

## 2 Sistemas multi-previsores

### 2.1. Introdução

Sistemas **multi-previsores** são sistemas constituídos de dois ou mais previsores, de mesma natureza ou não; nestes sistemas, há sempre algum tipo de integração entre os previsores, com o objetivo de compensar erros individuais e melhorar assim o desempenho final. É importante salientar o fato de que os métodos de previsão integrados em um sistema multi-previsor podem ter naturezas completamente diferentes, aspecto que vai ao encontro dos conceitos de diversidade e mitigação do risco.

Com relação à diversidade, pode-se propor uma classificação geral dos métodos de previsão em **quantitativos** (objetivos) ou **qualitativos** (subjetivos). Os métodos quantitativos são fundamentados em ciências exatas e geram previsões de maneira automática; já os métodos qualitativos são essencialmente manuais, com previsões derivadas do conhecimento tácito dos técnicos envolvidos. Na prática, os métodos quantitativos são ainda subdivididos em (i) univariados, que levam em conta apenas a variável que está sendo prevista (e versões defasadas dela), ou (ii) causais<sup>5</sup>, que levam em conta a variável que está sendo prevista, chamada endógena, e um conjunto de outras, chamadas exógenas. Em um segundo nível taxonômico, os métodos quantitativos também podem ser classificados de acordo com o tipo de previsão que realizam: direta ou iterativa (Marcellino et al. 2006). Fundamentos sobre as principais metodologias de previsão utilizadas na prática, principalmente do ponto de vista quantitativo, podem ser encontradas em Harvey (1991), Pankratz (1991), Makridakis et al. (1998), Souza & Camargo (2004) e Box et al. (2008).

---

<sup>5</sup> Também chamados de métodos multivariados, de função de transferência ou de regressão.

Os experimentos realizados neste trabalho utilizam apenas métodos quantitativos; contudo, a integração entre métodos quantitativos e qualitativos é um tema também estudado (Webby & O'Connor, 1996; Goodwin, 2002). Atualmente, a discussão prática a respeito da integração “quantitativo-qualitativo” gira em torno da seguinte questão: os ajustes subjetivos (manuais) estão realmente agregando valor sobre a linha de base fornecida pelas previsões quantitativas (estatísticas) (Chase, 2009; Fildes et al. 2009)? Embora parte da comunidade científica tenha apontado a incorporação de conhecimento especialista (humano) como um dos caminhos<sup>6</sup> para a melhoria na atividade de previsão (Franses, 2004), nem sempre os ajustes qualitativos têm contribuído para a melhoria do desempenho final (Chase, 2009). Isto valoriza a área de combinação de previsores como opção para a melhoria do desempenho de previsão, independente da natureza dos previsores disponíveis.

## 2.2. Taxonomia

Inspirado nos conceitos taxonômicos utilizados em Sharkey (1999), este trabalho sugere uma estrutura de classificação para os sistemas multi-previsores. Esta estrutura se divide em dois grandes grupos, de acordo com o tipo de associação existente entre a resposta consensual (previsão do sistema) e os previsores componentes:

1. **Sistemas de combinação modular.** Formados pela composição, e.g. soma, de previsores **não redundantes**. O termo não redundante quer dizer que cada previsor componente, independente de sua natureza ou capacidade real, é responsável apenas por uma parte da previsão. Em outras palavras, todo componente não redundante é parte essencial do esquema de combinação proposto.
2. **Sistemas de combinação por comitê (*ensembles*).** Formados pela agregação, e.g. média, de previsores **redundantes**. O termo redundante

---

<sup>6</sup> Outros caminhos citados foram (i) melhorar a extração de atributos relevantes e (ii) diminuir o vão entre universidade e empresa.

significa que cada previsor componente fornece por si só uma previsão completa, idealmente exclusiva, para a variável de interesse. A palavra exclusiva reforça que as previsões neste tipo de sistema não precisam ser idênticas ou mesmo correlacionadas; ao contrário, é até desejável que elas sejam **complementares** entre si, ou seja, compensem seus erros individuais. Com a redundância de seus componentes, os sistemas de comitê têm escalabilidade: pode-se adicionar ou remover previsores sem a descaracterização do esquema de combinação proposto.

A Figura 2 ilustra a taxonomia proposta. O presente trabalho está focado nos modelos do grupo (b), na medida em que propõe uma metodologia para combinação linear de previsores redundantes. Por outro lado, vale ressaltar que é sempre possível a construção de sistemas híbridos, reunindo características tanto de (a) quanto de (b).

