METR 5303 – Lecture #5

Global function fitting

We shall briefly discuss global function fitting, and then examine some of the pitfalls of function fitting in general. Recall that Lecture 3 gave an example of local (polynomial) fitting, where the fit of a cluster of data was accomplished solely to find the value at a grid point in the center of the cluster.

Global function fitting is the process of fitting specified functions to *all* observations in an analysis domain.

Again, let $f(\mathbf{r})$ be the dependent variable, where $\mathbf{r}$ is a 1-D, 2-D or 3-D spatial coordinate.

Assume the analyzed field $f_A(\mathbf{r})$ can be represented by a finite series of ordered basis functions $h_1(\mathbf{r})$, $h_2(\mathbf{r})$, ……. $h_m(\mathbf{r})$ …….. $h_M(\mathbf{r})$, $m = 0, \ldots, M$. Then our function model can be written as

$$f_A(\mathbf{r}) = \sum_{m=0}^{M} c_m \, h_m(\mathbf{r}) \qquad (1)$$

Note that m could be summing over 1, 2 or 3D space.

Assume K observations $f_o(\mathbf{r}_k)$, $k = 1, \ldots, K$ over the entire domain.

To obtain a global fit in the least squares sense, we need to minimize

$$I = \sum_{k=1}^{K} w_k \, d_k^2 = \sum_{k=1}^{K} w_k \left[\sum_{m=0}^{M} c_m \, h_m(\mathbf{r}_k) - f_O(\mathbf{r}_k) \right]^2 \qquad (2)$$

where $w_k = ½ < \varepsilon_o^2(\mathbf{r}_k) >^{-1}$ are the observational error variances.

To minimize (2), differentiate w.r.t. each coefficient c_m and set to zero:

$$\sum_{k=1}^{K} w_k h_m(\mathbf{r}_k) \left[\sum_{\mu=0}^{M} c_\mu h_\mu(\mathbf{r}_k) - f_O(\mathbf{r}_k) \right] = 0$$

or

$$\sum_{\mu=0}^{M} c_\mu \left[\sum_{k=1}^{K} w_k h_m(\mathbf{r}_k) h_\mu(\mathbf{r}_k) \right] = \sum_{k=1}^{K} w_k h_m(\mathbf{r}_k) f_O(\mathbf{r}_k) \qquad (3)$$

Eq. (3) are the <u>normal equations</u> for this problem, from which we need to solve for the c_μ

If we repeat this process in matrix form, eq. (1) becomes

$$\mathbf{f}_A = \mathbf{H}\,\mathbf{c} \qquad (4)$$

where $\mathbf{f}_A$ is a column vector of analysis values $f_A(\mathbf{r}_k)$ of length K and $\mathbf{H}$ is a K x (M+1) rectangular matrix with elements $h_{km} \sim h_m(\mathbf{r}_k)$.

The derivation of the normal equations starts with our usual quadratic form for I:

$$I \;=\; \tfrac{1}{2}\,\{\,[\mathbf{f}_A \;-\; \mathbf{f}_o]^T\,\mathbf{O}^{-1}\,[\mathbf{f}_A \;-\; \mathbf{f}_o]\,\}$$

where **O** is a K x K diagonal matrix whose elements are $< \varepsilon_o^2(\mathbf{r}_k) >$.

Using (4), we have

$$I \;=\; \tfrac{1}{2}\,\{\,[\mathbf{H}\,\mathbf{c} \;-\; \mathbf{f}_o]^T\,\mathbf{O}^{-1}\,[\mathbf{H}\,\mathbf{c} \;-\; \mathbf{f}_o]\,\}$$

which is equivalent to eq. (2). Taking the derivative w.r.t. **c** and setting to zero yields

$$\mathbf{H}^T\,\mathbf{O}^{-1}\,[\mathbf{H}\,\mathbf{c} \;-\; \mathbf{f}_o] \;=\; 0$$

or

$$\mathbf{H}^T\,\mathbf{O}^{-1}\,\mathbf{H}\,\mathbf{c} \;=\; \mathbf{H}^T\,\mathbf{O}^{-1}\,\mathbf{f}_o \qquad (5)$$

The $\mathbf{H}^T\,\mathbf{O}^{-1}\,\mathbf{H}$ term is a (M+1) x (M+1) square matrix called the Gram matrix **G**.

