rejeição no Teste J ao nível de significância escolhido é um indício de há necessidade de incremento na arquiterura, ou seja, que M deve ser aumentado. Neste sentido, o Teste J aplicado a RNAs estimadas pelo MGM, conforme o arcabouço geral aqui apresentado, pode ser interpretado como um teste para a necessidade de um neurônio adicional.

A estatística do Teste J para neurônio adicional é dada por:

$$J \equiv \frac{1}{N} \cdot \left(\sum_{n=1}^{N} f(x_n, \hat{\psi}_{MGM}) \right)' \cdot W_N \cdot \left(\sum_{n=1}^{N} f(x_n, \hat{\psi}_{MGM}) \right). \qquad (4\text{-}45)$$

[10] As condições de momento serão atendidas mesmo nos casos em que os regressores são endógenos, apesar de haver viés nos parâmetros estimados e da eventual perda do sentido de causalidade.

Com W_N convergindo em probabilidade para $W^* = E[f(x,\psi^*)' \cdot f(x,\psi^*)]^{-1}$, sob a hipótese nula, sabe-se que $J \xrightarrow{d} \mathcal{X}^2_{K-L}$.

Considerando as hipóteses dos TNLN apresentadas anteriormente:

- Hipótese nula: $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) = 1$ para algum $\vec{\alpha} \in \mathbb{R}^I$, contra
- Hipótese alternativa: $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) < 1$ para todo $\vec{\alpha} \in \mathbb{R}^I$.

e o caso particular no qual a RNA estimada possui unidades escondidas, ou seja, $M = 0$, será apresentada a proposição a seguir.

Proposição 1 *Nos casos em a RNA estimada tem $M = 0$, a hipótese alternativa do Teste J, $E[f(x,\psi)] \neq 0$, $\forall\ \psi \in \Psi$, será verdadeira somente se o verdadeiro modelo for não linear, ou seja, se $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) < 1$ para todo $\vec{\alpha} \in \mathbb{R}^I$[11] e, portanto, uma rejeição no Teste J é equivalente a uma rejeição em um TNLN.*

Prova 1 *Com $M = 0$, temos $f(\vec{x},\psi) = (y - G(\vec{x};\psi)) \cdot d^*(\vec{x}) = \vec{0}$ e*

$$d^*(\vec{x}) = \begin{bmatrix} \vec{x} \\ l(\vec{x}) \end{bmatrix}. \tag{4-46}$$

As condições de ortogonalidade $(y - G(\vec{x};\psi)) \cdot \vec{x}$ podem ser satisfeitas independentemente de o verdadeiro modelo ser linear ou não, pelos parâmetros $\vec{\alpha}$ de MQO. Portanto, a impossibilidade no atendimento das condições de momento é causada por $(y - G(\vec{x};\psi)) \cdot l(\vec{x})$. Caso o verdadeiro modelo seja linear, existe $\vec{\alpha}$ tal que $P(E[y|x] = \vec{\alpha}' \cdot \vec{x}) = 1$, logo, $E[(y - G(\vec{x};\psi)) \cdot l(\vec{x})] = E[y - G(\vec{x};\psi)] \cdot E[l(\vec{x})] = 0$. Isso significa que a linearidade do verdadeiro modelo é suficiente para garantir que todas as condições de momento possam ser atendidas e, consequentemente, que $E[f(x,\psi)] \neq 0\ \forall\ \psi \in \Psi$ só é possível se o verdadeiro modelo for não linear.

Sob a hipótese nula do Teste J, o modelo linear é capaz de satisfazer todas as condições de momento, inclusive as condições não-lineares derivadas da hipótese de exogeneidade, ou seja, o modelo linear aproxima adequadamente a esperança condicional. Portanto, nos casos em que $M = 0$, o Teste J funciona como um TNLN. As propriedades em amostra finita deste teste, denominado Teste J para NLN, serão objeto de um dos estudos de caso, apresentado mais adiante.

[11] Nos casos em que $M = 0$, o vetor de parâmetros ψ se resume a $\vec{\alpha}$. Para evitar lidar com possíveis soluções de canto no problema de estimação, será considerado o caso em que $\Psi \equiv \mathbb{R}^I$.

5
Estudos de Caso

Neste capítulo, serão apresentados três estudos de caso envolvendo os conceitos desenvolvidos no presente trabalho. Os dois primeiros comparam as estimações realizadas pelo MGM e pela QMV em dados simulados, no primeiro estudo de caso, e em dados reais de taxas de câmbio, no segundo. O terceiro estudo de caso compara as propriedades em amostra finita do Teste J para NLN às do LWG.

5.1
Experimentos de Monte Carlo para Estimação

Nesta seção, serão expostos os experimentos de Monte Carlo realizados a fim de mensurar a acurácia das estimações realizadas pelo MGM, vis-à-vis a QMV, em um ambiente controlado. As variáveis que serão observadas neste experimento são: (i) o tamanho de amostra (N); (ii) a razão de ruído (q); e (iii) o número de condições de momento do MGM (K).

5.1.1
Geração dos Dados

A esperança condicional da variável dependente foi gerada a partir de RNA, de modo a evitar problemas de aproximação. Esta RNA possui duas unidades escondidas ($M = 2$) e duas variáveis explicativas ($I = 2$), sendo uma constante e um variável aleatória sorteada de uma distribuição Uniforme no intervalo $[-1, 1]$. Os parâmetros ψ^* são:

$$\begin{aligned}
\vec{\alpha} &= \begin{bmatrix} -1.5927 & 1.0000 \end{bmatrix}', \\
\vec{\lambda} &= \begin{bmatrix} 1.5927 & 1.5927 \end{bmatrix}', \\
\vec{\omega}_1 &= \begin{bmatrix} 1.9520 & -10.1690 \end{bmatrix}', \\
\vec{\omega}_2 &= \begin{bmatrix} 1.9520 & 10.1690 \end{bmatrix}'.
\end{aligned}$$

Esta função aproxima uma curva Gaussiana sobre uma inclinação de 45 graus, como pode ser visto na figura 5.1.

O termo de ruído ε_n foi sorteado de uma distribuição Normal, com media nula e com variâncias tais que as razões de ruído fossem de 10%, 30%, 50%, 70% e 90%. Três tamanhos de amostra (50, 200 e 800) foram testados

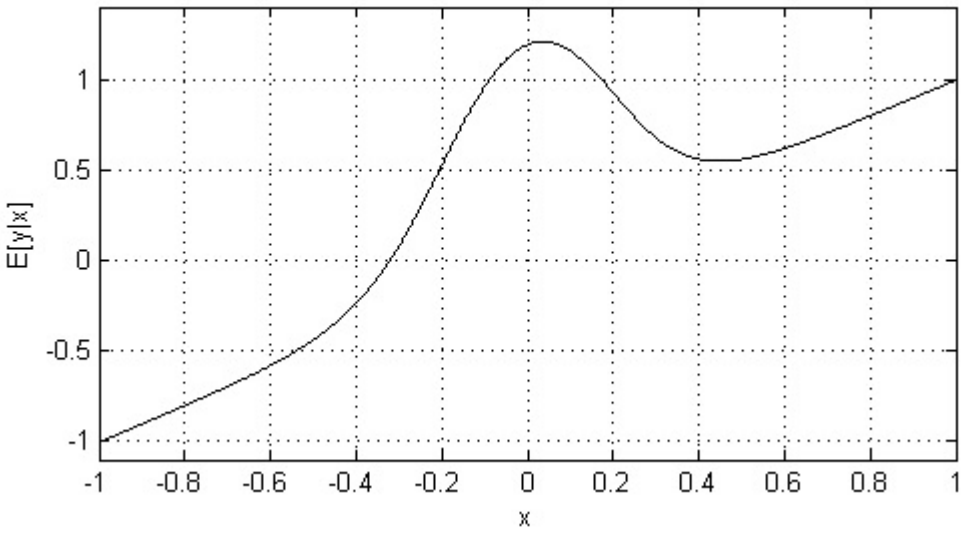

Figura 5.1: Esperança Condicional de y.

nos experimentos. Para cada uma das quinze combinações entre variância e tamanho de amostra, 1.000 repetições do experimento foram conduzidas.

É importante notar que, por conta dos ruídos Gaussianos, o estimador de QMV coincide com o de MV e, portanto, é assintoticamente eficiente. Logo, qualquer outro método só é capaz de obter melhor performance em amostras finitas.

5.1.2 Métodos Comparados

O método de estimação de referência nos testes realizados foi a QMV, aplicado através do procedimento em duas etapas apresentado no Capítulo 4, baseado em [24]. Para a etapa de inicialização, foram utilizados 1.000 conjuntos $\{\vec{\omega}_m\}_{m=1}^{M}$ aleatórios[1].