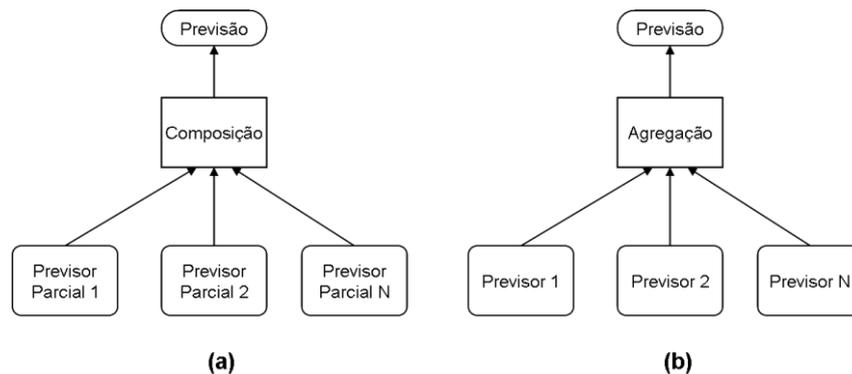


Figura 2 – Combinação modular (a) e combinação por comitê (b).

### 2.2.1. Combinação modular

Um exemplo de sistema modular está no trabalho de Lemos & Fogliatto (2008). Nesse trabalho, os autores escolhem um modelo estatístico (quantitativo) como linha de base para previsão de demanda e depois aplicam ajustes (qualitativos) baseados na metodologia Delphi (Wright & Giovinazzo, 2000; Gordon, 2007). Em outro exemplo, Zang (2003) propõe uma metodologia em dois passos: primeiro, um modelo ARIMA é utilizado para prever a parte linear de um problema; depois uma rede neural tenta agregar valor à previsão final, atuando

sobre os resíduos do primeiro passo. Outra aplicação de sistema modular, semelhante a este último exemplo, pode ser vista em Pai & Lin (2005).

### 2.2.2. Combinação por comitê (*ensembles*)

Chase (2009) apresenta um exemplo prático de comitê de previsores, combinando linearmente e com pesos iguais (média simples) três modelos quantitativos: dois univariados (amortecimento exponencial e ARIMA) e um causal. Outro exemplo interessante está no trabalho de Kurrle (2004), em que se combina (por média simples) previsões quantitativas (amortecimento exponencial e ARIMA) e subjetivas (Delphi).

A seguinte conclusão empírica – “mesmo que o melhor predictor possa ser identificado a cada instante, a combinação pode ainda ser uma estratégia atraente, por oferecer ganho potencial em diversificação” (Timmerman, 2006) – leva a uma segunda afirmativa: em um comitê é desejável que haja previsores complementares, i.e., previsores que compensem seus erros individuais ou, em outras palavras, que tenham baixa **correlação de erro** (correlação de erro é a correlação medida entre os erros de previsão). A busca pela combinação de previsores complementares pode seguir dois caminhos: (i) combinar previsores de naturezas diferentes ou (ii) combinar previsores que sejam variantes da mesma técnica. A obtenção de variantes pode ocorrer, por exemplo, das seguintes maneiras: somando ou subtraindo constantes nas previsões finais (seção 4.4.3), usando hiperparâmetros diferentes no algoritmo de previsão ou utilizando alguma técnica de reamostragem, e.g. *bootstrap* (Efron & Tibshirani, 1993), convenientemente adaptada (Souza & Camargo, 2004; Cordeiro & Neves, 2010).

## 2.3. Modelos de comitê

### 2.3.1. Linear

A combinação linear de  $N$  previsores no instante  $t + h$ , estimada com dados até  $t$ , tem a seguinte forma geral:

$$y_{t+h|t}^C = f(\hat{\mathbf{y}}_{t+h|t}, \hat{\mathbf{w}}_{t+h|t}) = f^{linear}(\hat{\mathbf{w}}_{t+h|t}) \quad (6)$$

onde  $\hat{\mathbf{y}}_{t+h|t}$  é o vetor de previsões disponíveis para  $t + h$  e  $\hat{\mathbf{w}}_{t+h|t}$  é o vetor de pesos estimados. Expandindo:

$$\hat{\mathbf{y}}_{t+h|t} = [\hat{y}_{t+h|t,1} \quad \hat{y}_{t+h|t,2} \quad \dots \quad \hat{y}_{t+h|t,N}]' \quad (7)$$

$$\hat{\mathbf{w}}_{t+h|t} = [\hat{w}_{t+h|t,1} \quad \hat{w}_{t+h|t,2} \quad \dots \quad \hat{w}_{t+h|t,N}]' \quad (8)$$

onde  $\hat{y}_{t+h|t,k}$  é a previsão  $h$  passos a frente feita pelo  $k$ -ésimo previsor, no instante  $t$ , e  $\hat{w}_{t+h|t,k}$  é o peso associado a esta previsão.

