Linear models

1 Linear functions

A linear model encodes the assumption that two quantities are linearly related. Mathematically, this is characterized using **linear functions**. A linear function is a function such that a linear combination of inputs is mapped to the same linear combination of the corresponding outputs.

Definition 1.1 (Linear function)**.** *A linear function $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^m$ is a function that maps vectors in $\mathbb{R}^n$ to vectors in R^m such that for any scalar $\alpha \in \mathbb{R}$ and any vectors $x_1, x_2 \in \mathbb{R}^n$*

$$\mathcal{T}(x_1 + x_2) = \mathcal{T}(x_1) + \mathcal{T}(x_2), \tag{1}$$

$$\mathcal{T}(\alpha\, x_1) = \alpha\, \mathcal{T}(x_1) \tag{2}$$

Multiplication with a matrix of dimensions $m \times n$ maps vectors in $\mathbb{R}^n$ to vectors in $\mathbb{R}^m$. For a fixed matrix, this is a linear function. Perhaps surprisingly, the converse is also true: *any* linear function between $\mathbb{R}^n$ and $\mathbb{R}^m$ corresponds to multiplication with a certain matrix. The proof is in Section A.1 of the appendix.

Theorem 1.2 (Equivalence between matrices and linear functions)**.** *For finite m, n every linear function $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^m$ can be represented by a matrix $T \in \mathbb{R}^{m \times n}$.*

This implies that in order to analyze linear models in finite-dimensional spaces we can restrict our attention to matrices.

1.1 Range and null space

The **range** of a matrix $A \in \mathbb{R}^{m \times n}$ is the set of all possible vectors in $\mathbb{R}^m$ that we can reach by applying the matrix to a vector in $\mathbb{R}^n$.

Definition 1.3 (Range)**.** *Let $A \in \mathbb{R}^{m \times n}$,*

$$\text{range}(A) := \{y \mid y = Ax \quad \textit{for some } x \in \mathbb{R}^n\}. \tag{3}$$

This set is a subspace of $\mathbb{R}^m$.

As we saw in the previous lecture notes (equation (43)), the product of a matrix and a vector is a linear combination of the columns of the matrix, which implies that for any matrix A

$$\text{range}\,(A) = \text{col}\,(A)\,. \tag{4}$$

In words, the range is spanned by the columns of the matrix.

The **null space** of a function is the set of vectors that are mapped to zero by the function. If we interpret Ax as data related linearly to x, then the null space corresponds to the set of vectors that are *invisible* under the measurement operator.

Definition 1.4 (Null space)**.** *The null space of $A \in \mathbb{R}^{m\times n}$ contains the vectors in $\mathbb{R}^n$ that A maps to the zero vector.*

$$\text{null}\,(A) := \{x \mid Ax = 0\}\,. \tag{5}$$

This set is a subspace of $\mathbb{R}^n$.

The following lemma shows that the null space is the orthogonal complement of the row space of the matrix.

Lemma 1.5. *For any matrix $A \in \mathbb{R}^{m\times n}$*

$$\text{null}\,(A) = \text{row}\,(A)^{\perp}\,. \tag{6}$$

The lemma, proved in Section A.2 of the appendix, implies that the matrix is invertible if we restrict the inputs to be in the row space of the matrix.

Corollary 1.6. *Any matrix $A \in \mathbb{R}^{m\times n}$ is invertible when acting on its row space. For any two nonzero vectors $x_1 \neq x_2$ in the row space of A*

$$Ax_1 \neq Ax_2. \tag{7}$$

Proof. Assume that for two different nonzero vectors x_1 and x_2 in the row space of A $Ax_1 = Ax_2$. Then $x_1 - x_2$ is a nonzero vector in the null space of A. By Lemma 1.5 this implies that $x_1 - x_2$ is orthogonal to the row space of A and consequently to itself, so that $x_1 = x_2$. □

This means that for every matrix $A \in \mathbb{R}^{m\times n}$ we can decompose any vector in $\mathbb{R}^n$ into two components: one is in the row space and is mapped to a nonzero vector in $\mathbb{R}^m$ that is unique in the sense that no other vector in $\text{row}\,(A)$ is mapped to it, the other is in the null space and is mapped to the zero vector.