Thus eq. (5) is

$$\mathbf{G}\,\mathbf{c} \;=\; \mathbf{H}^T\,\mathbf{O}^{-1}\,\mathbf{f}_o$$

In principle, G is invertible, so we can solve this equation for the unknown coefficients c

$$\mathbf{c} \;=\; \mathbf{G}^{-1}\,\mathbf{H}^T\,\mathbf{O}^{-1}\,\mathbf{f}_o \qquad (6)$$

We now use these values of **c** in eq. (4) [or eq. (1)], from which we can evaluate $f_A(\mathbf{r})$ anywhere in the domain.

If there are K observations, we could compute up to K coefficients from eq. (6), if $M+1 \leq K$

Note: If $M+1 = K$, **H** is a square matrix, and the normal equations (3) or (5) are said to be fully determined. In general, though, we prefer over-determined systems where $M+1 < K$. However, for some interpolation problems with limited data availability (or perfect data), we may prefer to use a fully determined approach.

Example of a fully determined system

Assume polynomial basis functions:

$$h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = x^2, \ldots\ldots\ldots, h_M(x) = x^M \qquad (7)$$

In this case we have K observations, x_k , $k = 1, \ldots, K$, and the **H** matrix can be written

$$\mathbf{H} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_K & x_K^2 & & x_K^M \end{bmatrix} \qquad (8)$$

K x (M + 1)

When $K = M + 1$, **H** is square, we can just solve $\mathbf{H}\,\mathbf{c} = \mathbf{f_o}$ via $\mathbf{c} = \mathbf{H}^{-1}\,\mathbf{f_o}$. (9)

So, for the case where $K = M + 1 = 2$ (2 obs. with 1 and x as the basis functions), use of (9) and (8) gives us

$$\begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}^{-1} \begin{bmatrix} f_0(x_1) \\ f_o(x_2) \end{bmatrix}. \qquad (10)$$

Class exercise: Solve eq. (10) to obtain:

$$c_0 = \frac{x_2 f_o(x_1) - x_1 f_o(x_2)}{x_2 - x_1}\ , \qquad c_1 = \frac{f_o(x_2) - f_o(x_1)}{x_2 - x_1}\ .$$

These coefficients can be used in the matrix form of the analysis problem, eq. (4), to write

$$\begin{bmatrix} f_A(x_1) \\ f_A(x_2) \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} \begin{bmatrix} \dfrac{x_2 f_o(x_1) - x_1 f_o(x_2)}{x_2 - x_1} \\ \dfrac{f_o(x_2) - f_o(x_1)}{x_2 - x_1} \end{bmatrix}. \qquad (11)$$

Finally, we can write the result in continuous form using our original equation (1), with M = 1, to obtain

$$f_A(x) = \frac{x_2 - x}{x_2 - x_1} f_o(x_1) + \frac{x - x_1}{x_2 - x_1} f_o(x_2), \qquad (12)$$

which is the Lagrange interpolating polynomial for K = 2. (see Appendix E in Daley)

Verify via (12) that f_A (x) fits the data points f_o (x_1) and f_o (x_2) at x_1 and x_2 exactly.

Therefore, a fully determined least squares minimization using polynomial basis functions is equivalent to Lagrange interpolation (for $K = 2$).

Note: Use of $M+1 = K$ is appropriate only if the data values are error free (e.g., as in the analysis step where one must interpolate from the grid to the station location – known as the forward model.)

When using observational data that contain error, we should use the over-determined approach where $K >> M + 1$.

To evaluate **c** in (6) the Gram matrix **G** has to be inverted. It can be inverted if all eigenvalues of G are positive. However, if one or more of the eigenvalues is vanishingly small, it can be numerically singular. This can happen when two of the basis functions, h_m (**r**) and h_μ (**r**) are similar but not identical, then two rows of the Gram matrix will become similar, and the matrix can be numerically singular.

See Daley, p. 43-45 for further discussion of more properties of the Gram matrix, especially w.r.t. determination of the condition number N of the matrix,

$$\mathrm{N} = \frac{\lambda_{\max}}{\lambda_{\min}}$$

where λ is an eigenvalue of **G**. $\lambda_{\max}$ is the largest eignevalue and $\lambda_{\min}$ is the smallest eigen value of **G**. The error of the matrix inversion will increase with the conditional number, and when $\lambda_{\min}$ is close to zero, the inversion may fail. When the condition number is high, the matrix is ill conditioned.