Combinações lineares são geralmente desejáveis por sua simplicidade matemática e facilidade de interpretação, principalmente pelo fato de haver uma relação biunívoca (um para um) entre pesos e previsores. Por outro lado, esquemas não lineares como os relatados por Donaldson & Kamstra (1996) e Mubwandarikwa (2007), apesar de apresentarem bons resultados em diversas situações, têm contra si o fato de serem complexos ou de difícil interpretação. O formato mais comum para combinação linear de previsores é o seguinte:

$$y_{t+h|t}^C = \sum_{k=1}^N \hat{w}_{t+h|t,k} \hat{y}_{t+h|t,k} \quad (9)$$

onde

$$\sum_{k=1}^N \hat{w}_{t+h|t,k} = 1 \quad e \quad \hat{w}_{t+h|t,k} \geq 0 \quad (10)$$

Além da facilidade na interpretação dos pesos, o esquema de combinação nas equações (9) e (10) é altamente desejável por dois outros motivos: (i) garantia de **não tendenciosidade** na previsão combinada, se os previsores envolvidos forem não tendenciosos e (ii) garantia de **convexidade** na combinação. As características de não tendenciosidade e convexidade são mais exploradas na seção 2.4.3. Ainda com relação à capacidade de interpretação dos pesos, ela ajuda

a justificar o estudo das combinações lineares neste trabalho, já que muitas vezes é um requisito determinante em aplicações reais.

### 2.3.2. Não linear

A equação (11) representa um esquema genérico de combinação não linear de previsores.

$$y_{t+h|t}^C = f(\hat{\mathbf{y}}_{t+h|t}, \hat{\mathbf{w}}_{t+h|t}) = f^{\text{não linear}}(\hat{\mathbf{w}}_{t+h|t}) \quad (11)$$

É importante observar que, no caso não linear, não há necessariamente uma relação biunívoca entre pesos e previsores. Em outras palavras, os pesos  $\hat{w}$  passam a ser parâmetros implícitos no modelo, sem interpretação direta. A dificuldade de interpretação dos pesos obtidos é uma das principais críticas aos modelos de combinação não lineares, tais como os baseados diretamente em redes neurais (Donaldson & Kamstra, 1996).

Um caso particular onde existe relação um para um entre pesos e previsões é a combinação geométrica, definida pelas equações (12) e (13). Este esquema de combinação foi usado com relativo sucesso em uma aplicação com previsões bayesianas (Mubwandarikwa, 2007).

$$y_{t+h|t}^C = \prod_{k=1}^N \hat{y}_{t+h|t,k}^{\hat{w}_{t+h|t,k}} \quad (12)$$

onde

$$\sum_{k=1}^N \hat{w}_{t+h|t,k} = 1 \quad (13)$$

### 2.3.3. Outros

A metodologia conhecida como *stacked generalization* (Wolpert, 1992), adaptada à tarefa de previsão na Figura 3, é um exemplo genérico de comitê. Nela, a etapa de agregação é executada por um previsor especial, situado em um nível hierárquico superior aos demais. Wolpert cita informalmente que neste esquema

os modelos de nível básico (*baselevel generalizers*) devem cobrir o espaço de generalização e serem mutuamente ortogonais; em outras palavras, eles devem ser redundantes e complementares, como na definição de comitê (seção 2.2).

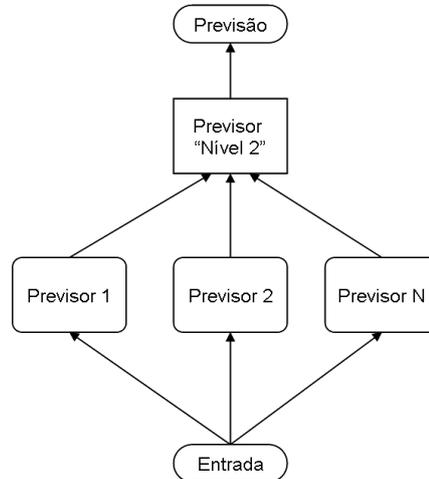


Figura 3 – *Stacked generalization*.

A Figura 4 exibe um esquema de combinação conhecido como mistura de *experts* (Jacobs et al. 1991). Neste esquema, há um mecanismo binário (*1 de N*) ativado pelo próprio dado de entrada e responsável pela seleção do previsor potencialmente melhor naquele momento (o previsor selecionado recebe peso **um** enquanto todos os outros recebem peso **zero**).

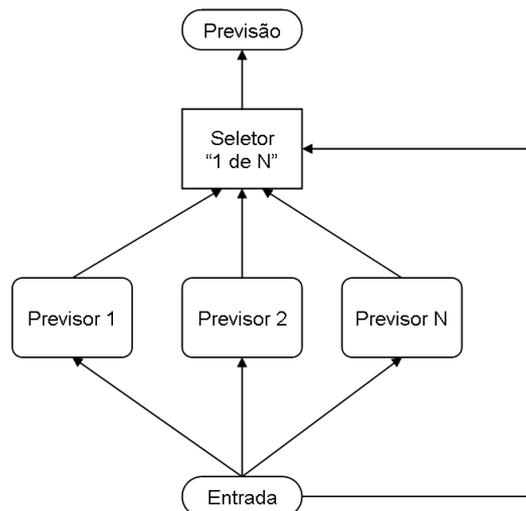


Figura 4 – Mistura de experts.

