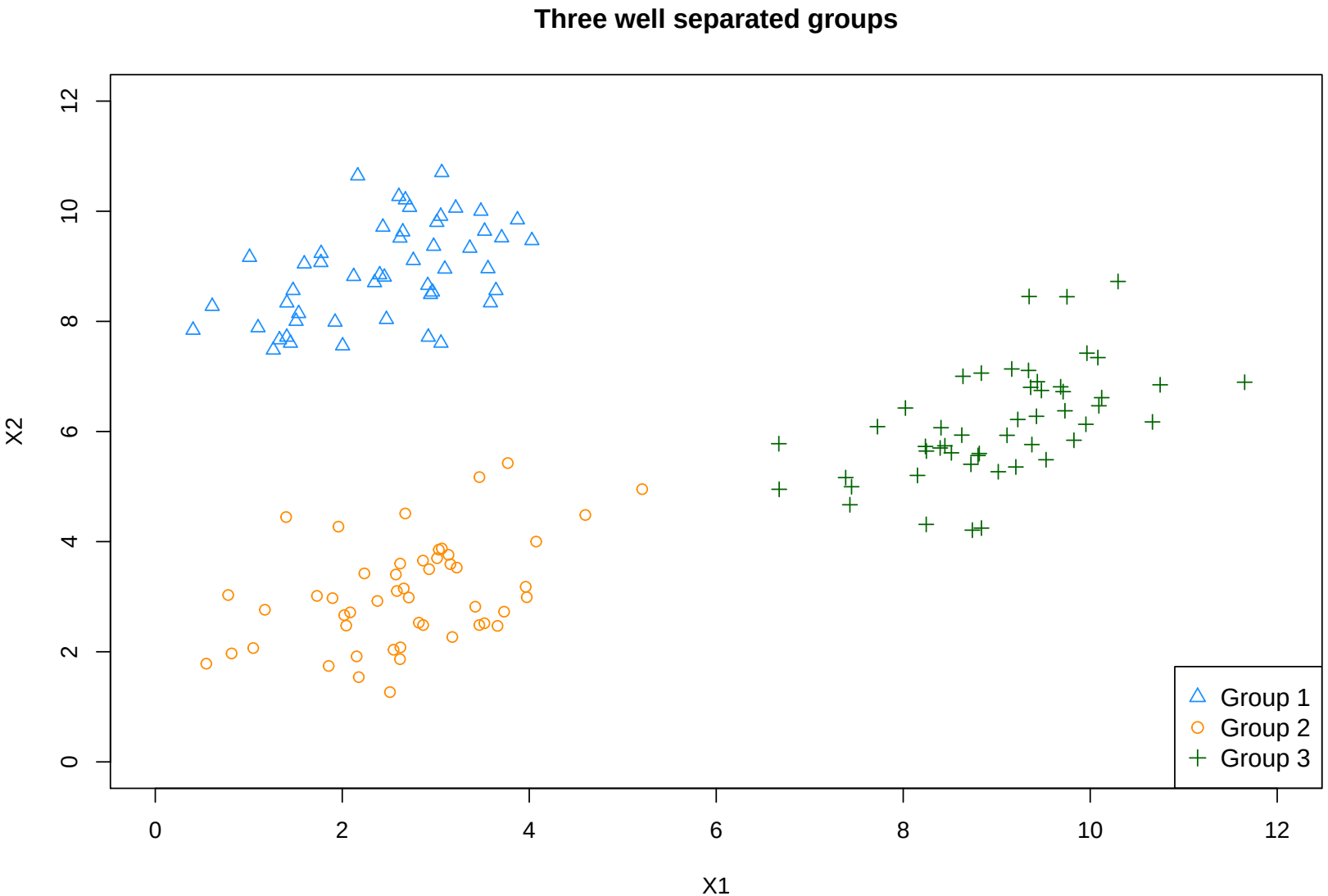
Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y Juan Carlos Salazar

Aprendizaje supervisado versus no supervisado

El siguiente gráfico ilustra 150 observaciones (simuladas) con mediciones en dos variables, X_1 y X_2. Cada observación corresponde a uno de tres grupos diferentes. Por ilustración, las observaciones se han coloreado para distinguirlas. Sin embargo, en la práctica, la membresía a un grupo particular no es conocida y el objetivo es determinar a que grupo pertenece cada observación. Si los grupos están claramente separados la tarea no es demasiado difícil.

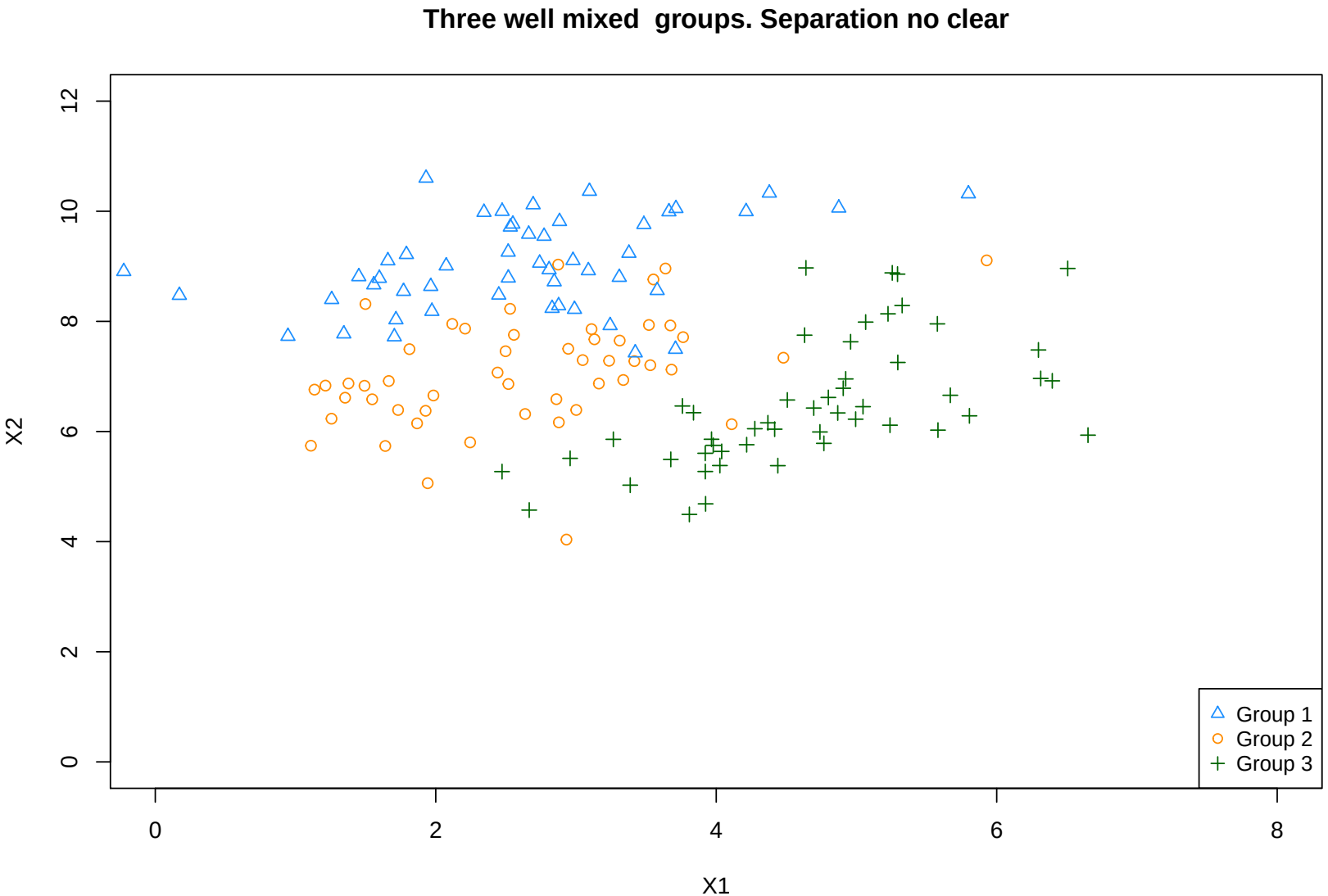
Aprendizaje supervisado versus no supervisado