O método alternativo, o MGM, foi implementado de acordo com o caso particular (Capítulo 4), com o procedimento de otimização em três etapas com uma única iteração na etapa (3), ou seja, um MGM Eficiente em Dois Estágios[2]. Oito versões deste estimador utilizando entre nove e dezesseis condições de momento, nomeados MGM-9, MGM-10 e assim sucessivamente, foram utilizadas no teste.

[1] O procedimento de geração dos $\{\vec{\omega}_m\}_{m=1}^{M}$ randômicos está descrito no apêndice.

[2] Este conjunto de experimentos é muito intensivo em tempo de processamento computacional, demorando aproximadamente dez dias para ser executada. A adição de mais iterações no cálculo do ponderador, apesar de desejável, foi suprimida para não tornar a execução ainda mais demorada. A mesma lógica serve para o número de $\{\vec{\omega}_m\}_{m=1}^{M}$ aleatórios utilizados na etapa inicial.

Para cada repetição do experimento, um mesmo conjunto de 1.000 sorteios aleatórios de $\left\{\vec{\omega}_m\right\}_{m=1}^{M}$ foi utilizado na etapa inicial do procedimento de otimização de todas as estimações. Da mesma forma, os mesmo algoritmos foram empregados em todas as etapas de otimização local. Seguindo-se a recomendação de [24], foi utilizado o método Broyden-Fletcher-Goldfarb-Shanno[3], detalhado no Apêndice. Com isso, pretendeu-se garantir as condições as mais iguais possíveis entre as metodologias comparadas, isolando, na medida do possível, as interferências externas.

5.1.3 Critério de Comparação

Cada RNA estimada foi analisada em termos da raiz quadrada do erro quadrático médio (REQM) em relação à RNA com os verdadeiros parâmetros ψ^* em um *grid* fino de valores da variável explicativa:

$$REQM_{Met} = \sqrt{\frac{\sum_{j=-1000}^{1000}\left[G\left(\left[1,\frac{j}{1000}\right];\psi^*\right) - G\left(\left[1,\frac{j}{1000}\right];\hat{\psi}_{Met}\right)\right]^2}{2001}}, \qquad (5\text{-}1)$$

onde o subíndice Met denota o método de estimação utilizado.

Os resultados reportados são os quantis 50% e 95% das REMQ ao longo das 1.000 realizações. Uma análise do viés nos parâmetros estimados é apresentado no apêndice E.

5.1.4 Resultados

Os resultados estão agrupados por tamanhos de amosta (N). A tabela 5.1 apresenta os resultados para $N = 50$.

A QMV obtém medianas menores para as razões de ruído mais baixas, enquanto os MGMs são dominantes sob alto ruído. Com $q = 90\%$, as medianas dos MGMs ficam por volta de 30% abaixo da mediana da QMV. Considerando-se o quantil de 95%, percebe-se que a QMV é melhor que os MGMs somente para $q = 10\%$ e que a sua performance se deteriora rapidamente com o aumento de q. Com $q = 30\%$, o quantil de 95% da QMV já é da ordem de duas vezes maior que o dos MGMs e este multiplicador atinge a ordem de cinquenta vezes quando $q = 90\%$.

A tabela 5.2 traz os resultados para $N = 200$. A QMV supera todos os MGMs, tanto em mediana quanto no quantil de 95%, quando $q = 10\%$. Por outro lado, os MGM com treze ou mais condições de momento são superiores

[3]Implementado na função *fminunc* do *software* Matlab.

Tabela 5.1: Tamanho de Amostra 50

		Razão de Ruído (q)				
Quantil	Método	10%	30%	50%	70%	90%
50%	QMV	0,226	0,345	0,497	0,77	1,435
	MGM-9	0,432	0,450	0,487	0,572	0,906
	MGM-10	0,434	0,450	0,489	0,581	0,93
	MGM-11	0,434	0,451	0,494	0,596	0,97
	MGM-12	0,262	0,352	0,449	0,587	0,99
	MGM-13	0,26	0,361	0,454	0,583	1,019
	MGM-14	0,263	0,352	0,45	0,59	1,012
	MGM-15	0,269	0,359	0,447	0,602	1,022
	MGM-16	0,315	0,376	0,456	0,593	1,005
95%	QMV	0,366	1,233	1,66	27,801	120,32
	MGM-9	0,507	0,59	0,7	1,057	1,954
	MGM-10	0,522	0,572	0,673	0,976	2,015
	MGM-11	0,567	0,652	0,926	1,095	2,1
	MGM-12	0,486	0,629	0,821	1,475	2,275
	MGM-13	0,46	0,598	0,718	0,991	2,426
	MGM-14	0,454	0,524	0,73	1,008	1,999
	MGM-15	0,468	0,544	0,706	1,134	1,964
	MGM-16	0,467	0,562	0,715	1,048	1,751

Tabela 5.2: Tamanho de Amostra 200

		Razão de Ruído (q)				
Quantil	Método	10%	30%	50%	70%	90%
50%	QMV	0,045	0,096	0,156	0,249	0,502
	MGM-9	0,053	0,11	0,168	0,232	0,381
	MGM-10	0,05	0,104	0,158	0,23	0,393
	MGM-11	0,05	0,1	0,153	0,225	0,394
	MGM-12	0,049	0,098	0,151	0,225	0,404
	MGM-13	0,049	0,096	0,146	0,221	0,403
	MGM-14	0,049	0,094	0,147	0,223	0,405
	MGM-15	0,049	0,091	0,144	0,219	0,4
	MGM-16	0,05	0,09	0,142	0,219	0,405
95%	QMV	0,067	0,151	0,24	0,376	0,753
	MGM-9	0,116	0,19	0,248	0,333	0,575
	MGM-10	0,091	0,175	0,249	0,346	0,613
	MGM-11	0,087	0,158	0,233	0,326	0,596
	MGM-12	0,074	0,159	0,226	0,325	0,603
	MGM-13	0,073	0,144	0,223	0,329	0,612
	MGM-14	0,072	0,143	0,221	0,324	0,593
	MGM-15	0,071	0,138	0,225	0,324	0,6
	MGM-16	0,073	0,142	0,23	0,323	0,62

à QMV em todos os demais casos, sob ambas as métricas de comparação. De maneira geral, os resultados para $N = 200$ são mais equilibrados do que para $N = 50$. Com $q = 90\%$, os resultados obtidos com o MGM ficaram entre 20% e 25% abaixo dos obtidos pela QMV.

Tabela 5.3: Tamanho de Amostra 800

		Razão de Ruído (q)				
Quantil	Método	10%	30%	50%	70%	90%
50%	QMV	0,023	0,044	0,07	0,113	0,248
	MGM-9	0,025	0,05	0,084	0,126	0,229
	MGM-10	0,024	0,048	0,078	0,12	0,231
	MGM-11	0,024	0,048	0,077	0,116	0,221
	MGM-12	0,025	0,047	0,073	0,114	0,221
	MGM-13	0,024	0,046	0,072	0,109	0,222
	MGM-14	0,024	0,046	0,071	0,109	0,218
	MGM-15	0,024	0,046	0,069	0,105	0,216
	MGM-16	0,025	0,046	0,069	0,105	0,216
95%	QMV	0,033	0,068	0,107	0,179	0,355
	MGM-9	0,08	0,112	0,161	0,207	0,321
	MGM-10	0,036	0,088	0,136	0,198	0,335
	MGM-11	0,037	0,085	0,116	0,175	0,314
	MGM-12	0,036	0,071	0,115	0,175	0,319
	MGM-13	0,036	0,071	0,106	0,17	0,316
	MGM-14	0,036	0,072	0,106	0,165	0,316
	MGM-15	0,036	0,07	0,101	0,164	0,323
	MGM-16	0,042	0,07	0,102	0,164	0,317

Para o maior tamanho de amostra testado (tabela 5.3), os resultados são os mais equilibrados. Apesar de a QMV ser superior a todos os MGMs para $q \leq 30\%$, as vantagens são pequenas se comparado aos resultados para amostras menores. Para este tamanho de amostra, existe uma relação clara entre o número de condições de momento e a performance da estimação pelo MGM. O MGM-15 e o MGM-16 superam a QMV para todo $q > 30\%$, com uma mediana aproximadamente 15% menor no caso em que $q = 90\%$. De fato, todos os MGMs foram superiores à QMV para $q = 90\%$, tanto em mediana quanto para o quantil de 95%.