Por último, apresentam-se os sistemas que realizam combinação/seleção de previsores através de regras (Arinze, 1994; Adya et al. 2001; Wang et. al. 2009;

Lemke & Gabrys, 2010). Este tipo de sistema utiliza um conjunto de regras **se-então** – conhecido como base de conhecimento – para ponderar a contribuição de cada um dos previsores disponíveis ou mesmo selecionar um único predictor, considerado mais adequado que todos os demais. A Figura 5 exibe uma base de conhecimento onde se realiza a seleção do melhor entre 6 previsores possíveis: média móvel (*mov\_avg*), amortecimento exponencial simples (*expon*), Holt simples (*holts*), Holt-Winters (*winters*), filtro adaptativo (*adap*) e decomposição (*time\_dec*); os *antecedentes* das regras são compostos por atributos como sazonalidade (*seasonality*), variabilidade ou número de pontos de inflexão (*variability*), tendência (*trend*), coeficiente de determinação (*r\_sq*) e erro quadrático médio (*MSE*).

```

---- start of rule ----
1: R_SQ??
2: | <0.2599:MSE??
3: | | <3.866:----- expon
4: | | | ≥3.866:VARIABILITY??
5: | | | | <9.:----- holts
6: | | | | | ≥9.:TREND??
7: | | | | | | <-11571.4:----- expon
8: | | | | | | | ≥-11571.4:R_SQ??
9: | | | | | | | | <0.05555:----- time_dec
10: | | | | | | | | | ≥0.05555:MSE??
11: | | | | | | | | | | <5890000000.:SEASONALITY??
12: | | | | | | | | | | | <-0.0825:----- holts
13: | | | | | | | | | | | | ≥-0.0825:----- winters
14: | | | | | | | | | | | | | ≥5890000000.:----- time_dec
15: | | | | | | | | | | | | | ≥0.2599:SEASONALITY??
16: | | | | | | | | | | | | | | <0.9021:MSE??
17: | | | | | | | | | | | | | | | <11.64:TREND??
18: | | | | | | | | | | | | | | | | <0.57505:----- holts
19: | | | | | | | | | | | | | | | | | ≥0.57505:VARIABILITY??
20: | | | | | | | | | | | | | | | | | | <6.:VARIABILITY??
21: | | | | | | | | | | | | | | | | | | | <3.5:----- time_dec
22: | | | | | | | | | | | | | | | | | | | | ≥3.5:----- mov_avg
23: | | | | | | | | | | | | | | | | | | | | | ≥6.:----- expon
24: | | | | | | | | | | | | | | | | | | | | | ≥11.64:TREND??
25: | | | | | | | | | | | | | | | | | | | | | | <0.20875:VARIABILITY??
26: | | | | | | | | | | | | | | | | | | | | | | | <7.:----- mov_avg
27: | | | | | | | | | | | | | | | | | | | | | | | | ≥7.:----- adap
28: | | | | | | | | | | | | | | | | | | | | | | | | ≥0.20875:TREND??
29: | | | | | | | | | | | | | | | | | | | | | | | | | <11595.5:SEASONALITY??
30: | | | | | | | | | | | | | | | | | | | | | | | | | | <0.8327:R_SQ??
31: | | | | | | | | | | | | | | | | | | | | | | | | | | | <0.9126:----- holts
32: | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥0.9126:VARIABILITY??
33: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | <0.5:----- adap
34: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥0.5:----- holts
35: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥0.8327:----- adap
36: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥11595.5:TREND??
37: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | <55250.:----- adap
38: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥55250.:----- expon
39: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥0.9021:TREND??
40: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | <45.6:----- holts
41: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥45.6:VARIABILITY??
42: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | <1.:----- holts
43: | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ≥1.:----- time_dec
---- end of rule ----

```

Figura 5 – Base de conhecimento para seleção de previsores (Arinze, 1994).

Uma grande dificuldade em sistemas dessa natureza é a definição das regras que vão compor a base de conhecimento e assim reger a combinação/seleção dos previsores disponíveis.

## 2.4. Métodos tradicionais para estimação de pesos lineares

Nesta seção, revisa-se alguns dos métodos de ponderação linear mais citados ou sugeridos na literatura. Excluindo-se a média simples (seção 2.4.1), uma característica comum destes métodos é que os pesos gerados múltiplos passos a frente são função dos desempenhos individuais dos previsores, medidos **dentro da amostra**. A Figura 6 exemplifica.

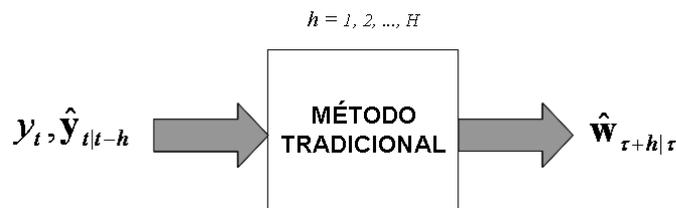


Figura 6 – Diagrama representativo de um método tradicional de ponderação. A informação é processada com janela de tempo fixa ou expansiva.