Aprendizaje supervisado versus no supervisado

Si los grupos no están claramente separados la tarea es más difícil ya que puede haber mucha superposición. En una situación como esta, se espera que un método de clustering no asigne los puntos de la intersección a su grupo de manera correcta. Una situación como está se ilustra en el siguiente gráfico

Aprendizaje supervisado versus no supervisado



Aprendizaje supervisado versus no supervisado

En todo caso, si hay solo dos variables, se podría inspeccionar visualmente las nubes de puntos a fin de identificar los clusters. Sin embargo, en la práctica frecuentemente se encuentran conjuntos de datos con más de dos variables y gráficar los datos ya no es tarea fácil. Por ejemplo, si hay p variables en el conjunto de datos, entonces hay $\begin{pmatrix} p \\ 2 \end{pmatrix} = \frac{p(p-1)}{2}$ nubes de puntos y por ende la inspección visual no es una forma práctica y viable de identificar clusters. **Por esta razón, los métodos de clustering automáticos son importantes**.

Aprendizaje supervisado versus no supervisado

Regresión versus problemas de clasificación. Las variables se pueden caracterizar como cuantitativas o cualitativas (categóricas). Las cuantitativas toman un valor numérico y las categorícas un valor discreto ordinal o nominal. Ejemplos de cuantitativas incluyen la edad, altura, ingreso, valor de un inmueble, precio de una acción. Ejemplos de cualitativas incluyen el género, la marca de un producto, cumplir con una obligación crediticia, tipo de cáncer, entre otras.

Aprendizaje supervisado versus no supervisado

Usualmente, los problemas con respuesta cuantitativa se conocen como **Problemas de regresión**, mientras que aquellos con respuesta cualitativa se conocen como **Problemas de clasificación**. Algunos métodos tales como el de K vecinos más cercanos, Naive Bayes y boosting (se discuten más adelante) se pueden usar ya sea con variables cuantitativas o cualitativas.

Evaluando la precisión de un modelo. ¿Porqué es necesario introducir tantos métodos de AE en vez de solo un buen método? La respuesta es que ningún método domina a los otros en presencia de todos los posibles conjuntos de datos. Cada conjunto de datos, tiene sus características propias que hacen que un método que funcione bien para un conjunto similar puede no hacerlo para él. **Por lo tanto es una tarea importante decidir, para un conjunto de datos particular, cuál método produce los mejores resultados. Seleccionar la mejor aproximación, es en general uno de los elementos más retadores al implementar AE en la práctica**.

Midiendo la calidad del ajuste.

- Se necesita cuantificar el grado al cual la respuesta predicha, dada una observación, está cerca de la respuesta verdadera para esa observación.
- En regresión, la medida de uso más común es el **Error Cuadrático Medio** (MSE: Mean Squared Error, en inglés):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2$$

Donde $\widehat{f}(x_i)$ es la predicción que $\widehat{f}$ genera para la i-ésima observación y_i. El MSE será pequeño si la predicción está cerca de la respuesta verdadera, y será grande, si para alguna de las observaciones, el valor predicho y verdadero difieren sustancialmente.

El MSE se calcula usando los datos de entrenamiento y por lo tanto se conoce como **MSE de entrenamiento** o **Training MSE**.

$$Training - MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2$$

Pero en general, el interés no se centra en si el método funciona bien en los datos de entrenamiento, sino en la precisión de la predicción que se obtiene cuando el método se aplica a datos nuevos (que no han sido observados aún, Test Data, o datos de prueba). El MSE con los datos de prueba se llama **MSE de prueba** o **Test MSE**.

Formalmente, suponga que se ajusta un método de AE al conjunto de entrenamiento $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ y se obtiene $\widehat{f}$. Se puede entonces calcular $\widehat{f}(x_1), \ldots, \widehat{f}(x_n)$. Si cada uno de estos valores es cercano a $y_1, \ldots, y_n$ entonces el training MSE es pequeño. Pero no se esta interesado en si $\widehat{f}(x_i) \approx y_i$; en vez de esto, interesa saber si $\widehat{f}(x_0)$ es aproximadamente igual a y_0, donde (x_0, y_0) es una **observación no registrada previamente y que no fue usada para entrenar el método de AE**. **Se quiere seleccionar el método que produzca el menor test MSE, y no el que produzca el menor training MSE**.

Modelos paramétricos versus no paramétricos

En otras palabras, si se tiene un número grande de observaciones de prueba, se podría calcular la cantidad

$$Average\left(y_0 - \widehat{f}(x_0)\right)^2$$

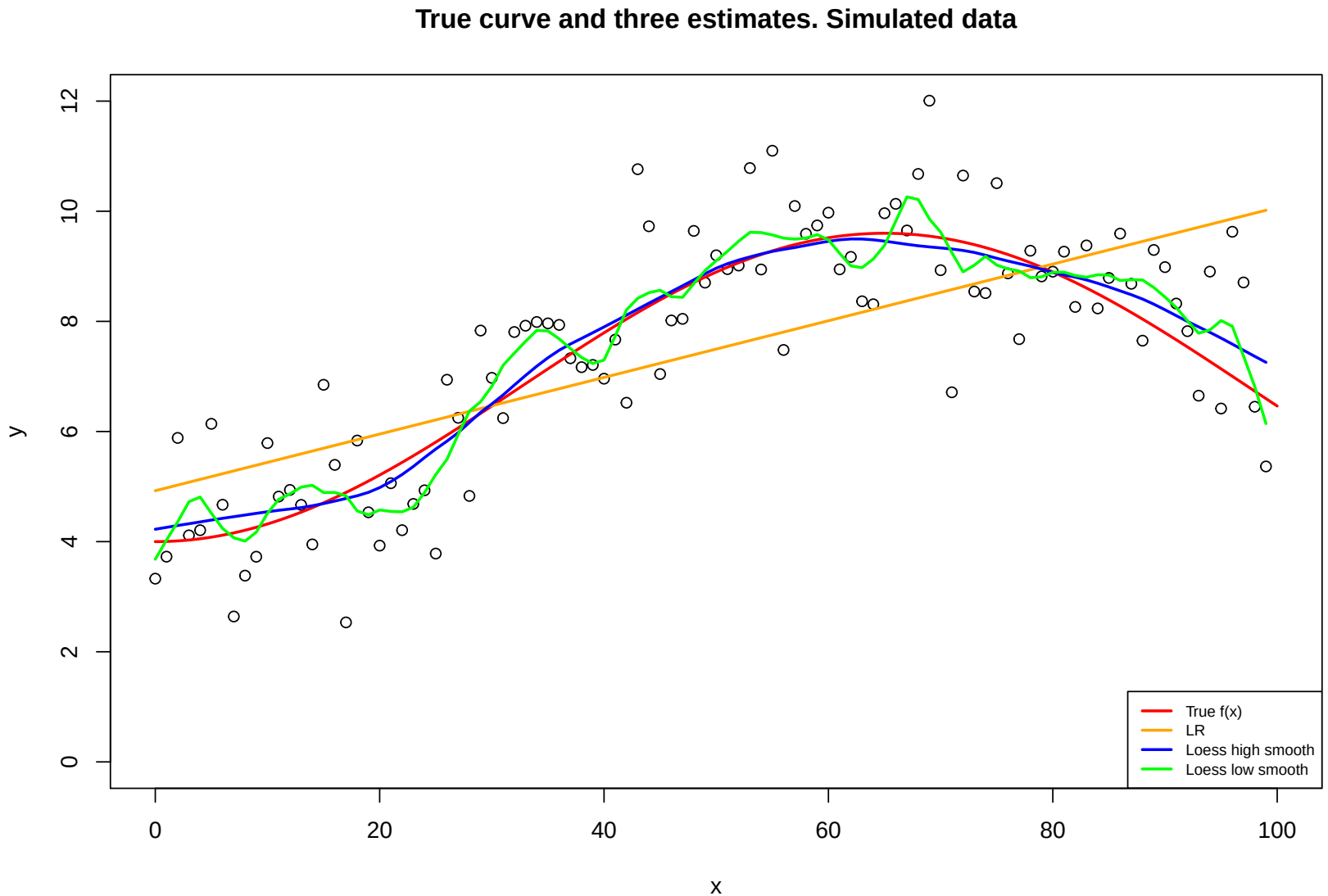
Es decir, si se tienen N observaciones de prueba $\{(x_{01}, y_{01}), \ldots, (x_{0N}, y_{0N})\}$