Em resumo, os MGM entregaram melhores estimações nos casos em que a razão de ruído era elevada. Esta vantagem foi mais significativa quanto menor foi a amostra usada na estimação. Estes resultados são compatíveis com a hipótese de que o MGM sobreidentificado seria menos sujeito ao sobreajuste. Sob outra perspectiva, a divisão dos casos em que o MGM ou a QMV foram superiores está em linha com a conjectura feita a respeito dos

resultados de [66, 67] segundo os quais a adição de condições de momento redundantes deteriorariam a estimação. De forma geral, pode-se afirmar que as performances dos MGMs foram positivamente correlacionadas ao número de condições de momento.

Os casos em que a superioridade dos MGMs em relação à QMV foi mais pronunciada, com pequena amostra e alto ruído, são, justamente, as situações mais complicadas para a estimação, porém não são incomuns. Alguns campos de aplicação relevantes, como finanças e detecção de fraudes, as séries de dados costumam apresentar altos níveis de ruído. Aplicações de série temporais, especialmente com periodicidade anual, em muitos casos, possuem apenas algumas dezenas de dados, por conta de descontinuidades ou quebras estruturais. O MGM pode ser uma ferramenta valiosa para estimação de RNAs nestes e em outros casos.

5.2
Experimentos de Previsão de Taxas de Câmbio

Desde o início dos anos 1980 [70, 71], as taxas de câmbio vêm desafiando economistas, econometristas e outros incautos que tentam prevê-las. Tanto os modelos estruturais econômicos como os modelos econométricos lineares são incapazes de obter significativas melhorias de previsão em relação ao passeio aleatório, modelo segundo o qual a melhor previsão para o câmbio no futuro é o seu valor atual. Algumas possíveis razões para este fenômeno foram levantadas em [72–74].

Por outro lado, diversos estudos apontaram a existência de não linearidades relevantes na dinâmica das taxas de câmbio [75–78]. Em particular, RNAs foram aplicadas para previsão de taxas de câmbio em [79–84], todos com algum grau de sucesso, porém nenhum deles com resultados fortes o suficiente para ser apontado como uma solução definitiva para o problema.

As taxas de câmbio são extremamente importantes em economia e finanças, possuem baixa previsibilidade, ou seja, alto ruído, e há evidências de que apresentam não linearidades em sua dinâmica. Por conta destes fatores, foram escolhidas para os experimentos comparando as estimações de RNAs por QMV e MGM. Para fins de demarcação, convém explicitar que o objetivo destes experimentos não é propor uma nova metodologia para previsão de taxas de câmbio, e sim comparar as performances de RNAs estimadas pelos dois métodos.

5.2.1

Dados

Foram utilizadas séries de taxas de câmbio de dez moedas em relação ao dólar americano para as quais [77] não rejeitou a hipótese de não linearidade negligenciada (quatorze séries foram testadas ao todo). Os países cujas moedas foram consideradas são: Áustria, Bélgica, Finlândia, França, Alemanha, Suécia, Austrália, Espanha, Sri Lanka e Índia.

Uma descrição completa das séries e do TNLN realizados pode ser encontrada em [77], com a única diferença que as séries foram atualizada, incorporando dados de julho de 2000 até dezembro de 2001. As séries são mensais e iniciam-se em janeiro de 1971, com exceção de Espanha e Sri Lanca, cujas séries começam em janeiro de 1973.

As previsões dos retornos logarítmicos das taxas de câmbio (variáveis dependentes) foram realizado para o período de seis anos entre janeiro de 1996 e dezembro de 2001 (72 observações), com base no log-retorno do mês imediatamente anterior(variáveis explicativas)[4]. Três esquemas de reestimação foram testados: (i) uma única estimação para todos os 72 períodos (sem reestimações); (ii) reestimações anuais; e (iii) reestimações mensais. Todas as reestimações utilizam toda a respectiva série pregressa.

5.2.2 Métodos Comparados

As RNAs foram estimadas com uma única unidade escondida ($M = 1$), a fim de evitar a possibilidade de haver unidades demais. O número de parâmetros estimados é $L = 5$. Os métodos comparados foram os mesmos utilizados nos experimentos de Monte Carlo descritos anteriormente, QMV e MGM, com as seguintes diferenças nas implementações:

- os MGM foram testados com seis a nove condições de momento, seguindo a mesma lógica de nomenclatura; e
- nas etapas de inicialização foram gerados 10.000 conjuntos $\left\{\vec{\omega}_m\right\}_{m=1}^{M}$ aleatórios, e não apenas 1.000;
- na terceira etapa do processo de estimação pelo MGM, foi incluída uma segunda iteração do processo de otimização local, ou seja, trata-se de um MGM Eficiente Iterado com critério de parada na segunda iteração.

A primeira diferença foi motivada pelo fato de haver menos parâmetros a serem estimados, fazendo sentido o uso de menos condições de momento. As outras duas mudanças foram realizadas porque, uma vez que este estudo

[4] A primeira defasagem é o único regressor comum a todas as abordagens de previsão de taxas de câmbio com RNAs supracitadas.

de caso envolve uma quantidade de estimações muito menor que o anterior (menos de um centésimo), o custo computacional destas alterações é tolerável.

5.2.3 Critério de Comparação

Cada série de previsões teve a sua REQM em relação aos valores realizados calculado de forma análoga a equação (5-1). A fim de facilitar as comparações entre séries de diferentes taxas de câmbio, os resultados foram apresentados em termos da diferença percentual (DP) dos REQM do MGM em relação à QMV:

$$DP_K = \left(\frac{REQM_{MGM-K}}{REQM_{QMV}} - 1 \right) \cdot 100\%. \tag{5-2}$$

O subescrito K refere-se ao número de condições de momento do estimador de MGM. Valores positivos de DP_K indicam que a QMV produziu previsões mais próximas dos valores realizados.

Foram incluídas notações de significância estatística padrão para DP_K (* para significativo ao nível de 10%, ** para 5% e *** para 1%), baseadas em intervalos de confiança unicaudais obtidos utilizando-se um *bootstrap* simples.

5.2.4 Resultados

Os resultados estão apresentados de forma agregada por esquema de reestimação.

Começando com o caso sem reestimações (tabela 5.4), todos os MGM geraram previsões mais precisas na previsão do que a QMV para nove das dez séries do experimento, tendo, como única exceção, o dólar australiano. Os MGMs com 7 ou mais condições de momento obtiveram reduções da REQM significativas ao nível de 10% para a moeda do Sri Lanka e, para o MGM_9, a redução foi significativa à 1%. Em média, os MGMs tiveram uma REQM 4,46% menor que a QMV.

Com as reestimações anuais (tabela 5.5), os MGMs continuam dominantes. O MGM_8 e o MGM_9 superam a QMV em todas as séries. Para a moeda do Sri Lanka, o MGM_7 e o MGM_9 obtiveram vantagens significativas ao nível de 5%. Em média, a redução da REQM obtida com os MGMs foi de 4,64%.

Para as reestimações mensais (tabela 5.6), o padrão observado é o obtido com reestimações anuais. Difere o fato de que a redução da REQM para o MGM_9 foi significativa ao nível de 1%. A redução média do REQM obtida com os MGMs foi de 4,52%. Para as reestimações mensais (tabela 5.6), o padrão observado é o obtido com reestimações anuais. Difere o fato de que a

Tabela 5.4: Previsões Sem Reestimações

País	DP_6	DP_7	DP_8	DP_9
Áustria	-2,876%	-2,986%	-2,896%	-4,371%
Bélgica	-3,848%	-3,56%	-5,136%	-5,622%
Finlândia	-5,687%	-5,737%*	-5,658%	-5,669%
França	-4,789%	-4,879%	-4,684%	-4,221%
Alemanha	-3,957%	-2,819%	-2,808%	-4,578%
Suécia	-2,612%	-2,606%	-2,584%	-1,806%
Austrália	1,781%	0,944%	1,384%	0,703%
Espanha	-4,822%	-4,849%	-4,8%	-5,951%
Sri Lanka	-6,772%	-15,327%**	-14,386%*	-18,447%***
Índia	-2,111%	-2,882%	-3,266%	-3,315%

Tabela 5.5: Previsões Com Reestimações Anuais

País	DP_6	DP_7	DP_8	DP_9
Áustria	-3,454%	-3,488%	-3,618%	-4,64%
Bélgica	-4,225%	-4,834%	-5,464%	-5,751%
Finlândia	-5,726%	-5,749%	-5,715%	-5,789%
França	-4,115%	-4,022%	-4,078%	-4,962%
Alemanha	-4,241%	-3,36%	-3,35%	-4,792%
Suécia	-2,988%	-2,935%	-2,87%	-1,801%
Austrália	1,148%	0,299%	-0,404%	-0,232%
Espanha	-4,805%	-4,832%	-4,8%	-5,855%
Sri Lanka	-7,672%	-15,729%**	-11,135%	-17,561%**
Índia	-4,141%	-2,694%	-3,336%	-1,982%

redução da REQM para o MGM_9 foi significativa ao nível de 1%. A redução média do REQM obtida com os MGMs foi de 4,52%.