Unfortunately, the matrices encountered in function fitting often have high condition numbers, and for the polynomial functions, the condition number is high except when M is small. $M = 7$ is about the practical limit for polynomial fitting. This is not a problem for local fitting where M is generally small, but it is a serious limitation in global fitting.

Even when **G** is well conditioned, its inversion is still expensive.

Problems with function fitting

Even if **G** is invertible and a solution obtained, potential problems exist with function fitting (which have counterparts in nearly all analysis methods). Examples are:

A. Overfitting

Consider Fig. 2.3 (taken from book Thiebaux and Pedder (1987)). The true signal is $1 + x$. However, the obs. are generated from $f_o(x_j) = 1 + x_j + \varepsilon_j$, $j = 1, \ldots, 7$

where ε_j are random, Gaussian errors. Now we attempt to fit these data using polynomial basis functions:

$$f_A(x) = \sum_{k=0}^{M} c_k x^k \quad , \quad M = 6.$$

Thus 7 observations allow us to fit up to a 6th-order polynomial to the data.

Fig. 2.3 below shows fits for $M = 1, 2, 3, 4, 5$, and 6 (k in the figure), where $M = 6$ represents the exact fit (the others represent over-determined cases).

Discuss figure. One might argue that the fitted curves from $M = 2$, 3 and 4 are just as good as when $M = 1$ but the derivative of f_A, $\frac{\partial f_A(x)}{\partial x}$, starts behaving poorly by $M = 2$ or 3.

When $M = 6$, we see we have an exact fit to the data, which would be great if the obs were perfect, but if the obs contain error, this "perfect" fit comes at great cost to the values of the field between the observations, and to the derivatives.

Message: Don't fit the data containing error too closely or you will be fitting the noise or sampling errors as well.