$$Test - MSE = \frac{1}{N}\sum_{j=1}^{N}\left(y_{0j} - \widehat{f}(x_{0j})\right)^2$$

Se quiere seleccionar el modelo o método para el cual el Test MSE sea el más pequeño posible.

- **Problema: No hay garantía de que un método con un training MSE pequeño también tenga un test MSE pequeño**.

Modelos paramétricos versus no paramétricos



Modelos paramétricos versus no paramétricos

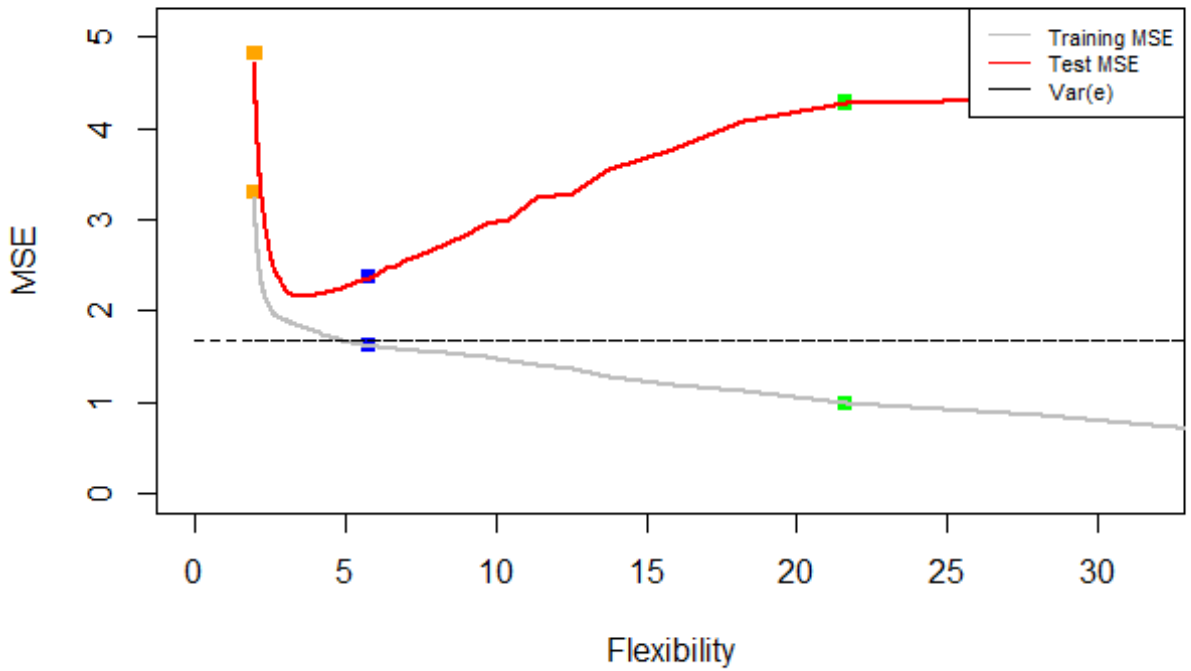


Figura 1: Flexibilidad versus MSE

Observe que el training MSE disminuye a medida que se incrementa la flexibilidad (grados de libertad asociados al modelo) del modelo, pero esto no es cierto con el test MSE. El problema es que muchos métodos estadísticos estiman coeficientes que minimizan el training MSE, pero no hay garantía de que el test MSE también siga siendo pequeño tal y como se ilustra en el gráfico anterior.

La recta de regresión lineal simple es el modelo menos flexible (2 grados de libertad). El training MSE decrece monótonamente a medida que se incrementa la flexibilidad (grados de libertad de cada modelo). En este ejemplo, la verdadera f es no lineal, por lo que esta recta de regresión hace un trabajo de estimación muy pobre y además no es lo suficientemente flexible para capturar las fluctuaciones de f. Por su parte la curva verde, tiene el menor training MSE ya que es la más flexible de las tres (flexibilidad arriba de 20 grados de libertad). la línea negra es el error irreducible que es el menor test MSE que se puede alcanzar.

La curva azul minimiza el test MSE (la curva roja), lo cual no sorprende, ya que del gráfico se observa que de los tres modelos es la que mejor estima a f (su nivel de flexibilidad es un poco mayor a 5). Esto se confirma observando la curva del test MSE la cual decrece a medida que el nivel de flexibilidad aumenta, luego se estabiliza y después empieza a crecer. Esto hace que los modelos con mayor test MSE sean el modelo naranja (regresión lineal) y el verde (el suavizamiento con muchos grados de libertad). De aquí se concluye que el mejor modelo es el de la curva azul (spline suavizado con un poco más de 5 grados de libertad)

NOTAS:

- **El patrón descrito en forma de U por la curva del test MSE es una propiedad fundamental de AE que siempre se cumple sin importar el conjunto de datos ni el método estadístico usado..**
- **A medida que la flexibilidad aumenta, el training MSE decrece, pero el test MSE no necesariamente. Cuando un método produce un training MSE pequeño pero un test MSE grande, se dice que el modelo sobreajusta los datos**[1].