Como acontece em diversas metodologias relacionadas à séries temporais – incluindo o sistema NEW (capítulo 3) – os métodos tradicionais de ponderação processam as informações disponíveis considerando uma **janela de tempo**, que pode ser **fixa** ou **expansiva**. Considerando janela fixa de tamanho  $v$ , o cálculo do vetor de pesos para o instante  $t$  leva em conta informações anteriores (ou mesmo simultâneas) a  $t$ , mas apenas as  $v$  informações mais recentes; informações “antigas” são descartadas. Por outro lado, considerando janela expansiva, o cálculo do vetor de pesos para o instante  $t$  leva em conta toda a informação anterior (ou simultânea) a  $t$ ; não há descarte de informação.

### 2.4.1. Média simples

Tirar a média das previsões é o método mais simples de combinação linear de previsores; ainda assim, é difícil de ser superado (Timmermann, 2006). Fazer a média simples no horizonte  $h$  é, em outras palavras, utilizar o seguinte valor de peso para cada um dos  $N$  previsores disponíveis:

$$\hat{w}_{t+h|t,k} = 1/N \quad (k = 1, 2, \dots, N) \quad (14)$$

### 2.4.2. Mínimos quadrados irrestritos

A estimação de pesos por mínimos quadrados irrestritos é um exemplo clássico de otimização aplicada à estatística. A ideia é que se obtenha os pesos minimizando o erro quadrático ao longo das observações disponíveis. No caso linear esta minimização é expressa como se segue, considerando janela de tempo de tamanho  $v$  e horizonte de previsão  $h$ :

$$\hat{\mathbf{w}}_{t+h|t}(v) = \begin{bmatrix} \hat{w}_{t+h|t,1} \\ \hat{w}_{t+h|t,2} \\ \vdots \\ \hat{w}_{t+h|t,N} \end{bmatrix} = \min_w \sum_{i=0}^{v-1} \left( y_{t-i} - \sum_{k=1}^N w_{t-i|t-i-h,k} \cdot \hat{y}_{t-i|t-i-h,k} \right)^2 \quad (15)$$

A expressão acima tem solução exata (Johnson & Wichern, 2007):

$$\hat{\mathbf{w}}_{t+h|t} = (\hat{\mathbf{Y}}_{t|t-h} \hat{\mathbf{Y}}_{t|t-h}')^{-1} \hat{\mathbf{Y}}_{t|t-h}' y^{t|t-v} \quad (16)$$

onde

$$\hat{\mathbf{Y}}_{t|t-h} = \hat{\mathbf{Y}}_{t|t-h}(v) = \begin{bmatrix} \hat{y}_{t-1|t-h-1+1,1} & \hat{y}_{t-1|t-h-1+1,2} & \cdots & \hat{y}_{t-1|t-h-1+1,N} \\ \hat{y}_{t-2|t-h-2+1,1} & \hat{y}_{t-2|t-h-2+1,2} & \cdots & \hat{y}_{t-2|t-h-2+1,N} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{y}_{t-v+1|t-h-v+1,1} & \hat{y}_{t-v+1|t-h-v+1,2} & \cdots & \hat{y}_{t-v+1|t-h-v+1,N} \end{bmatrix} \quad (17)$$

$$y^{t|t-v} = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-v+1} \end{bmatrix} \quad (18)$$

As equações acima podem ser adaptadas para considerar janela expansiva. Neste caso, a equação (15) deve ser reescrita como mostra a equação (19).

$$\hat{\mathbf{w}}_{t+h|t} = \begin{bmatrix} \hat{w}_{t+h|t,1} \\ \hat{w}_{t+h|t,2} \\ \vdots \\ \hat{w}_{t+h|t,N} \end{bmatrix} = \min_w \sum_{i=h+1}^t \left( y_i - \sum_{k=1}^N w_{i|t-h,k} \cdot \hat{y}_{i|t-h,k} \right)^2 \quad (19)$$

### 2.4.3. Mínimos quadrados restritos

O uso de mínimos quadrados irrestritos garante, dentro da amostra, a melhor solução matemática para a estimação de pesos em combinações lineares (seção 2.4.2). Entretanto, os vetores de pesos gerados por aquele método, além de não terem uma interpretação tão direta, podem trazer problemas práticos fora da amostra, uma vez que não atendem a duas restrições desejáveis (Timmermann, 2006): (a) terem componentes somando 1 e (b) não terem componentes negativos.

A restrição (a) (somar 1) garante que a previsão por combinação linear seja não tendenciosa, desde que os previsores sendo combinados, individualmente, sejam não tendenciosos. Um predictor é não tendencioso quando o valor esperado ( $E$ ) de uma previsão equivale à própria realização da série:

$$E(\hat{y}_{t+h|t,k}) \equiv y_{t+h} \quad (20)$$

Para demonstrar o efeito da restrição (a), deve-se aplicar o valor esperado na equação (9):

$$E(y_{t+h|t}^C) = E\left(\sum_{k=1}^N \hat{w}_{t+h|t,k} \hat{y}_{t+h|t,k}\right) \quad (21)$$

$$E(y_{t+h|t}^C) = \hat{w}_{t+h|t,1} E(\hat{y}_{t+h|t,1}) + \hat{w}_{t+h|t,2} E(\hat{y}_{t+h|t,2}) + \dots + \hat{w}_{t+h|t,N} E(\hat{y}_{t+h|t,N}) \quad (22)$$