Estes resultados mostram uma clara vantagem do MGM sobre a QMV. As previsões realizadas com o MGM foram mais precisas na enorme maioria das vezes e, em alguns casos, com diferenças de precisão estatisticamente significativas. Em média, os MGMs reduziram a REQM do QMV em 4,54% e, nos casos extremos, foi além dos 15%.

No esquema de reestimação menos desfavorável à QMV (sem reestimações), a QMV foi superior aos MGMs em uma dentre dez séries. Considerando-se a previsão de cada de série como um experimento independente das demais previsões, a hipótese de que o MGM e a QMV são equivalentes (no sentido em que ambos os métodos têm a mesma probabilidade de gerar a melhor previsão) tem um p-valor de 1,074219%. Portanto, dados os resultados observados, é extremamente improvável que a QMV seja equivalente ao MGM.

Duas séries possuem resultados que chamam a atenção. A série do câmbio

Tabela 5.6: Previsões Com Reestimações Mensais

País	DP_6	DP_7	DP_8	DP_9
Áustria	-3,744%	-3,422%	-3,572%	-4,317%
Bélgica	-3,419%	-3,912%	-5,324%	-5,512%
Finlândia	-5,407%	-5,295%	-4,911%	-5,674%
França	-4,757%	-4,234%	-4,338%	-4,085%
Alemanha	-4,014%	-3,273%	-3,252%	-4,499%
Suécia	-2,766%	-2,722%	-2,647%	-1,375%
Austrália	0,877%	0,244%	-0,086%	-0,282%
Espanha	-4,631%	-4,424%	-4,379%	-5,695%
Sri Lanka	-7,708%	-15,595%**	-12,683%	-17,927%***
Índia	-4,21%	-2,778%	-3,108%	-1,992%

da Austrália concentra todas as vitórias da QMV. São vitórias por uma margem pequena (sempre menor que 2%) e não significativas, e uma explicação para plausível para elas é a baixa variância da série (somente maior que a da moeda indiana). No extremo oposto, há os resultados para a moeda do Sri Lanka, para as quais os MGMs obtêm vantagens expressivas, algumas maiores que 10%, e significativas. Trata-se da série com maiores variância, assimetria e curtose, além do maior valor absoluto (0,37805), ou seja, apresenta condições propícias para o sobreajuste. A vantagem dos MGMs pode ser explicada pela menor suscetibilidade ao sobreajuste.

Via de regra, a performance dos MGMs foi melhor quanto mais condições de momento ele possuísse. Dado o alto nível de ruído nas séries de taxas de câmbio, a supremacia do MGM sobre a QMV está perfeitamente alinhada com os resultados dos experimentos de Monte Carlo apresentados na seção anterior.

5.3 Experimentos de Monte Carlo para os TNLNs

Neste último estudo de caso, serão apresentados os experimentos de Monte Carlo conduzidos a fim de mensurar as propriedades em amostra finita do Teste J para NLN, a saber, o tamanho e a potência, e compará-las com o do teste LWG. Como já foi dito anteriormente, o teste LWG é uma boa referência porque, além de ser um dos mais utilizados, obteve resultados positivos quando comparado a diversos outros TNLNs [27, 30].

O tamanho de um teste estatístico é a probabilidade de rejeição da hipótese nula nos casos em que ela é verdadeira. Diz-se que um teste tem tamanho adequado quando a probabilidade de rejeição é igual ou próxima do seu nível de significância nominal. Um teste com tamanho inadequado é dito viesado. Já a potência é a probabilidade de rejeição da hipótese nula nos casos

e que ela é falsa. Portanto, é desejável que a potência de um teste não viesado seja a mais alta possível. A potência de um teste varia dependendo do tipo de violação da hipótese nula.

5.3.1 Geração dos Dados

Nos experimentos, foram utilizados os mesmos processos geradores de dados (PGD) de [85]. Seus onze PGD são divididos em dois blocos: (1-5) são séries temporais; e (6-11) são dados transversais. O primeiro PGD de cada bloco é linear (hipótese nula verdadeira) e serve para mensurar o tamanho dos testes, ou seja, se o nível de rejeições dos testes é compatível com o nível de significância escolhido. Os demais PGD são não lineares, e são utilizados para medir a potência (ou poder) dos testes, que vem a ser a frequência de rejeição da hipótese nula uma vez que ela é falsa. O tamanho de amostra utilizado foi $N = 200$.

Para descrever os PGD, convém definir:

- $\{\varepsilon_n\}_{n=1}^{N}$ são realizações independentes de uma distribuição normal com média nula e variância igual a dois;
- $\{x_{1,n}, x_{2,n}\}_{n=1}^{N}$ são realizações independentes (para os diferentes n) de uma distribuição normal bivariada com média nula, variância unitária e correlação ρ;
- $\{S_n\}_{n=1}^{N}$ é uma cadeia de Markov com dois estados, assumindo valores zero e um, com probabilidade de trasição de 30%.

Isto posto, pode-se definir os PGDs:

1. $y_n = 0.6 \cdot y_{n-1} + \varepsilon_n$;

2. $y_n = \begin{cases} 0.9 \cdot y_{n-1} + \varepsilon_n \ , & \text{se } |y_{n-1}| \leq 1 \\ -0.3 \cdot y_{n-1} + \varepsilon_n, & \text{caso contrário;} \end{cases}$

3. $y_n = \begin{cases} 1 + \varepsilon_n \ , & \text{se } y_{n-1} \geq 0 \\ -1 + \varepsilon_n, & \text{caso contrário;} \end{cases}$

4. $y_n = \dfrac{0.7 \cdot |y_{n-1}|}{|y_{n-1}| + 2} + \varepsilon_n$;

5. $y_n = 0.6 \cdot y_{n-1} \cdot (1 - S_n) - 0.3 \cdot y_{n-1} \cdot S_n + \varepsilon_n$;

6. $y_n = 1 + x_{1,n} + x_{2,n} + \varepsilon_n$, com $\rho = 0$;

7. $y_n = 1 + x_{1,n} + x_{2,n} + \varepsilon_n$, com $\rho = 0.7$;

8. $y_n = 1 + x_{1,n} + x_{2,n} + 0.2 \cdot x_{1,n} \cdot x_{2,n} + \varepsilon_n$, com $\rho = 0$;

9. $y_n = 1 + x_{1,n} + x_{2,n} + 0.2 \cdot x_{1,n} \cdot x_{2,n} + \varepsilon_n$, com $\rho = 0.7$;

10. $y_n = 1 + x_{1,n} + x_{2,n} + 0.2 \cdot x_{2,n}^2 + \varepsilon_n$, com $\rho = 0$;

11. $y_n = 1 + x_{1,n} + x_{2,n} + 0.2 \cdot x_{2,n}^2 + \varepsilon_n$, com $\rho = 0.7$.

5.3.2 TNLNs Comparados

Tanto o teste LWG como o caso particular de MGM para estimação[5] de RNAs aqui proposto dependem da geração de unidades escondidas com parâmetros aleatórios. Para ambos os casos, a implementação aqui apresentada seguiu [26] e gerou-os a partir de uma distribuição uniforme no intervalo $[-2, 2]$. Outro paralelo que pode ser feito é entre o número de componentes principais utilizado no teste LWG (q^*) e o número de condições de momento adicionais da estimação por MGM ($K-L$). Seguindo [85], foram testados $q^* = K-L = 2$ e $q^* = K - L = 3$, descartando-se o primeiro componente principal no caso do LWG. No caso das séries temporais, a variável explicativa utilizada foi a primeira defasagem, enquanto as variáveis x_1 e x_2 fizeram este papel para os dados transversais. O teste LWG foi implementado na versão da equação (4-33).

5.3.3 Resultados

A tabela 5.7 traz as proporções de rejeições do teste ao nível de significância de 10%[6]. Para o primeiro e o sexto PGD, espera-se rejeições próximas a 10% enquanto, para os demais, valores mais elevados são preferíveis.