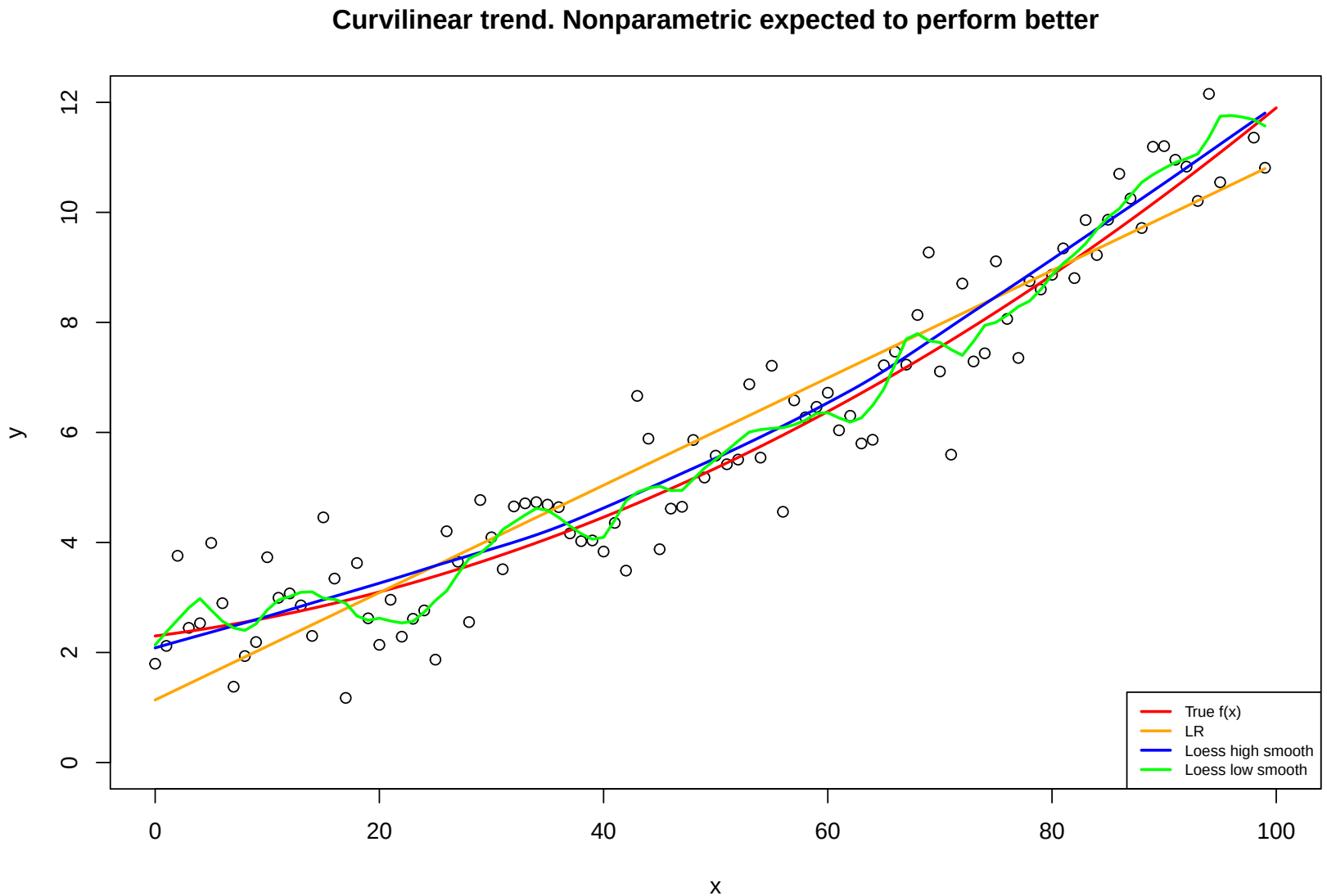
[1] el método de AE se esfuerza en encontrar la señal, pero termina explicando la fluctuaciones causadas por el ruido y no por la señal

Modelos paramétricos versus no paramétricos

Cuando se sobreajustan los datos[2], el test MSE es muy grande ya que el supuesto patrón que encontró el método con los datos de entrenamiento simplemente no existen en el conjunto de prueba. Casi siempre, se espera que el training MSE sea menor que el test MSE ya que la mayoría de métodos de AE buscan minimizar el training MSE.

[2] el método de AE se esfuerza en encontrar la señal, pero termina explicando la fluctuaciones causadas por el ruido y no por la señal

Otro ejemplo.



Modelos paramétricos versus no paramétricos

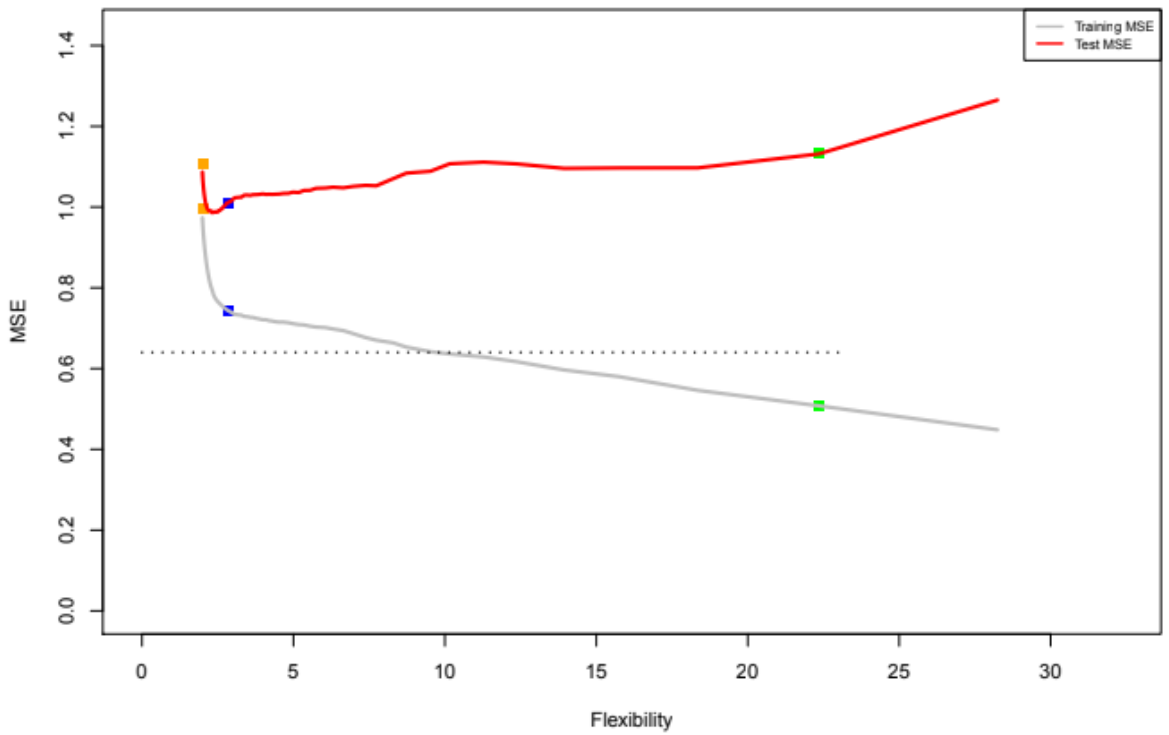
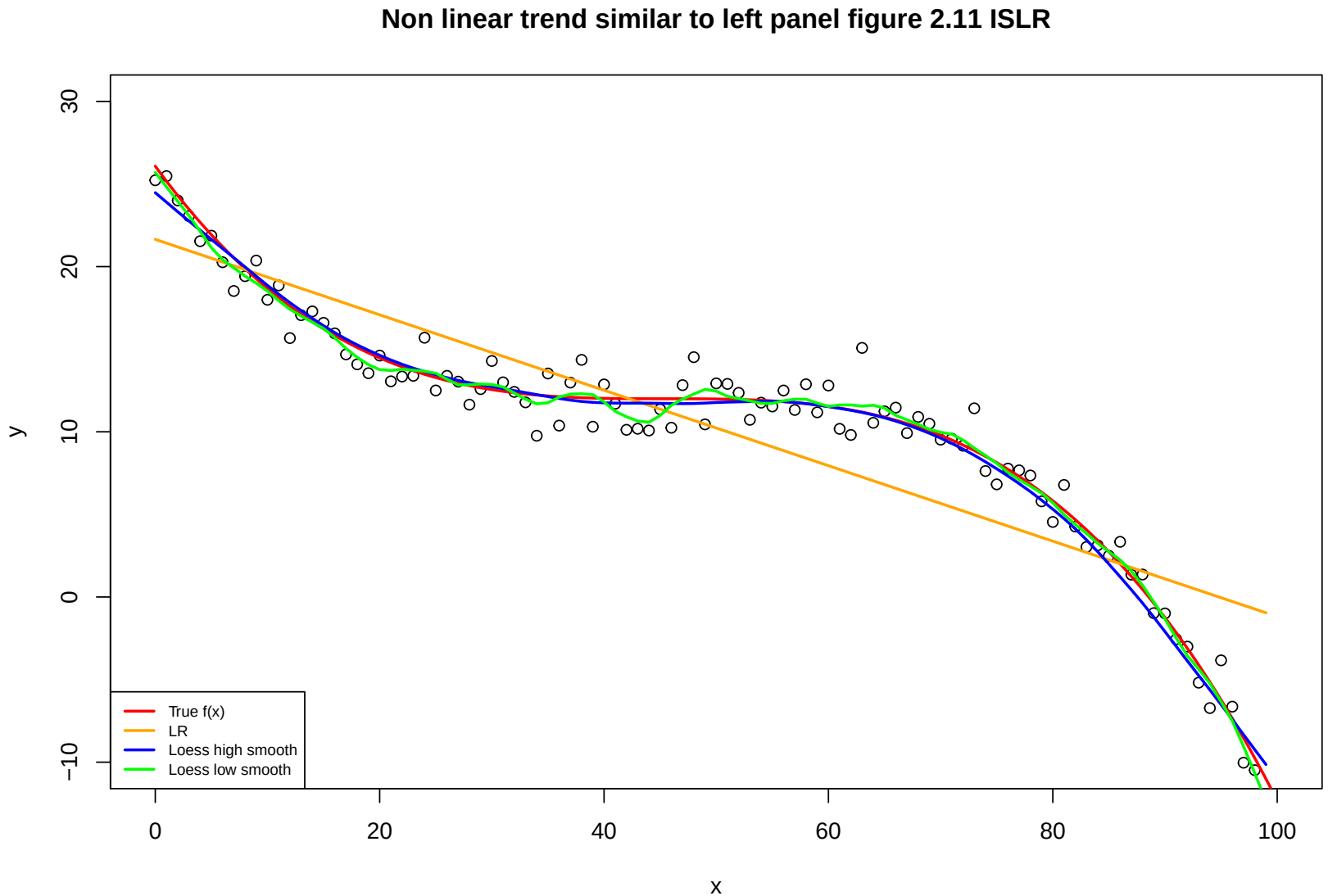