Se os previsores combinados forem não tendenciosos, a equação (22) pode ser convenientemente reduzida para a equação (23); pela equação, fica claro que a previsão combinada só será não tendenciosa se  $\hat{w}_{t+h|t,1} + \hat{w}_{t+h|t,2} + \dots + \hat{w}_{t+h|t,N} = 1$ .

$$E(y_{t+h}^C) = y_{t+h} (\hat{w}_{t+h|t,1} + \hat{w}_{t+h|t,2} + \dots + \hat{w}_{t+h|t,N}) \quad (23)$$

A restrição (b) (ser positivo), associada à condição de não tendenciosidade, torna a combinação linear convexa, fazendo com que a magnitude da previsão combinada esteja dentro da região de valores definida pelas previsões individuais. A Figura 7 ilustra esta questão. Além de fornecer melhor interpretação dos pesos, outra vantagem das combinações convexas é que elas tendem a ser mais **estáveis** (ou de menor variabilidade) do que as não convexas (seção 4.4.2.1).

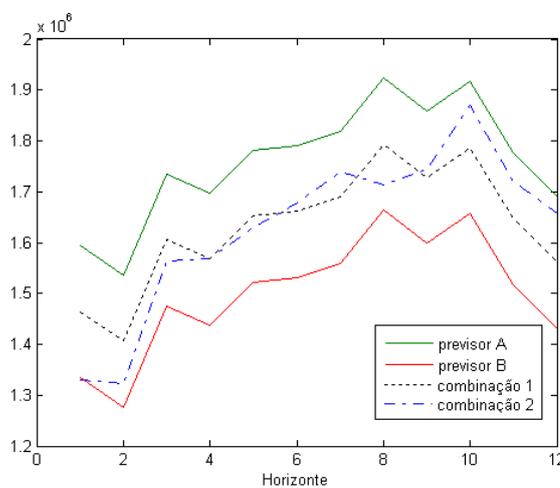


Figura 7 – Todas as combinações convexas entre os previsores A e B geram previsões limitadas à região por eles definida. Neste exemplo, mostram-se duas combinações possíveis.

Para transformar o problema de mínimos quadrados irrestritos em um problema de mínimos quadrados restritos, reescreve-se a equação (15) da seguinte forma:

$$\hat{\mathbf{w}}_{t+h|t}(v) = \begin{bmatrix} \hat{w}_{t+h|t,1} \\ \hat{w}_{t+h|t,2} \\ \vdots \\ \hat{w}_{t+h|t,N} \end{bmatrix} = \min_w \sum_{i=0}^{v-1} \left( y_{t-i} - \sum_{k=1}^N w_{t-i|t-i-h,k} \cdot \hat{y}_{t-i|t-i-h,k} \right)^2 \quad (24)$$

sujeito a

$$\sum_{k=1}^N w_{t|t-h,k} = 1 \quad e \quad w_{t|t-h,k} \geq 0 \quad (25)$$

Para este tipo de otimização não há solução exata; os algoritmos utilizados são iterativos (Gill et al. 1984; Mathworks, 2010a).

#### 2.4.4. Bates & Granger simples (BG1)

Outro método simples (mas também difícil de ser superado) é considerar que os pesos são inversamente proporcionais ao erro quadrático médio (MSE) (Bates & Granger, 1969; Newbold & Granger, 1974; Stock & Watson, 2001):

$$\hat{w}_{t+h|t,k}(v) = \frac{\left[ v^{-1} \sum_{i=0}^{v-1} (y_{t-i} - \hat{y}_{t-i|t-i-h,k})^2 \right]^{-1}}{\sum_{j=1}^N \left[ v^{-1} \sum_{i=0}^{v-1} (y_{t-i} - \hat{y}_{t-i|t-i-h,j})^2 \right]^{-1}} \quad (26)$$

sendo  $v \in 1,2,3\dots$  o tamanho da janela considerada e  $h$  o horizonte de previsão. O denominador na equação (26) garante que os pesos gerados em  $t$  somem 1 e, portanto, sejam convexos (seção 2.4.3), uma vez que um valor de MSE nunca é negativo. Assim como na seção 2.4.2, a equação (26) também pode ser escrita com janela expansiva:

$$\hat{w}_{t+h|t,k} = \frac{\left[ \frac{1}{t-h} \sum_{i=h+1}^t (y_i - \hat{y}_{i|t-h,k})^2 \right]^{-1}}{\sum_{j=1}^N \left[ \frac{1}{t-h} \sum_{i=h+1}^t (y_i - \hat{y}_{i|t-h,j})^2 \right]^{-1}} \quad (27)$$