Para o PGD (1), ambos os testes possuem tamanho adequado $q^* = 2$, mas apresentam um viés de sub-rejeição para $q^* = 3$. Todos os testes apresentaram poder semelhante para os PGD de (2) até (5). No PGD (6), o Teste J

[5]No caso do Teste J para NLN, o modelo estimado é linear e a estimação tem baixo custo computacional. Por conta disso, é viável a realização de um grande número de iterações da estimação do ponderador W. Neste experimento, foram aplicadas 10 iterações, apesar de, na maioria dos casos observados, a convergência ocorria por volta da quinta iteração. A cada iteração, estimação do ponderador foi feita através da forma utilizada na equação (3-11).

[6]Assim como em [85], também foram calculadas as rejeições ao nível de significância de 5%, porém optou-se por não apresentá-los uma vez que seguiram exatamente o mesmo padrão observado para o nível de 10% e levariam às mesmas conclusões.

Tabela 5.7: Proporção de Rejeições ao Nível de 10%

	Teste LWG		Teste J	
PGD	$q^* = 2$	$q^* = 3$	$q^* = 2$	$q^* = 2$
1	10.100%	7.900%	10.400%	8.100%
2	38.600%	35.600%	39.300%	36.700%
3	92.100%	92.500%	90.200%	91.900%
4	20.100%	20.000%	20.000%	19.600%
5	13.300%	12.900%	11.200%	10.800%
6	3.000%	4.200%	10.200%	9.900%
7	3.400%	4.300%	9.900%	9.600%
8	7.600%	19.500%	31.100%	26.100%
9	25.800%	22.000%	35.800%	30.300%
10	16.500%	17.000%	29.600%	39.400%
11	27.800%	31.900%	43.900%	38.300%

apresentou um tamanho adequado enquanto o LWG apresentou uma proporção de rejeição muito abaixo do nível de 10%. Nos PGD de (7) a (11), o Teste J mostrou um significativo incremente de potência em relação ao LWG.

Em resumo, pode-se dizer que os TNLNs obtiveram resultados similares para as séries temporais, mas o Teste J saiu-se melhor nos dados transversias. Este resultado é favorável ao Teste J para NLN e, indiretamente, para a estimação de RNAs pelo MGM.

6
Conclusões

Este trabalho teve como núcleo a proposição de um arcabouço geral para a estimação dos parâmetros de RNAs utilizando o MGM sobre-identificado. O principal apelo desta abordagem seria a possibilidade de gerar estimativas em amostra finita mais precisas que a QMV, particularmente por conta de uma menor propensão ao sobreajuste, sem abrir mão das boas propriedades assintóticas, a consistência e a normalidade. Outra potencial vantagem da utilização do MGM seria o Teste J para detecção de não linearidade negligenciada. Tanto o arcabouço geral para estimação de RNAs pelo MGM como o Teste J para não linearidade negligenciada são contribuições inéditas do presente trabalho.

Para fins de aplicação, foi proposto, também, um caso particular de MGM para estimação de RNAs, utilizando condições de ortogonalidade entre os resíduos e combinações lineares de unidades escondidas com parâmetros gerados aleatoriamente. Esta formulação, acoplada ao processo de otimização em três etapas aqui proposto, foi comparada à QMV quanto à precisão na estimação em dois estudos de caso.

O primeiro estudo de caso, utilizando dados simulados a partir de uma RNA acrescida de um ruído gaussiano, indicou que o MGM gera estimações mais precisas que a QMV nos casos em que a razão de ruído é elevada, e que esta vantagem do MGM é maior quanto menor é a amostra utilizada na estimação. Estes resultados são coerentes com a hipótese de que o MGM proposto é menos suscetível a sobreajuste que a QMV.

O segundo estudo de caso utilizou dados reais de dez séries de taxas de câmbio, para as quais há evidências de não linearidade e alto ruído. Os resultados indicaram uma clara vantagem para as RNAs estimadas pelo MGM, corroborando os resultados dos experimentos de Monte Carlo que indicaram a superioridade do MGM em estimações sob elevada razão de ruído.

Houve, ainda, um terceiro estudo de caso, com experimentos de Monte Carlo realizados a fim de comparar as propriedades em pequena amostra do Teste J para NLN com as dos teste LWG, amplamente conhecido e utilizado. Os dois testes tiveram performances semelhantes nas séries temporais, porém, o Teste J saiu-se melhor com os dados transversais.

O conjunto destes resultados configura uma forte evidência a favor da estimação de RNAs pelo MGM em detrimento da QMV, particularmente em

circunstâncias que envolvam altos níveis de ruído.

Dentre as contribuições secundárias deste trabalho, pode-se dizer que a derivação de um limite inferior aproximado para a existência de soluções com sobreajuste em RNAs estimadas por QMV trouxe um grau de formalismo maior ao estudo deste fenômeno, além de desconstruir a ideia amplamente difundida segundo a qual o sobreajuste somente aconteceria em RNAs com excesso de unidades escondidas. Já a formulação da otimização da QMV para RNAs com função de ativação do tipo rampa como um problema de PQIM é a única abordagem já proposta com garantia de otimalidade global, além de poder ser utilizada como método de obtenção do ponto inicial para casos em que outras funções de ativação são utilizadas.

Em suma, o presente trabalho estende estado da arte no campo de abordagens econométricas para RNAs [21, 22] ao proporcionar um método de estimação capaz de conter o sobreajuste, mantendo as propriedades estatísticas desejáveis não encontradas em outros métodos construídos com esta mesma finalidade. Os resultados apresentados dos estudos de caso sedimentam este entendimento. O trabalho pode ser visto como uma nova ponte entre as RNAs o campo da estatística e econometria, através do MGM, que abre possibilidades para futuros desenvolvimentos.

For fim, pode-se mencionar alguns pontos como possíveis desdobramentos deste trabalho. Evidentemente, em relação aos experimentos de Monte Carlo apresentados no primeiro e no terceiro estudos de caso, é sempre possível elaborar novas configurações dos PGDs que podem acrescentar algum grau de entendimento ao problema estudado. Duas outra questões inter-relacionadas que podem ser exploradas futuramente são a otimização dos parâmetros, utilizando outros métodos disponíveis na literatura, e as diferenças entre versões do MGM Eficiente. Outro avanço poderia advir da aplicação de métodos de seleção de condições de momento, como os propostos em [66, 67]. Em termos de aplicações em dados reais, como as do segundo estudo de caso, outras séries financeiras são candidatas a terem melhoras significativas na estimação com a utilização do MGM por conta dos altos níveis de ruído presentes nestes tipos de dados. O mesmo raciocínio vale para modelos de resposta binária, como os utilizados para detecção de fraudes em operações com cartão de crédito, por exemplo.

Referências Bibliográficas

[1] HAYKIN, S.. **Neural Networks: A Comprehensive Foundation.** 2nd Edition, Prentice-Hall, Ontario, 1994.

[2] MCCULLOCH, W. ; PITTS, W.. **A Logical Calculus Of The Ideas Immanent In Nervous Activity.** Bulletin of Mathematical Biophysics, Vol 5, pp. 115-133, 1943.

[3] MINSKY, M.. **Theory Of Neural-Analog Reinforcement Systems And Its Application To The Brain Model Problem.** Tese de Doutorado, Departamento de Matemática, Princeton University, Princeton, NJ, 1954.

[4] ROSENBLATT, F.. **The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain.** Psychological Review, Vol 65, pp. 386-408, 1958.

[5] MINSKY, M.; PAPERT S.. **Perceptrons** The MIT Press, Cambridge, MA, 1969.

[6] OLAZARAN, M.. **A Sociological Study Of The Official History Of The Perceptrons Controversy.** Social Studies of Science 26, no. 3:611-659, 1996.

[7] HOPFIELD, J.. **Neural Networks And Physical Systems With Emergent Collective Computational Abilities.** Proc. of the National Academy of Sciences, USA, Vol 79, pp. 2554-2558, 1982.

[8] RUMELHART, D.; HINTON, G.; WILLIAMS, R.. **Learning Internal Representations By Error Propagation.** Neurocomputing: foundations of research, pp. 673-695, MIT Press Cambridge, MA, USA, 1988.

[9] CYBENKO, G.. **Approximation By Superpositions Of A Sigmoidal Function.** Mathematics of Control, Signals and Systems, Vol 2.4, 303-314, 1989.

[10] SARLE, W.. **Stopped Training And Other Remedies For Overfitting.** Proc. of the 27th Symposium on the Interface of Computing Science and Statistics, pp. 352-360, Fairfax, VA, 1995.

[11] HUNT, K.; SBARBARO, D.; ZBIKOWSKI, R.; GAWTHROP, P.. **Neural Networks For Control Systems - A Survey.** Automatica, Vol 28.6, 1083-1112, 1992.