Figura 2: Flexibilidad versus MSE

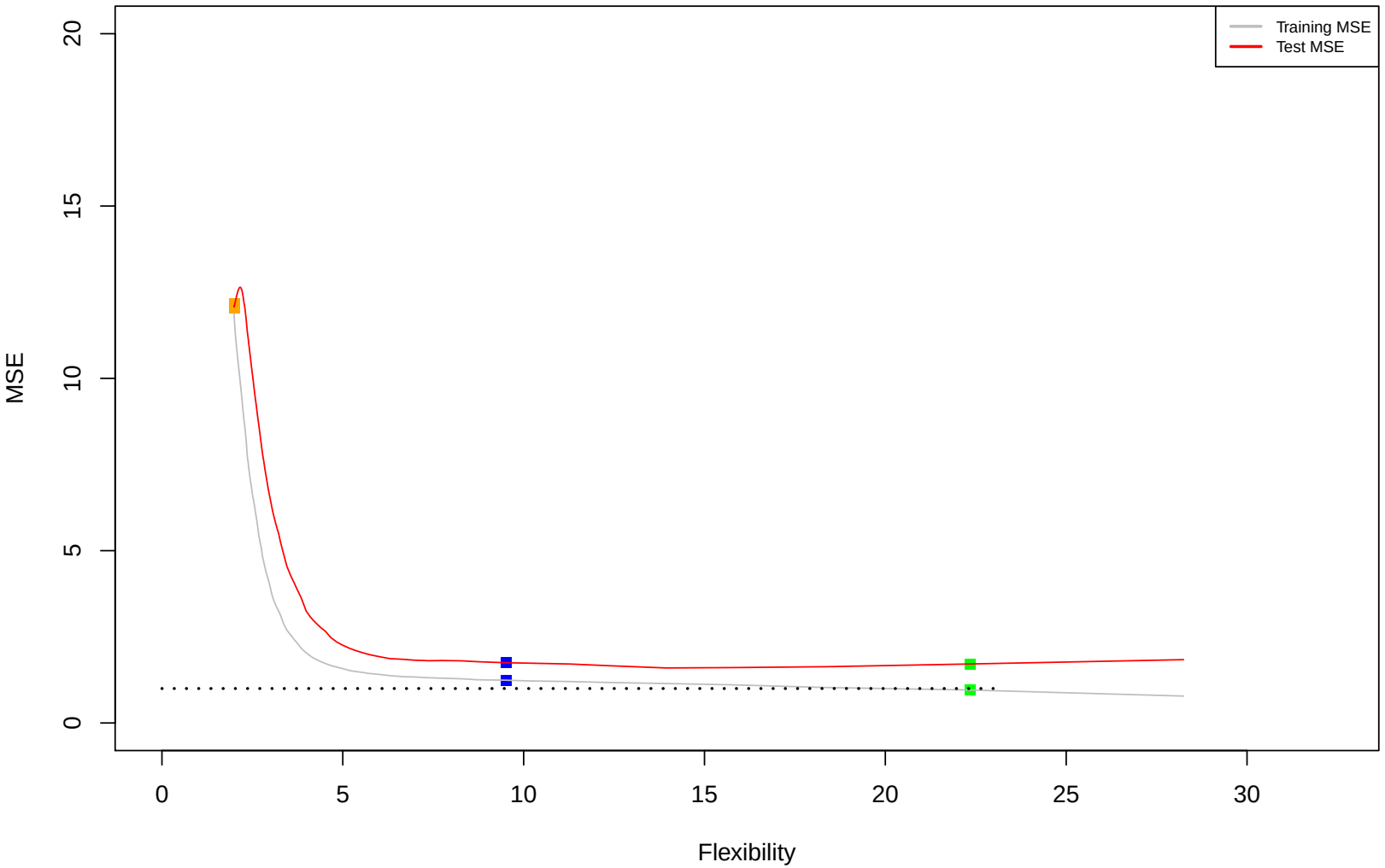
En este ejemplo se observa de nuevo que el training MSE decrece a medida que la flexibilidad aumenta, y también la forma de U del test MSE. Sin embargo, debido a que la verdadera f está cercana a una función lineal, el modelo lineal (naranja) tiene un mejor desempeño que el altamente no lineal modelo verde. En este caso, de nuevo el modelo azul tiene un menor test MSE entre los tres considerados y por lo tanto su ajuste se considera mejor que el de los otros dos.