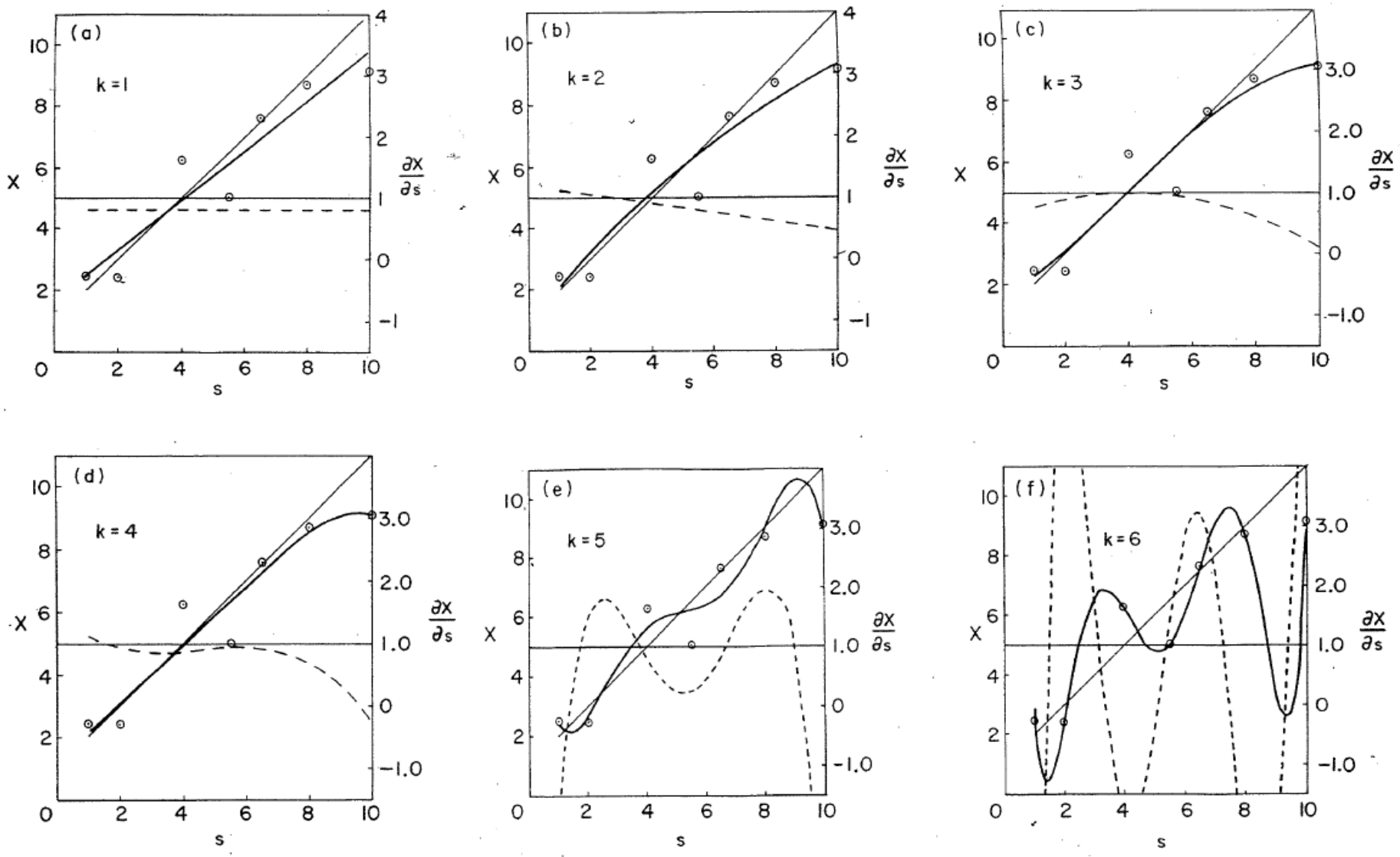

Fig. 2.3 The effect of fitting progressively higher-order (k) polynomial models by OLS regression to observations scattered randomly about the signal function $1 + s$. The bold full curves represent the analysis $X^a(s)$. The broken curves represent $\partial X^a(s)/\partial s$.

B. Underfitting

Examine Fig. 2.4 of Thiebaux and Pedder. Here the true signal is given by $1 + x^2$ (quadratic) but it is estimated with the linear polynomial $f_A = c_0 + c_1 x$.

The two linear misfits have different slopes depending on the distribution of the observations.

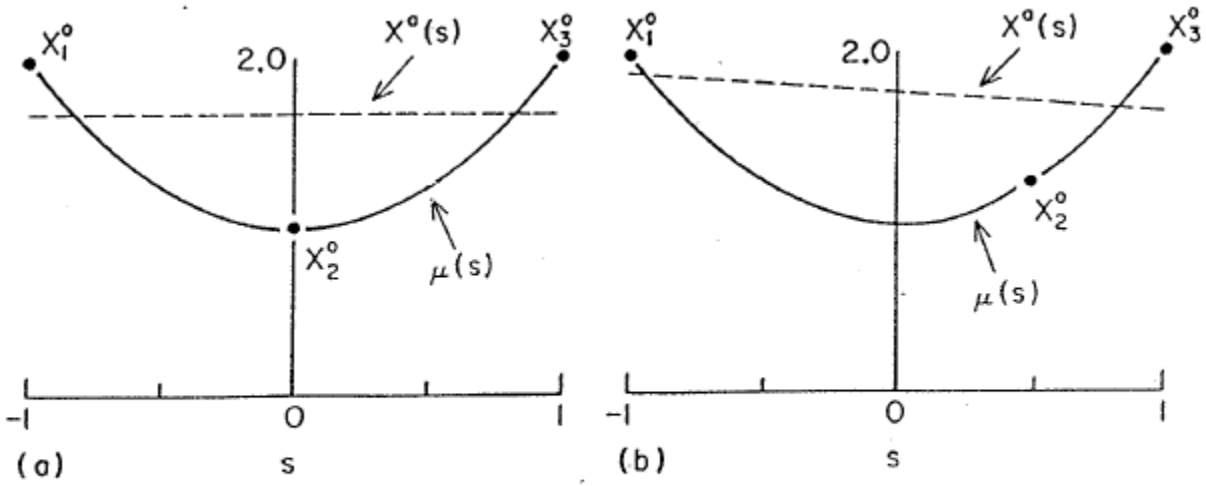

Fig. 2.4 The effect of aliasing from a quadratic component of variation in $\mu(s)$ on an OLS straight line analysis $X^a(s)$ for two arrangements of three observing locations.