[12] VELLIDO, A.; LISBOA, P.; VAUGHAN, J.. **Neural Networks In Business: A Survey Of Applications (1992-1998).** Expert Systems with Applications, Vol. 17.1, 51-70, 1999.

[13] WONG, B.; LAI, V.; LAM, J.. **A Bibliography Of Neural Network Business Applications Research: 1994-1998.** Computers Operations Research, Vol. 27.11, 1045-1076, 2000.

[14] EGMONT-PETERSEN, M.; DE RIDDE, D., HANDELS, H.. **Image Processing With Neural Networks - A Review.** Pattern Recognition, Vol. 35.10, 2279-2301, 2002.

[15] WONG, B.; SELVI, Y.. **Neural Network Applications In Finance: A Review And Analysis Of Literature (1990-1996)** Information Management, Vol. 34.3, 129-139, 1998.

[16] PAPIK, K.; MOLNAR, B.; SCHAEFER, R.; DOMBOVARI, Z.; TULASSAY, Z.; FEHER, J.. **Application Of Neural Networks In Medicine - A Review.** Med Sci Monit, Vol. 4(3), 538-546, 1998.

[17] KUAN, C.; WHITE, H.. **Artificial Neural Networks: An Econometric Perspective.** Econometric Reviews, Vol. 13.1, 1-91,1994.

[18] KALOGIROU, S.. **Artiticial Neural Networks In Renewable Energy Systems Applications: A Review.** Renewable and Sustainable Energy Reviews, Vol. 5.4, 373-401, 2001.

[19] NADARAYA, E.. **On Estimating Regression** Theory Of Probability & Its Applications, Vol. 9.1, 141-142, 1964.

[20] WATSON, G.. **Smooth Regression Analysis.** Sankhy a: The Indian Journal of Statistics, Series A, 359-372, 1964.

[21] WHITE, H.. **Learning In Artificial Neural Networks: A Statistical Perspective.** Neural Computation, Vol. 1.4, 425-464, 1989.

[22] WHITE, H.. **Some Asymptotic Results For Learning In Single Hidden-Layer Feedforward Network Models.** Journal of the American Statistical Association, Vol. 84.408, 1003-1013, 1989.

[23] WHITE, H.. **An Additional Hidden Unit Test For Neglected Nonlinearity In Multilayer Feedforward Networks.** Neural Networks. IJCNN., International Joint Conference on. IEEE, 1989.

[24] MEDEIROS, M.; TERÄSVIRTA, T.; RECH, G.. **Building Neural Network Models For Time Series: A Statistical Approach.** Journal of Forecasting, Vol. 25.1, 49-75, 2006.

[25] TERSVIRTA, T.; LIN, C.. **Determining The Number Of Hidden Units In A Single Hidden-Layer Neural Network Model.** Research Report, Bank of Norway, Oslo, 1993.

[26] LEE, T.; WHITE, H; GRANGER, C.. **Testing For Neglected Nonlinearity In Time Series Models: A Comparison Of Neural Network Methods And Alternative Tests.** Journal of Econometrics, Vol. 56.3, 269-290, 1993.

[27] TERSVIRTA, T.; LIN, C.; GRANGER, C.. **Power Of The Neural Network Linearity Test.** Journal of Time Series Analysis, Vol. 14.2, 209-220, 1993.

[28] ZHENG, J.. **A Consistent Test Of Functional Form Via Nonparametric Estimation Techniques.** Journal of Econometrics, Vol. 75.2, 263-289, 1996.

[29] LI, Q.; WANG, S.. **A Simple Consistent Bootstrap Test For A Parametric Regression Function.** Journal of Econometrics, Vol. 87.1, 145-165, 1998.

[30] LEE, T.. **Neural Network Test And Nonparametric Kernel Test For Neglected Nonlinearity In Regression Models.** Studies in Nonlinear Dynamics Econometrics, Vol. 4.4, 169-182, 2001.

[31] SCHWARZ, G.. **Estimating The Dimension Of A Model.** The Annals of Statistics 6, no. 2:461-464, 1978.

[32] ANDERSEN, T.; MARTINEZ, T.. **Cross Validation And Mlp Architecture Selection.** International Joint Conference on Neural Networks, 3, 1614-1619, 1999.

[33] MURATA, N.; YOSHIZAWA, S.; AMARI, S.. **Network Information Criterion-Determining The Number Of Hidden Units For An Artificial Neural Network Model.** IEEE Transactions on Neural Networks, 5.6, 865-872, 1994.

[34] BALKIN, S.; ORD, J.. **Automatic Neural Network Modeling For Univariate Time Series.** International Journal of Forecasting, 16.4, 509-515, 2000.

[35] ANDERS, U.; KORN, O.. **Model Selection In Neural Networks.** Neural Networks, 12.2, 309-323, 1999.

[36] HANSEN, L.. **Large Sample Properties of Generalized Method of Moments Estimators.** Econometrica: Journal of the Econometric Society, 1029-1054, 1982.‘

[37] FAN, J.; PENG, H.. **Nonconcave Penalized Likelihood With A Diverging Number Of Parameters.** The Annals of Statistics, 32.3, 928-961, 2004.

[38] CHONG, G.. **Penalized Likelihood Hazard Estimation.** Purdue University, Department of Statistics, 1991.

[39] HORNIK, K.; STINCHCOMBE, M.; WHITE, H.. **Multilayer Feedforward Networks Are Universal Approximators.** Neural Networks 2.5, 359-366, 1989.

[40] FISHER, R.. **On The Mathematical Foundations Of Theoretical Statistics.** Philosophical Transactions of the Royal Society of London, Série A, 309-368, 1922.

[41] FISHER, R.. **Theory Of Statistical Estimation.** Mathematical Proceedings of the Cambridge Philosophical Society, 22.05, 700-725, 1925.

[42] DAVIDSON, R.; MACKINNON J.. **Estimation And Inference In Econometrics.** Oxford University Press, 1993.

[43] WHITE, H.. **Maximum Likelihood Estimation Of Misspecified Models.** Econometrica, 50, 125, 1982.

[44] BERK, R.. **Limiting Behavior Of Posterior Distributions When The Model Is Incorrect.** The Annals of Mathematical Statistics 37.1, 51-58, 1966.

[45] HUBER, P.. **The Behavior Of Maximum Likelihood Estimates Under Nonstandard Conditions.** Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1967.

[46] SOUZA, G.; GALLANT, R.. **Statistical Inference Based On M-Estimators For The Multivariate Nonlinear Regression Model In**

Implicit Form. Institute of Statistics, North Carolina State University, 1979.

[47] GOURIEROUX, C.; MONFORT A.; TROGNON A.. **Pseudo Maximum Likelihood Methods: Theory.** Econometrica, 681-700, 1984.

[48] FLÔRES JUNIOR, R. G.. **O Método Generalizado dos Momentos (MGM): Conceitos Básicos.** EPGE - Ensaios Econômicos, Rio de Janeiro, 2003.

[49] WOOLDRIDGE, J.. **Applications Of Generalized Method Of Moments Estimation.** The Journal of Economic Perspectives, 15.4, 87-100, 2001.

[50] NEWEY, W.; WEST, K.. **Hypothesis Testing With Efficient Method Of Moments Estimation.** International Economic Review, 777-787, 1987.

[51] GILCHRIST, S.; HIMMELBERG, C.. **Hypothesis Testing With Two-Step GMM Estimators.** No. 95-02, 1995.

[52] FLÔRES JUNIOR, R. G.. **Variáveis Intrumentais E O MGM: Uso De Momentos Condicionais.** EPGE - Ensaios Econômicos, Rio de Janeiro, 2003.

[53] HANSEN, L. P.; HEATON, J.; YARON, A.. **Finite-Sample Properties Of Some Alternative GMM Estimators.** Journal of Business Economic Statistics, 14.3, 262-280, 1996.

[54] NEWEY, K.; MCFADDEN, D.. **Large Sample Estimation And Hypothesis Testing.** Handbook of econometrics, Vol. 4, Cap. 36, 2111-2245, 1994.

[55] NEWEY, W.; WEST, K.. **A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.** Econometrica, 55, 703708, 1986.

[56] NEWEY, W.. **Generalized Method Of Moments Specification Testing.** Journal of Econometrics, 29.3, 229-256, 1985.

[57] HAYAKAWA, K.. **Alternative Over-Identifying Restriction Test In GMM Estimation Of Panel Data Models.** ISER Seminar Series, 2013.

[58] LUENBERGER, D.; YE, Y.. **Linear And Nonlinear Programming**, Vol. 116, Springer Science Business Media, 2008.

[59] SHANG, Y.; WAH, B.. **Global Optimization For Neural Network Training.** Computer 29, no. 3, 45-54, 1996.