Modelos paramétricos versus no paramétricos

Finalmente, considere una situación, donde f es no lineal:

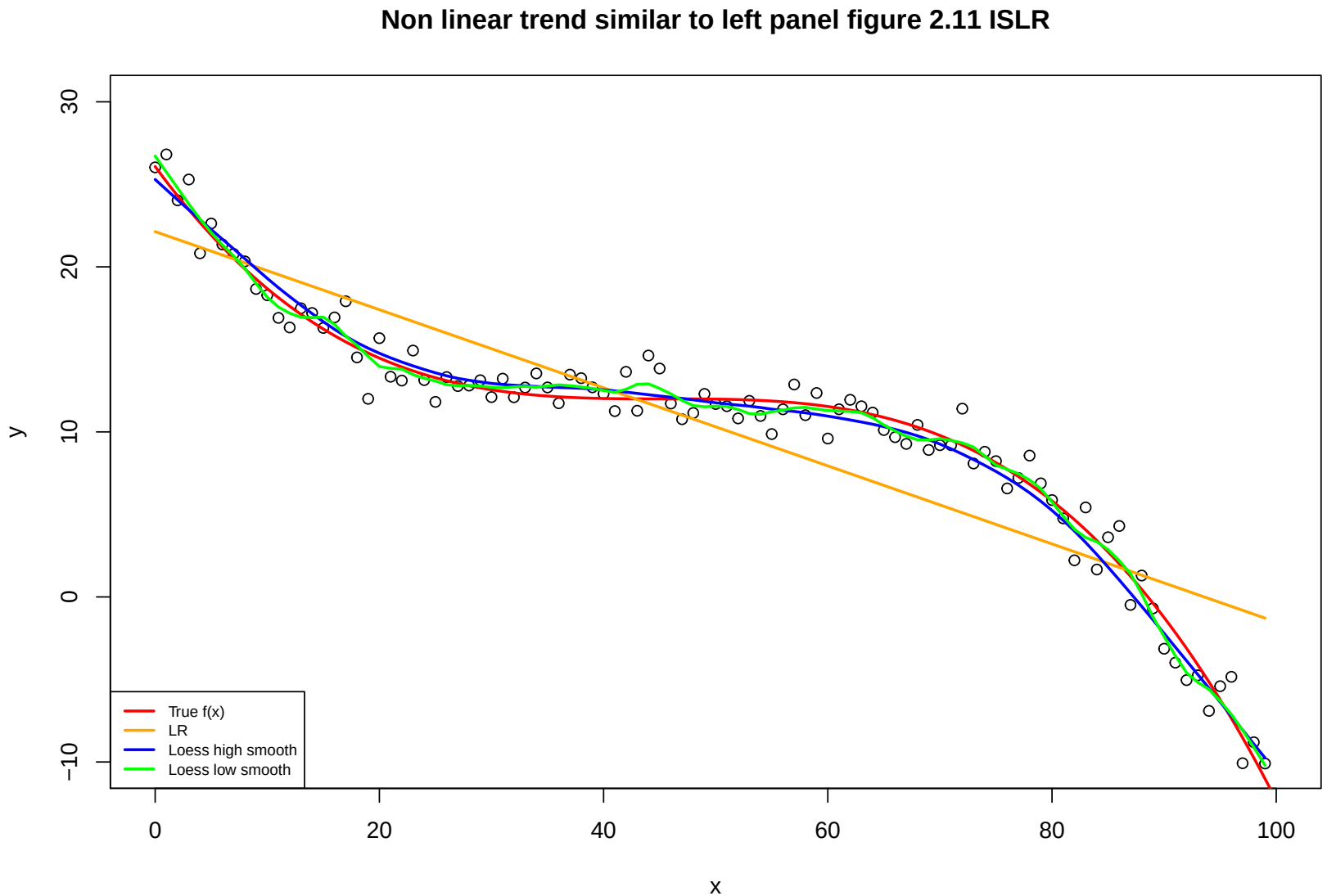


Modelos paramétricos versus no paramétricos

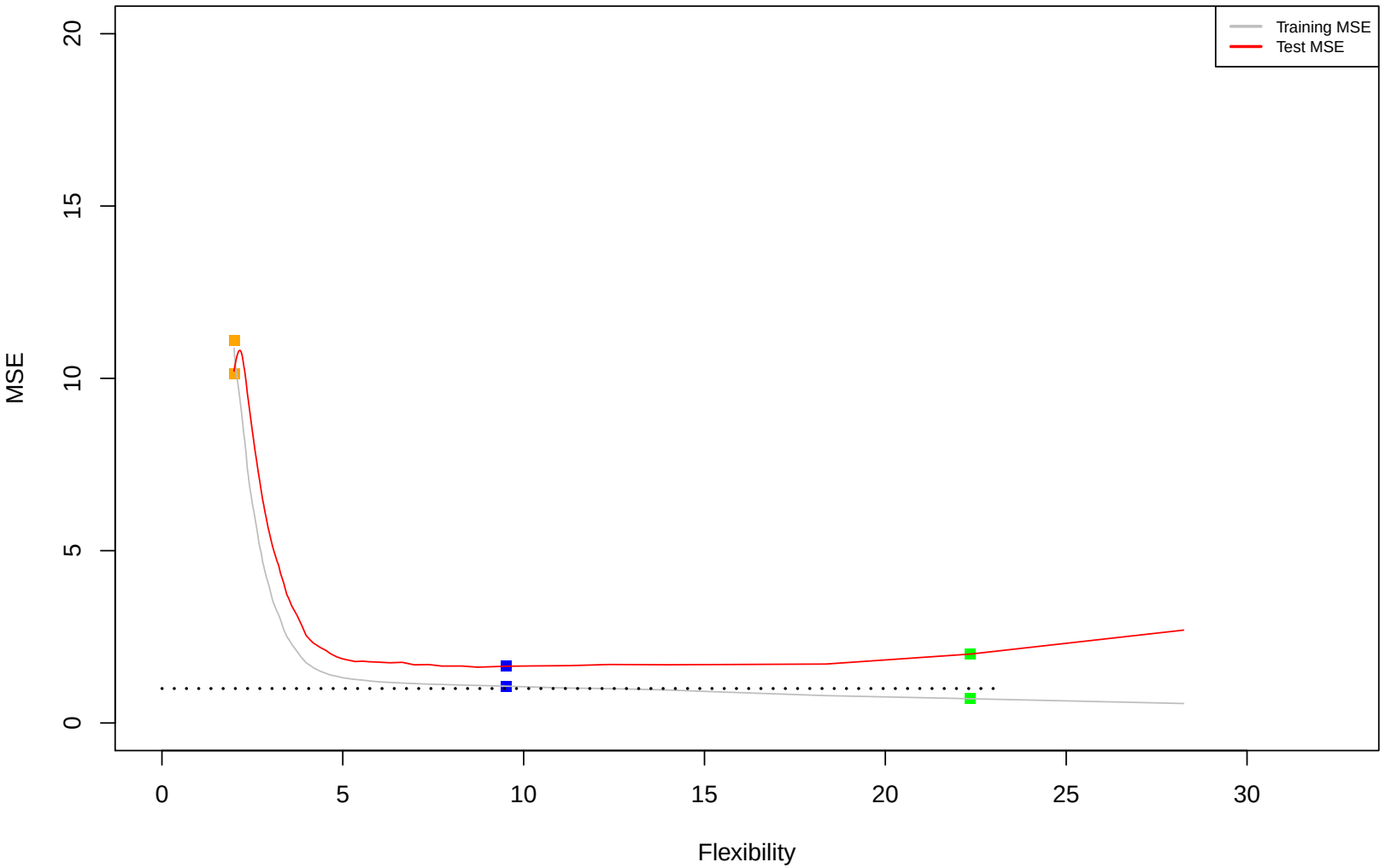


LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

Considere la función $y = -0.00021999(0.88x - 40)^3 + 12 + e$:



LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF



LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

La forma característica en U del test MSE resulta ser el resultado de dos propiedades de AE que compiten entre si: El sesgo (Bias) y la varianza. Es posible mostrar (Ejercicio) que el Test MSE, para un valor dado x_0, siempre se puede descomponer en la suma de tres cantidades fundamentales: La varianza de $\widehat{f}(x_0)$, el sesgo de $\widehat{f}(x_0)$ al cuadrado y la varianza del error. Es decir,

$$E\left(y_0 - \widehat{f}(x_0)\right)^2 = Var\left(\widehat{f}(x_0)\right) + \left[Bias\left(\widehat{f}(x_0)\right)\right]^2 + Var\left(\varepsilon\right)$$

En este contexto,

$$Bias(\widehat{f}(x)) = E\left(\widehat{f}(x)\right) - f(x)$$

y

$$Var\left(\widehat{f}(x)\right) = E\left(\widehat{f}(x) - E\left(\widehat{f}(x)\right)\right)^2$$

La notación $E\left(y_0 - \widehat{f}(x_0)\right)^2$ define el Test MSE esperado y se refiere al test MSE promedio que se podría obtener si se estima repetidamente f usando un gran número de conjuntos de entrenamiento y probando cada uno con x_0. El test MSE promedio global se puede calcular al promediar $E\left(y_0 - \widehat{f}(x_0)\right)^2$ sobre todos los posibles valores de x_0 en el conjunto de prueba.

EJEMPLO: CHEQUEANDO LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

Usando los tres modelos anteriores (LR, loess high smooth and low smooth):

```
## [1] "CHECKING THE BIAS-VARIANCE EQUATION FOR EACH MODEL:"

## [1] "sqrBias + variance + var(e), FOR EACH MODEL:"

## [1] 13.698405  1.537036  1.688157

## [1] "MSE FOR EACH MODEL:"

## [1] 13.586712  1.528946  1.681228

## [1] "Approximation is considered a good one!!!!"
```

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

La ecuación bias-variance tradeoff lo que dice es que a fin de minimizar el test MSE esperado, se requiere seleccionar un método de AE que simultáneamente alcance baja varianza y bajo sesgo. Note que la varianza y el cuadrado del sesgo son cantidades no negativas, por lo tanto, el test MSE esperado nunca puede caer abajo de $Var(\varepsilon)$.

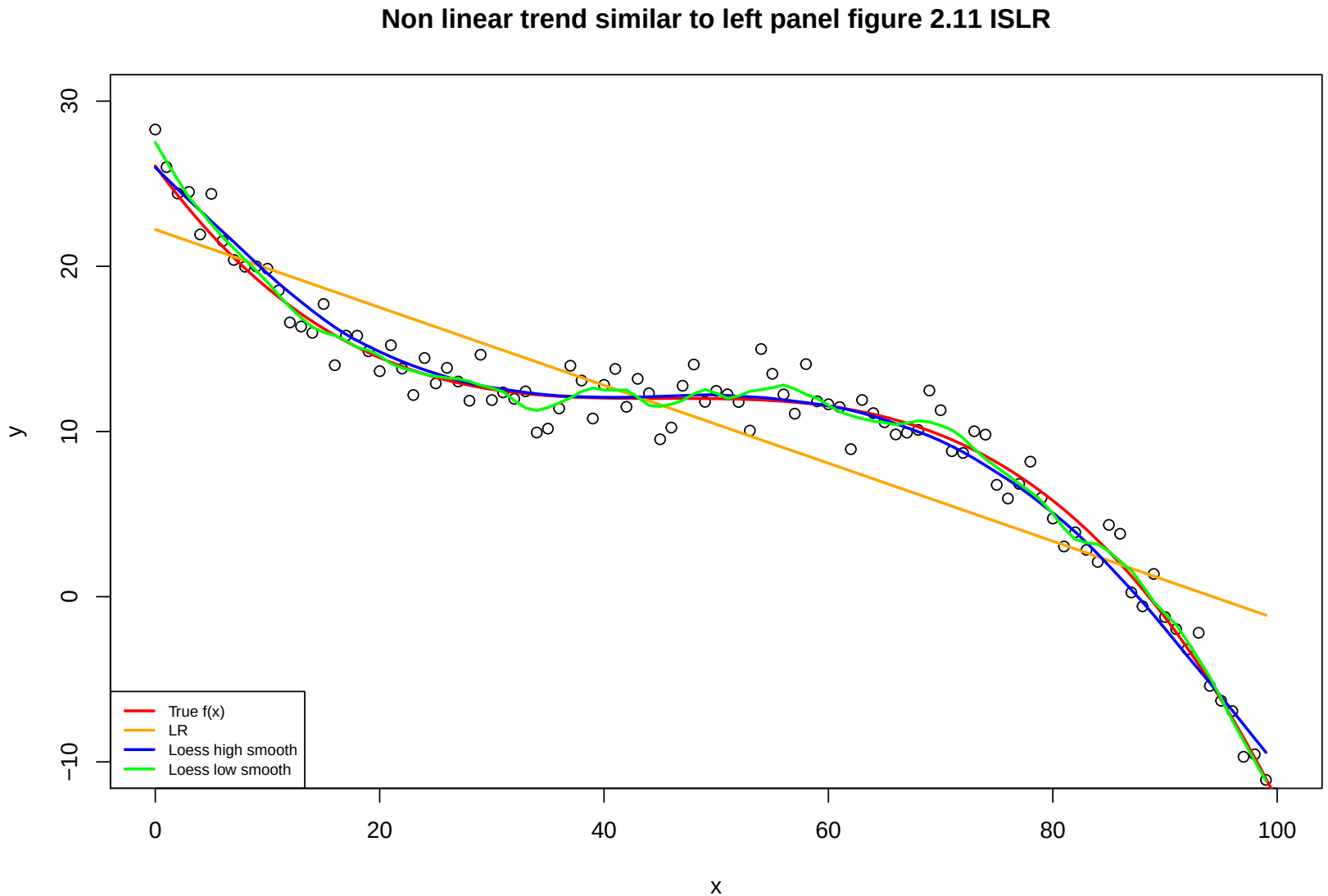
LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

¿Qué significa la varianza y el sesgo de un método de AE? La varianza se refiere a la cantidad que $\widehat{f}$ cambiaría si se estima usando un conjunto de entrenamiento diferente. Puesto que el conjunto de entrenamiento se usa para ajustar el método de AE, diferentes conjuntos de entrenamiento producirán distintos $\widehat{f}$. Sin embargo, si un método tiene varianza grande entonces, pequeños cambios en el conjunto de entrenamiento resultarán en grandes cambios en $\widehat{f}$. En general, métodos de AE más flexibles, tienen mayor varianza.