#### 2.4.5. Bates & Granger correlacionado (BG2)

Este método estende o método anterior (seção 2.4.4), estabelecendo que os pesos devem ser inversamente proporcionais à soma das correlações de erro (seção 2.2.2) entre os previsores sendo combinados (Bates & Granger, 1969; Newbold & Granger, 1974). O conceito básico envolvido é dar maior peso aos previsores que têm pouca correlação (de erro) com os demais, e, ao mesmo tempo,

diminuir o peso dos previsores que são altamente correlacionados. Este conceito de ponderação também foi explorado por Mubwandarikwa (2007), como uma forma de privilegiar previsores menos **duplicados** e mais **independentes**. A equação (28) reflete esta questão.

$$\hat{\mathbf{w}}_{t+h|t}(\hat{\Sigma}) = \begin{bmatrix} \hat{w}_{t+h|t,1} \\ \hat{w}_{t+h|t,2} \\ \vdots \\ \hat{w}_{t+h|t,N} \end{bmatrix} = \frac{\hat{\Sigma}_{t|t-h}^{-1} \mathbf{1}}{\mathbf{1}' \hat{\Sigma}_{t|t-h}^{-1} \mathbf{1}} \quad (28)$$

onde:

- (i)  $\hat{\Sigma}_{t|t-h}^{-1}$  é o inverso da matriz  $N \times N$  de covariância dos erros de previsão, condicionada a um intervalo de tempo definido em instantes  $\leq t$ ;
- (ii)  $\mathbf{1}$  é um vetor  $N \times 1$  de **uns**.

Condicionando-se o cálculo da matriz de covariância  $\hat{\Sigma}_{t|t-h}$  a uma janela de tempo fixa de tamanho  $v$ , cada elemento  $[i, j]$  na matriz terá a seguinte forma:

$$\hat{\Sigma}_{t|t-h}[i, j] = \hat{\Sigma}_{t|t-h}(v)[i, j] = v^{-1} \sum_{u=0}^{v-1} \varepsilon_{t-u|t-u-h,i} \varepsilon_{t-u|t-u-h,j} \quad (29)$$

A variável  $\varepsilon_{t|t-h,k}$  é o erro de previsão em  $t$ , calculado com a previsão feita em  $t - h$  pelo predictor  $k$ :

$$\varepsilon_{t|t-h,k} = (y_t - \hat{y}_{t|t-h,k}) \quad (30)$$

Assim como no BG1, uma variante à utilização de janelas de tempo é a utilização de janela expansiva. Nesta variante, os elementos da matriz de covariância devem ser calculados de outra forma:

$$\hat{\Sigma}_{t|t-h}[i, j] = \frac{1}{t-h} \sum_{u=h+1}^t \varepsilon_{u|u-h,i} \varepsilon_{u|u-h,j} \quad (31)$$

Finalmente, é interessante observar que o denominador na equação (28) garante que os pesos gerados em  $t$  somem 1, o que assegura uma combinação linear não tendenciosa (seção 2.4.3). Por outro lado, o método não garante a condição de convexidade: pode haver pesos negativos.

#### 2.4.6. AFTER

No método AFTER – *Aggregated Forecast Through Exponential Re-weighting* (Yang, 2004), o peso atual é obtido recursivamente, aplicando-se um fator sobre o peso mais recente:

$$\hat{w}_{t+h|t,k}(\hat{\sigma}) = \frac{\hat{w}_{t-1+h|t-1,k} \hat{\sigma}_{t|t-h,k}^{-1/2} \exp\{-(y_t - \hat{y}_{t|t-h,k})^2 / 2\hat{\sigma}_{t|t-h,k}\}}{\sum_{j=1}^N \hat{w}_{t-1+h|t-1,j} \hat{\sigma}_{t|t-h,j}^{-1/2} \exp\{-(y_t - \hat{y}_{t|t-h,j})^2 / 2\hat{\sigma}_{t|t-h,j}\}} \quad (32)$$

O escalar  $\hat{\sigma}_{t|t-h,k}$  é um estimador da variância condicional dos erros do previsor  $k$ . Por exemplo, condicionando-se o estimador a uma janela fixa de tamanho  $v$  tem-se:

$$\hat{\sigma}_{t|t-h,k} = \hat{\sigma}_{t|t-h,k}(v) = v^{-1} \sum_{i=0}^{v-1} \varepsilon_{t-i|t-i-h,k}^2 \quad (33)$$

Os erros  $\varepsilon_{t|t-h,k}$  são calculados como na equação (30). Alternativamente, pode-se condicionar o cálculo a todo o histórico, i.e., usar janela expansiva:

$$\hat{\sigma}_{t|t-h,k} = \frac{1}{t-h} \sum_{u=h+1}^t \varepsilon_{u|u-h,k}^2 \quad (34)$$

Como observado por Sanchez (2008), a contribuição de um previsor no AFTER é pesada duas vezes (*re-weighted*): primeiro, pelo seu desempenho médio, através do termo  $\hat{\sigma}_{t|t-h,k}$ ; depois, pelo seu último desempenho, através do termo exponencial. O denominador na equação (32) garante que o vetor de pesos  $\hat{\mathbf{W}}_{t+h|t} = [\hat{w}_{t+h|t,1} \hat{w}_{t+h|t,2} \dots \hat{w}_{t+h|t,N}]'$  será convexo (seção 2.4.3).