C. Mis-specification of the function basis

This is illustrated by Fig. 2.6 of Thiebaux and Pedder. The truth is given by a sine wave. The estimate is made with polynomial basis functions. One is a cubic fit (k = 3; overdetermined since have 5 data points). The other (k = 4; quartic model) is fully determined - so it provides an exact fit to the observations but is a poorer estimate of the truth between the observations. Therefore, success in fitting the data exactly (or very closely) is no guarantee of success in knowing the field between observations.

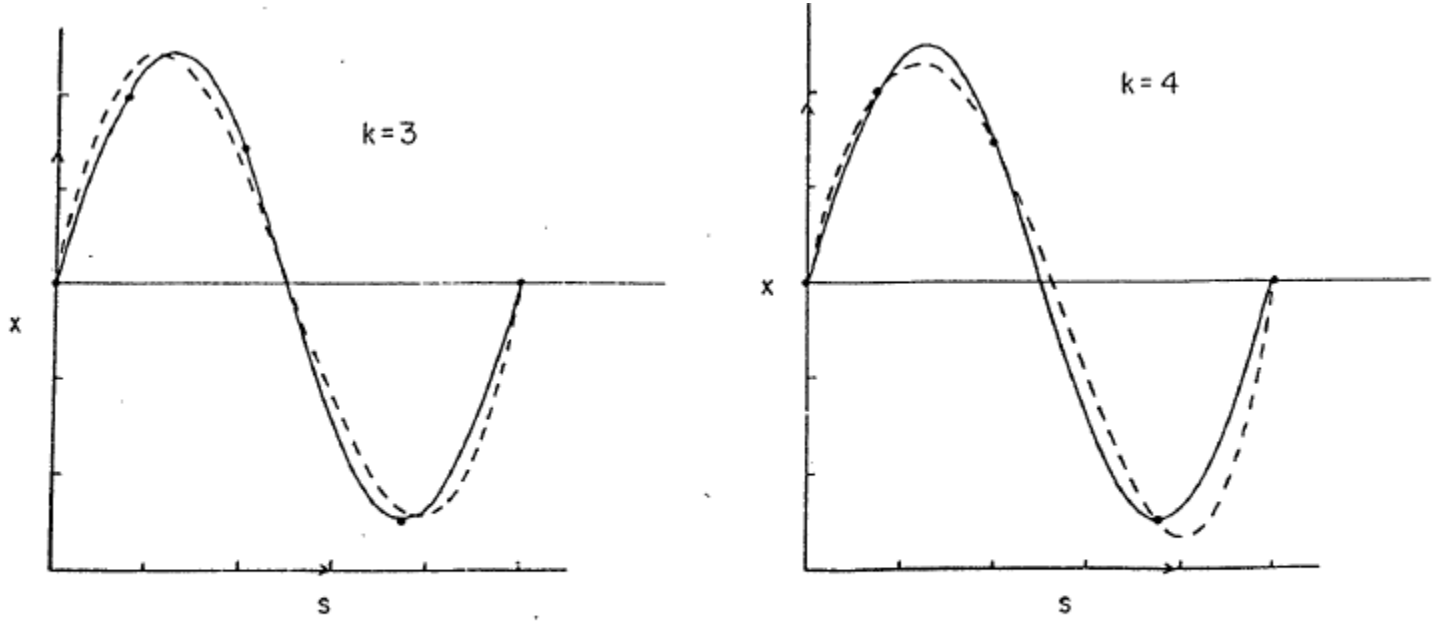

Fig. 2.6 The effect of regressing polynomial models (broken curves) onto observations lying on a sine curve (full curves).

D. Example in Daley, p. 45-49: Underfitting and overfitting

Assume the analysis $f_A(x)$ is expanded in trigonometric basis function

$$f_A(x) = \frac{a_0}{2} + \sum_{m=1}^{M} a_m \cos(mx) + b_m \sin(mx)$$

Or in a complex form

$$f_A(x) = \sum_{m=-M}^{M} c_m \exp(imx)$$

Where $c_m = 0.5(a_m - ib_m)$, and $c_{-m} = c_m^*$, where (*) indicates complex conjugation.

The truth is from trigonometric basis functions with $M = 2$. 9 observations were also taken from this model with random, Gaussian errors added. The truth ($M = 2$) implies 5 degrees of freedom: $-2 \leq m \leq 2$.