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

Sesgo (o Bias) se refiere al error que se introduce al tratar de aproximar un problema de la vida real, el cual puede ser muy complicado, usando un modelo simple. Considere de nuevo la siguiente situación:

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

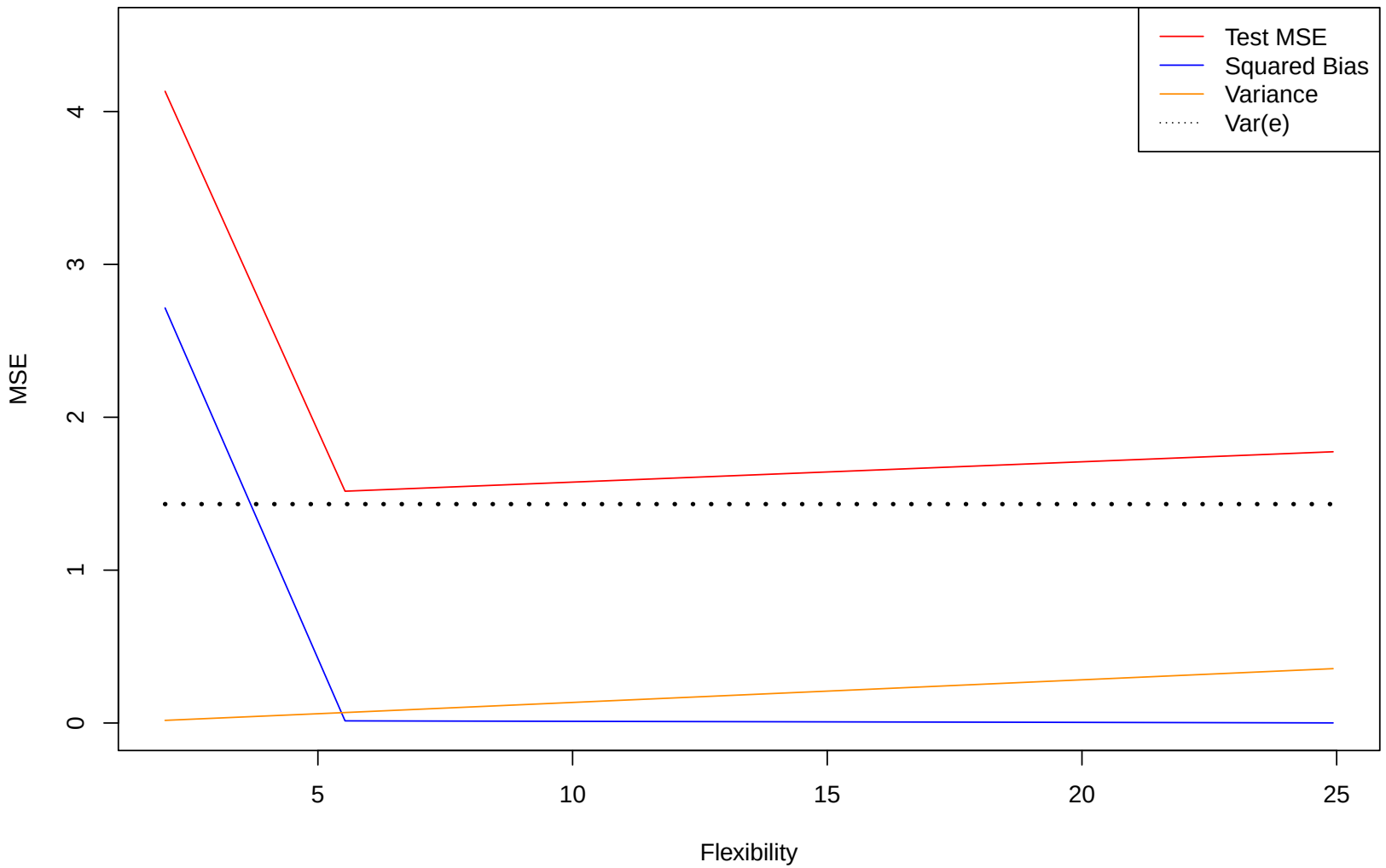


En el gráfico anterior, la verdadera f es altamente nolineal, por lo tanto, no importa cuántas observaciones de entrenamiento se tengan, no será posible obtener una buena estimación usando regresión lineal (modelo naranja, regresión lineal tiene un sesgo grande en este caso). Pero el suavizamiento, tiene un sesgo menor En general, métodos de AE más flexibles tienen menos sesgo.

Como una regla general, si se usan métodos de AE muy flexibles, la varianza se incrementará y el sesgo decrecerá. La tasa relativa de cambio entre estas dos cantidades determina si el test MSE decrece o crece. A medida que se incrementa la flexibilidad de una clase de métodos, el sesgo tiende inicialmente a decrecer más rápido que el incremento en la varianza. Sin embargo, en algún punto, incrementar la flexibilidad tiene poco impacto en el sesgo, pero empieza a incrementar de manera importante la varianza. Cuando esto pasa, el Test MSE se incrementa.

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

Gráfico del Bias-Variance tradeoff para el modelo $y = -2.8 \times cos((2\pi/130)x) + 6.8$:



LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

El gráfico anterior ilustra una situación donde la varianza crece mientras que el sesgo al cuadrado disminuye y la típica curva en U del test MSE. Esta situación se conoce como el Bias-Variance Tradeoff. Un buen método de AE debe tener baja varianza y bajo sesgo al cuadrado y este es el reto en AE.

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

El Bias-Variance tradeoff siempre debe tenerse en mente, aún cuando en situaciones de la vida real, en las cuales no se observa f, no es sencillo o posible calcular explicitamente el test MSE, el sesgo y la varianza para un método de AE. Sin embargo, con un método conocido como Cross-Validation (Validación Cruzada, se discute más adelante) es posible calcular el test MSE usando los datos de entrenamiento.

El problema de clasificación. Hasta ahora, la situación se ha centrado en el problema de regresión. Conceptos tales como Bias-Variance Tradeoff (Intercambio entre sesgo y varianza) se pueden tambien adaptar al problema de clasificación teniendo en cuenta que y_i ya no es numérica. Suponga que se quiere estimar a f con base en una muestra de entrenamiento $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ donde las $y_1, y_2, \ldots, y_n$ son cualitativas o categóricas.

La forma más común de identificar la precisión en la estimación $\widehat{f}$ es usando **La tasa de error de entrenamiento (Training error rate)**, que es la proporción de errores que se cometen si se aplica el estimador $\widehat{f}$ a las observaciones de entrenamiento:

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \widehat{y}_i\right)$$

Fracción de clasificaciones incorrectas. Se calcula con los datos de entrenamiento.

Aquí, $\widehat{y}_i$ es la clase predicha (o el sello o label de la clase predicha) para la i-ésima observación usando $\widehat{f}$ y $I(y_i \neq \widehat{y}_i)$ es una variable indicadora que vale 1 si $y_i \neq \widehat{y}_i$ y cero en caso contrario. Si vale cero, la observación se clasificó correctamente. **El interés se centra en tasas de error que resultan al aplicar el clasificador $\widehat{f}$ a observaciones de prueba que no se usaron como datos de entrenamiento**.

LA ECUACIÓN DEL BIAS-VARIANCE TRADEOFF

La tasa de error de prueba (Test error rate) asociada con un conjunto de n_0 observaciones de prueba de la forma $(x_0,\ y_0)$ está dada por:

$$Average\left(I\left(y_0 \neq \widehat{y}_0\right)\right) = \frac{1}{n_0}\sum_{j=1}^{n_0} I\left(y_{0j} \neq \widehat{y}_{0j}\right)$$

Donde $\widehat{y}_0$ es la clase predicha que resulta al aplicar el clasificador $\widehat{f}$ a la observación con predictor x_0. **Un buen clasificador será aquel para el cual** $Average\left(I\left(y_0 \neq \widehat{y}_0\right)\right)$ **es el más pequeño**.