Uma característica do AFTER é que se algum peso  $\hat{w}_{t+h|t,k}$  for **zerado** ele permanece assim por todo o horizonte subsequente; isto é particularmente indesejado em séries não estacionárias (Sanchez, 2008).

#### 2.4.7. Shrinkage

*Shrinkage* (Diebold & Pauly, 1990) não é um método de ponderação isolado, mas uma metodologia de integração de métodos. A ideia básica é que o vetor de pesos final seja uma média entre o vetor de pesos calculado pelo método de ponderação escolhido e um vetor de pesos definido *a priori* (Stock & Watson, 2006). Tipicamente o vetor imposto *a priori* é um vetor de pesos iguais, e neste caso diz-se que os pesos originais são **dilatados** na direção da combinação por média simples:

$$\hat{w}_{t+h|t,k}(\alpha) = \alpha \hat{w}_{t+h|t,k}^{\text{MÉTODO}} + (1-\alpha)(1/N) \quad (35)$$

O hiperparâmetro  $\alpha$  varia no intervalo [0,1] e controla a quantidade de dilatação na direção dos pesos iguais; no caso mais trivial,  $\alpha$  vale 0.5. Um pequeno aprofundamento na metodologia *shrinkage* conduz a um resultado algébrico simples, mas muitas vezes desaperecebido, válido para toda combinação linear: calcular um vetor médio entre dois (ou mais) vetores de pesos isolados e depois usar este vetor para combinar previsões (linearmente) é o mesmo que calcular a média entre duas (ou mais) previsões obtidas por combinações (lineares), definidas pelos mesmos vetores de pesos isolados usados anteriormente. Considere-se por exemplo três diferentes combinações lineares (por simplicidade, nas equações (36) a (42) fez-se  $\hat{w}_{t|t-h,k} = \hat{w}_{t,k}$  e  $\hat{y}_{t|t-h,k} = \hat{y}_{t,k}$ ):

$$y_t^{C(1)} = \hat{w}_{t,1}^{(1)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(1)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(1)} \hat{y}_{t,N} \quad (36)$$

$$y_t^{C(2)} = \hat{w}_{t,1}^{(2)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(2)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(2)} \hat{y}_{t,N} \quad (37)$$

$$y_t^{C(3)} = \hat{w}_{t,1}^{(3)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(3)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(3)} \hat{y}_{t,N} \quad (38)$$

Uma quarta combinação, definida pelo vetor de pesos médios, assumiria a forma da equação (39). Esta equação pode então ser reescrita – equações (40) e (41) – para mostrar o que se pretende.

$$y_t^{C(4)} = \frac{(\hat{w}_{t,1}^{(1)} + \hat{w}_{t,1}^{(2)} + \hat{w}_{t,1}^{(3)})}{3} \hat{y}_{t,1} + \frac{(\hat{w}_{t,2}^{(1)} + \hat{w}_{t,2}^{(2)} + \hat{w}_{t,2}^{(3)})}{3} \hat{y}_{t,2} + \dots + \frac{(\hat{w}_{t,N}^{(1)} + \hat{w}_{t,N}^{(2)} + \hat{w}_{t,N}^{(3)})}{3} \hat{y}_{t,N} \quad (39)$$

$$3y_t^{C(4)} = \hat{w}_{t,1}^{(1)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(1)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(1)} \hat{y}_{t,N} + \hat{w}_{t,1}^{(2)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(2)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(2)} \hat{y}_{t,N} + \hat{w}_{t,1}^{(3)} \cdot \hat{y}_{t,1} + \hat{w}_{t,2}^{(3)} \cdot \hat{y}_{t,2} + \dots + \hat{w}_{t,N}^{(3)} \hat{y}_{t,N} \quad (40)$$

$$y_t^{C(4)} = \frac{y_t^{C(1)} + y_t^{C(2)} + y_t^{C(3)}}{3} \quad (41)$$

Outro resultado importante é que se cada um dos vetores de pesos originais somar 1, o vetor de pesos médio também somará 1. Isto pode ser observado a partir da equação (39), impondo-se a seguinte igualdade:

$$\frac{(\hat{w}_{t,1}^{(1)} + \hat{w}_{t,1}^{(2)} + \hat{w}_{t,1}^{(3)})}{3} + \frac{(\hat{w}_{t,2}^{(1)} + \hat{w}_{t,2}^{(2)} + \hat{w}_{t,2}^{(3)})}{3} + \dots + \frac{(\hat{w}_{t,N}^{(1)} + \hat{w}_{t,N}^{(2)} + \hat{w}_{t,N}^{(3)})}{3} = 1 \quad (42)$$

Admitindo-se como verdadeira a hipótese de que a soma dos componentes de cada um dos vetores de pesos originais é 1, i.e.,  $\hat{w}_{t,1}^{(i)} + \hat{w}_{t,2}^{(i)} + \dots + \hat{w}_{t,N}^{(i)} = 1$ ,  $i = 1, 2, 3$ , é fácil concluir que também é verdadeira a igualdade na equação (42).