If $M = 0$ or 1: Causes underfitting (Fig. 2.2b in Daley) - analysis is too smooth

If $M = 4$: Represents a fully determined system (since we have $2M + 1$ weights and 9 observations). Fig. 2.2c clearly shows that this is overfitting.

In this case, a choice of $M = 2$ provides the best fit. Thus an over-determined fit is preferred.

Also note that the fit is better where the observations are dense (left side) then where the obs are sparse (right side).

Note on derivatives of trigonometric vs polynomial basis functions

Derivatives of trigonometric expansions have more variance (weight) at the smaller scales - as would be expected from experience [recall changes as we differentiate the height field once (to get velocity) or twice (to get vorticity)].

However, derivatives of polynomials produce smoother and smoother results (ultimately going to zero) – which is counter to experience.

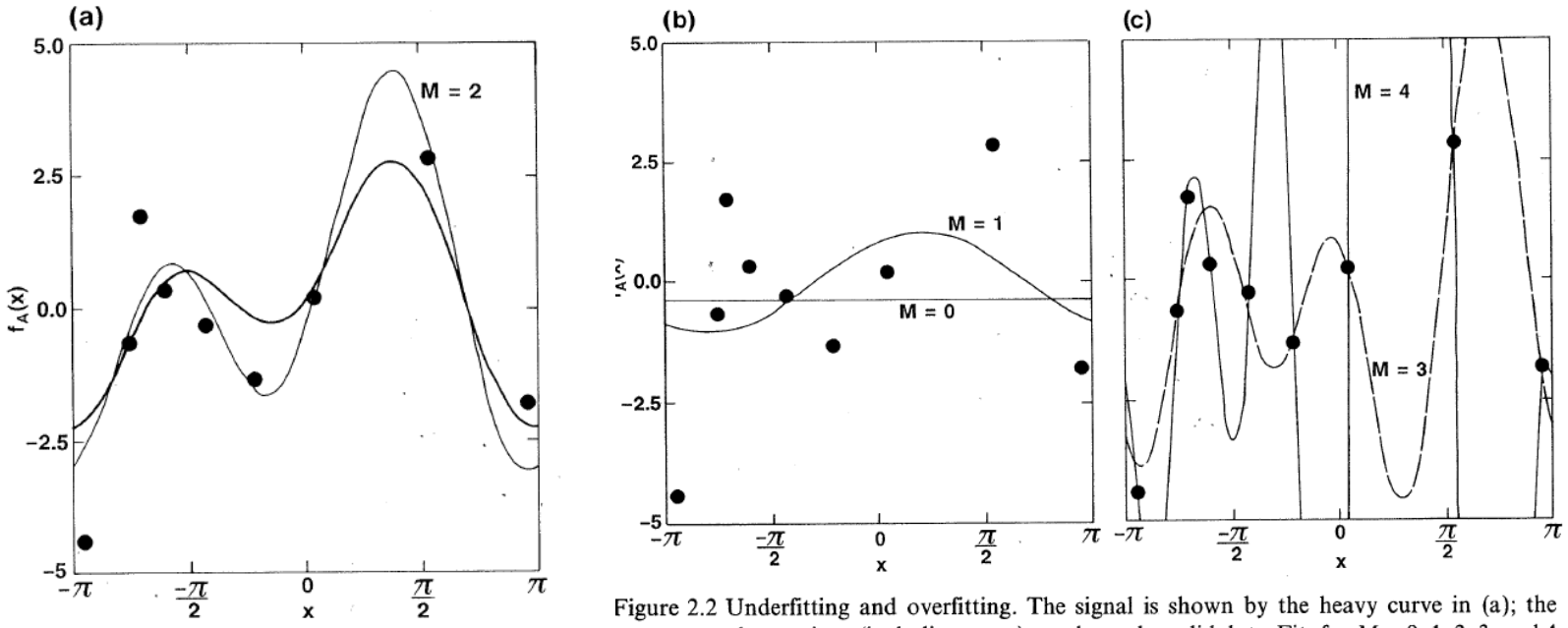

Figure 2.2 Underfitting and overfitting. The signal is shown by the heavy curve in (a); the observations (including error) are shown by solid dots. Fits for $M = 0, 1, 2, 3$, and 4 are shown in (a), (b), and (c).