[60] SEXTON, R.; ALIDAEE, B.; DORSEY, R.; JOHNSON, J.. **Global Optimization For Artificial Neural Networks: A Tabu Search Application.** European Journal of Operational Research 106, no. 2, 570-584, 1998.

[61] ZANCHETTIN, C. **Otimização Global em Redes Neurais Artificiais.** Tese de Doutorado, Centro de Informática, Universidade Federal de Pernambuco, Recife, 2008.

[62] GAD, E.; ATIYA, A.; SHAHEEN, S.; EL-DESSOUKI, A.. **A New Algorithm For Learning In Piecewise-Linear Neural Networks.** Neural Networks 13, no. 4, 485-505, 2000.

[63] KOZUB, D.; HOLENA, M.. **Learning Of Multilayer Perceptrons With Piecewise-Linear Activation Functions.** Dissertação de Mestrado, Faculdade de Informática e Ciências, Czech Technical University, Praga, 2007.

[64] LAND, A.; DOIG, A.. **An Automatic Method Of Solving Discrete Programming Problems.** Econometrica: Journal of the Econometric Society, 497-520, 1960.

[65] LEE, T.; XI, Z;,ZHANG, R.. **"Testing For Neglected Nonlinearity Using Artificial Neural Networks With Many Randomized Hidden Unit Activations.** Journal of Time Series Econometrics 5, no. 1, 61-86, 2012.

[66] HALL, A.; PEIXE, F.. **A Consistent Method For The Selection Of Relevant Instruments.** Econometric Reviews, 22(3), 269-287, 2003.

[67] HALL, A.; INOUE, A.; JANA, K.; SHIN, C.. **Information In Generalized Method Of Moments Estimation And Entropy-based Moment Selection.** Journal of Econometrics, 138(2), 488-512, 2007.

[68] ELSTER, C.; NEUMAIER, A.. **A Trust-Region Method For The Optimization Of Noisy Functions.** Computing, v. 58, p. 31-46, 1997.

[69] ANDREWS, D.. **A Stopping Rule For The Computation Of Generalized Method Of Moments Estimators.** Econometrica, 55, 913-931, 1997.

[70] MEESE, R.; ROGOFF, K.. **Empirical Exchange Rate Models Of The Seventies: Do They Fit Out Of Sample?** Journal of International Economics, 14.1, 3-24, 1983.

[71] MEESE, R.; ROGOFF, K.. **The Out-Of-Sample Failure Of Empirical Exchange Rate Models: Sampling Error Or Misspecification?** Exchange Rates and International Macroeconomics, University of Chicago Press, 67-112, 1983.

[72] TAYLOR, M.. **The Economics Of Exchange Rates.** Journal of Economic Literature, 13-47, 1995.

[73] CHEUNG, Y; CHINN, D.. **Macroeconomic Implications Of The Beliefs And Behavior Of Foreign Exchange Traders.** No. 7417. National Bureau of Economic Research, 1999.

[74] KILIAN, L.; TAYLOR , M.. **Why Is It So Difficult To Beat The Random Walk Forecast Of Exchange Rates?.** Journal of International Economics, 60.1, 85-107, 2003.

[75] BALKE, N.; FOMBY, T.. **Threshold Cointegration.** International Economic Review, 627-645, 1997.

[76] TAYLOR, M.; PEEL, D.. **Nonlinear Adjustment, Long-Run Equilibrium And Exchange Rate Fundamentals.** Journal of International Money and Finance, 19.1, 33-53, 2000.

[77] MEDEIROS, M.; VEIGA, A.; PEDREIRA, C.. **Modeling Exchange Rates: Smooth Transitions, Neural Networks, And Linear Models.** IEEE Transactions on Neural Networks , 12.4, 755-764, 2001.

[78] BOERO, G.; MARROCU,E.. **The Performance Of Non-Linear Exchange Rate Models: A Forecasting Comparison.** Journal of Forecasting, 21.7, 513-542, 2002.

[79] KUAN, C.;,LIU, T.. **Forecasting Exchange Rates Using Feedforward And Recurrent Neural Networks.** Journal of Applied Econometrics, 10.4, 347-364, 1995.

[80] YAO, J.; POH, H.; JASIC, T. **Foreign Exchange Rates Forecasting With Neural Networks.** International Conference on Neural Information Processing, Hong Kong, 1996.

[81] ANDREOU, A.; GEORGOPOULOS, E.;. LIKOTHANASSIS, E.. **Exchange-Rates Forecasting: A Hybrid Algorithm Based On Genetically Optimized Adaptive Neural Networks.** Computational Economics, 20.3, 191-210, 2002.

[82] NAG, A.; MITRA, A.. **Forecasting Daily Foreign Exchange Rates Using Genetically Optimized Neural Networks.** Journal of Forecasting, 21.7, 501-511, 2002.

[83] KAMRUZZAMAN, J.; SARKER, R.. **Forecasting Of Currency Exchange Rates Using Ann: A Case Study.** Proceedings of the 2003 International Conference on Neural Networks and Signal Processing, Vol. 1. IEEE, Nanjing, China, 2003.

[84] YU, L.; LAI, K.; WANG, S.. **Multistage RBF Neural Network Ensemble Learning For Exchange Rates Forecasting.** Neurocomputing, 71.16,3295-3302, 2008.

[85] LEE, T,; XI, Z.; ZHANG, R.. **Testing For Neglected Nonlinearity Using Artificial Neural Networks With Many Randomized Hidden Unit Activations** Journal of Time Series Econometrics, Vol. 5.1, 61-86, 2012.

A
Geração Unidades Escondidas com Parâmetros Aleatórios

Nos processos de inicialização das otimizações das estimações, tanto por QMV como pelo MGM, precisa-se de um conjunto de $\{\vec{\omega}_m\}_{m=1}^{M}$ gerados aleatoriamente. Para facilitar a imposição de restrições sobre os parâmetros, convém defini-los considerando que os regressores não constantes estão normalizados entre zero e um. A álgebra para encontrar os $\{\vec{\omega}_m\}_{m=1}^{M}$ equivalentes para os regressores não normalizados é trivial.

O processo é inicializado pelo sorteio dos $\{\omega_{m,1}\}_{m=1}^{M}$, que multiplicam as constantes (ou os vieses na nomenclatura de RNAs tradicional). Eles são sorteados de uma distribuição uniforme no intervalo $[-25, 0]$, e associados às unidades escondidas em ordem crescente de acordo com m, a fim de evitar múltiplas parametrizações redundantes.

Em seguida, para cada unidade escondida, os demais parâmetros são sorteados de uma distribuição uniforme no intervalo $[-25 - \omega_{m,1}, 25 - \omega_{m,1}]$ e, em seguida, são testado para checar se eles atendem às seguintes restrições:

- $min\left(\omega_{m,1}, \sum_{i\leq 1} \omega_{m,i}\right) \leq 3$;
- $\max\left(\omega_{m,1}, \sum_{i\leq 1} \omega_{m,i}\right) \geq -3$;
- $\max\left(|\omega_{m,1}|, \left|\sum_{i\leq 1} \omega_{m,i}\right|\right) \geq 3$; e
- $\sum_{i\leq 1} abs(\omega_{m,i}) \geq 2$.

As três primeiras restrições previnem que todos os pontos fiquem longe de zero, na região mais plana de função de ativação. A quarta restrição evita pontos muito próximos na função de ativação, garantindo um mínimo de não linearidade. Se uma das restrições não for atendida, novos sorteios são realizados, até a obtenção de um vetor de parâmetros válido.

B
Eficiência da Máxima Verossimilhança

A eficiência do estimador de MV depende do atendimento das chamadas condições de regularidade sobre a função de log-verossimilhaça $l(\psi) = log(L(\psi))$:

1. $l(\psi)$ é três vezes diferenciável e as derivadas são contínuas e limitadas para todo $\psi \in \Psi$;

2. o valor esperado das duas primeiras derivadas de $l(\psi)$ existem e são finitas;

3. existe uma função com valor esperado finito que é uma cota superior do módulo da terceira derivada de $l(\psi)$ para todo $\psi \in \Psi$.

Atendidas estas condições de regularidade, garante-se que a covariância do estimador de MV convergirá assintóticamente para a inversa da Matriz de Informação de Fisher, que vem a ser igual a menos a esperança da matriz hessiana da função de log-verossimilhança. Esta é covariância a mínima entre todos os estimadores consistentes e assintoticamente normais, por isso é conhecida o Limite Inferior de Cramér-Rao.

A título de exemplo, no caso de estimações em que a amostra obtida de uma distribuição gaussiana, independente e identicamente distribuída, as condições de regularidade são atendidas e o estimador de MV é eficiente. Por outro lado, se a distribuição for uniforme, não é dificil perceber que que a primeira condição de regularidade não será atendida, uma vez a função de verossimilhança atinge valor zero para alguns parâmetros, o que faz com que a log-verossimilhança seja ilimitada.

Apesar de a eficiência ter sido um grande impulsionador da utilização da MV, as condições de regularidade são muitas vezes negligenciadas.

C
Viés do MGM em Pequena Amostra

Um pequeno experimento foi montado a fim de exemplificar o viés do MGM em pequenas amostras e a influência do número de condições de momento pode interferir. Neste experimento, 500 amostras de tamanho 200 foram geradas a partir de uma distribuição de Poisson com parâmetro $\lambda = 1$. Para cada amostra, o parâmetro foi estimado através MGM com um a seis condições de momento[1]. Foram utilizadas condições de momento utilizadas da forma $E[x^k]$, onde as constantes k foram sorteadas (para cada estimação) de uma distribuição uniforme entre 0 e 3. Os resultados estão expostos na tabela C.1.

Tabela C.1: Resultados dos Experimentos

Condições de Momento	Viés Absoluto	Desvio Padrão	REQM
1	0,001877	0,080082	0,080104
2	0,004054	0,073557	0,073669
3	0,00821	0,074867	0,075316
4	0,013864	0,076519	0,077765
5	0,017523	0,077897	0,079844
6	0,021645	0,07885	0,081767

A tabela C.1 deixa clara a relação direta entre o viés e a quantidade de condições de momento. O MGM com uma única condição de momento teve o maior desvio padrão entre os MGMs testados. Para duas ou mais condições de momento, também observa-se uma relação direta entre o desvio padrão e o número de condições de momento. Apesar de ter o menor viés, o MGM com uma condição de momento obteve a segunda maior REQM.

No caso deste experimento, um ponto razoável de equilíbrio entre viés e variância seria o MGM com duas condições de momento. Convém ressaltar que este resultado é extremamente particular.

[1]Os ponderadores foram obtidos pelo método do MGM eficiente continuamente atualizado.

D
O Algoritmo BFGS

Em otimização numérica, o algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS), comumente referido pela sigla BFGS, enquadra-se na categoria de métodos Quasi-Newton, que abrange variantes do método Newton-Rhapson.

O Método Newton-Rhapson para minimização (sem perda de generalidade) é baseado em um processo iterativo de substituição do último ponto pelo mínimo da Expansão de Taylor de segunda ordem ao seu redor, que podem ser calculadas analiticamente. O procedimento pode ser descrito como:

$$x_{k+1} = x_k + \Delta x_k, \tag{D-1}$$

onde x_k é o ponto na k-ésima iteração, $\Delta f(x_k)$, conhecido como passo, satisfaz $B \cdot \Delta f(x_k) = \nabla f(x_k)$, sendo B a matriz Hessiana e $\nabla f(\cdot)$.Geralmente, método de Newton-Rhapson converge em um número de iterações que os métodos de gradiente, e com menos tempo de processamento.

Não obstante, o cálculo de B pode evolver um alto custo computacional. Os métodos Quasi-Newton utilizam aproximações iterativamente calculados de B ou de B^{-1}. A principal diferença entre os diferentes métodos desta categoria está na forma como esse processo iterativo é feito.

No caso do algoritmo BFGS, iniciando-se com em um ponto x_k com uma matriz Hessiana B_k (que pode ser uma matriz identidade ou a Hessiana propriamente avaliada x_0) para $k = 0$, o processo é dado pela repetição iterativa, até a convergência, dos seguintes passos:

1. calcula-se a direção do passo p_k resolvendo $p_k = B_k^{-1} \cdot \nabla f(x_k)$;

2. obtém-se um cumprimento de passo α_k que atenda às Condições de Wolf{Par de restrições que garante a positividade definida de B_k ao longo das iterações. Detalhes em [58]. ;

3. atualiza-se o ponto $x_{k+1} = x_k + s_k$, onde $s_k = \alpha_k \cdot p_k$;

4. define-se $y_k = \nabla f(x_{k+1}) \cdot \nabla f(x_k)$;

5. atualiza-se a aproximação da matriz Hessiana:

$$B_{k+1} = B_k + \frac{y_k \cdot y_k'}{y_k' \cdot s_k} - \frac{B_k \cdot s_k \cdot s_k' \cdot B_k}{s_k' \cdot B_k \cdot s_k}. \tag{D-2}$$

Em particular, para problemas relacionados a RNAs, o BFGS é um método bastante popular. Hoje em dia, encontra-se implementado em pacotes comerciais, como o *Neural Networks Toolbox*, do *Matlab*, e em *softwares* livres, como no pacote *nnet*, que roda sobre a linguagem de programação *R*.

E
Análise Viés do MGM nos Expeimentos da Seção 5.1

Os experimentos da seção 5.1 compararam estimações por QMV e MGM em termos de aderência entre a função esperança condicional estimada e a verdadeira. Porém, como foi mencionado na seção 3.4 e exemplificado no apêndice C, em muitos casos, a inclusão de condições de momento acarretar aumento viés dos estimadores dos parâmetros.

Este apêndice traz um pequeno experimento auxiliar elaborado a fim de verificar a ocorrência desse fenômeno nos experimentos da seção 5.1. Nos experimento auxiliar foram incluídas amostrasde 100 e 400 observações, o número de repetições foi reduzido de 1000 para 100 por configuração e foram anotados o viés (absoluto) e a variância de cada estimador do MGM em cada configuração.

Em seguida, para cada parâmetro, foram rodadas as seguintes regressões:

$$log(Viés_{M,N,q}) = \beta_0 + \beta_1 \cdot (M-8) + \beta_2 \cdot I_{N=100} + \beta_3 \cdot I_{N=200} + \beta_4 \cdot I_{N=400} + \\ + \beta_5 \cdot I_{N=800} + \beta_6 \cdot I_{q=30\%} + \beta_7 \cdot I_{q=50\%} + \beta_8 \cdot I_{q=70\%} + \beta_9 \cdot I_{q=90\%} \quad \text{(E-1)}$$

e

$$log(Variância_{M,N,q}) = \phi_0 + \phi_1 \cdot (M-8) + \phi_2 \cdot I_{N=100} + \phi_3 \cdot I_{N=200} + \phi_4 \cdot I_{N=400} + \\ + \phi_5 \cdot I_{N=800} + \phi_6 \cdot I_{q=30\%} + \phi_7 \cdot I_{q=50\%} + \phi_8 \cdot I_{q=70\%} + \phi_9 \cdot I_{q=90\%}, \quad \text{(E-2)}$$

onde M é o número de condições de momento, N é o tamanho de amostra, q é a razão de ruído e I são variáveis do tipo *dummy*[1]. Os resultados para β_1 e $\phi_{1]}$, qu são apresentados na tabela E.1.

A análise dos parâmetros β_1 indica que o viés é negativamente relacionado à quantidade de condições de momento nos experimentos da seção 5.1. A negatividade de todos os β_1 foi significativa a 5% em todas as regressões. Da mesma forma, os valores negativos (significativos a 1%) de ϕ_1 indicam que a variância dos estimadores também é negativamente relacionada ao número de condições de momento.

O resultado referente ao viés é interessante, uma vez que contradiz a ideia de que mais condições de momento implicam em mais viés. Pode-se conjecturar

[1]Formulações alternativas, sem a utilização de variáveis *dummy* também foram testadas com resultados similares.

Tabela E.1: Resultados das Regressões do Experimento Auxiliar

Parâmetro	β_1	Prob$[\beta_1 > 0]$	ϕ_1	Prob$[\phi_1 > 0]$
α_1	-0,86523	$< 10^-6$	-1,73218	$< 10^-6$
α_2	-0,34433	$< 10^-6$	-0,64891	$< 10^-6$
λ_1	-0,59375	$< 10^-6$	-1,09526	$< 10^-6$
λ_2	-0,84514	$< 10^-6$	-1,72592	$< 10^-6$
$\omega_{1,1}$	-0,08029	0,011211	-0,21309	$< 10^-6$
$\omega_{1,2}$	-0,07372	0,020964	-0,22859	$< 10^-6$
$\omega_{2,1}$	-0,11897	0,000105	-0,22736	$< 10^-6$
$\omega_{2,2}$	-0,10067	0,000729	-0,27087	$< 10^-6$

que este resultado inusitado se deve às características particulares da estimação de RNAs, em particular, ao sobreajuste, ou fato de que cada condição de momento adicional apresenta uma forma mais complexa de não-linearidade.