1.2 Interpretation using the SVD

Recall that the left singular vectors of a matrix A that correspond to nonzero singular values are a basis of the column space of A. It follows that they are also a basis for the range. The right singular vectors corresponding to nonzero singular values are a basis of the row space. As a result any orthonormal set of vectors that forms a basis of $\mathbb{R}^n$ together with these singular vectors is a basis of the null space of the matrix. We can therefore write any matrix A such that $m \geq n$ as

$$A = [\underbrace{u_1 \; u_2 \; \cdots \; u_r}_{\text{Basis of range}(A)} \; u_{r+1} \; \cdots \; u_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & & & \cdots & & & \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} [\underbrace{v_1 \; v_2 \; \cdots \; v_r}_{\text{Basis of row}(A)} \; \underbrace{v_{r+1} \; \cdots \; v_n}_{\text{Basis of null}(A)}]^T.$$

Note that the vectors $u_{r+1}, \ldots, u_n$ are a subset of an orthonormal basis of the orthogonal complement of the range, which has dimension $m - r$.

The SVD provides a very intuitive characterization of the mapping between $x \in \mathbb{R}^n$ and $Ax \in \mathbb{R}^m$ for any matrix $A \in \mathbb{R}^{m \times n}$ with rank r,

$$Ax = \sum_{i=1}^{r} \left(\sigma_i v_i^T x \right) u_i. \tag{8}$$

The linear function can be decomposed into four simple steps:

1. Compute the projection of x onto the right singular vectors of A: $v_1^T x$, $v_2^T x$, $\ldots$, $v_r^T x$.
2. Scale the projections using the corresponding singular value: $\sigma_1 v_1^T x$, $\sigma_2 v_2^T x$, $\ldots$, $\sigma_r v_r^T x$.
3. Multiply each scaled projection with the corresponding left singular vector u_i.
4. Sum all the scaled left singular vectors.

1.3 Systems of equations

In a linear model we assume that the data $y \in \mathbb{R}^m$ can be represented as the result of applying a linear function or matrix $A \in \mathbb{R}^{m \times n}$ to an **unknown** vector $x \in \mathbb{R}^n$,

$$Ax = y. \tag{9}$$

The aim is to determine x from the measurements. Depending on the structure of A and y this may or may not be possible.

If we expand the matrix-vector product, the linear model is equivalent to a system of linear equations

$$A_{11}x[1] + A_{12}x[2] + \ldots + A_{1n}x[n] = y[1] \tag{10}$$
$$A_{21}x[1] + A_{22}x[2] + \ldots + A_{2n}x[n] = y[2] \tag{11}$$
$$\ldots \tag{12}$$
$$A_{m1}x[1] + A_{m2}x[2] + \ldots + A_{mn}x[n] = y[m]. \tag{13}$$

If the number of equations m is greater than the number of unknowns n the system is said to be **overdetermined**. If there are more unknowns than equation $n > m$ then the system is **underdetermined**.

Recall that $\text{range}\,(A)$ is the set of vectors that can be reached by applying A. If y does not belong to this set, then the system cannot have a solution.

Lemma 1.7. *The system $y = Ax$ has one or multiple solutions if and only if $y \in \text{range}\,(A)$.*

Proof. If $y = Ax$ has a solution then $y \in \text{range}\,(A)$ by (4). If $y \in \text{range}\,(A)$ then there is a linear combination of the columns of A that yield y by (4) so the system has at least one solution. □

If the null space of the matrix has dimension greater than 0, then the system cannot have a unique solution.

Lemma 1.8. *If $\dim(\text{null}\,(A)) > 0$, then if $Ax = y$ has a solution, the system has an infinite number of solutions.*

Proof. The null space has at least dimension one, so it contains an infinite number of vectors h such that for any solution x for which $y = Ax$ $x + h$ is also a solution. □

In the critical case $m = n$, linear systems may have a unique solution if the matrix is **full rank**, i.e. if all its rows (and its columns) are linearly independent. This means that the data in the linear model completely specify the unknown vector of interest, which can be recovered by inverting the matrix.

Lemma 1.9. *For any square matrix $A \in \mathbb{R}^{n \times n}$, the following statements are equivalent.*

1. $\text{null}\,(A) = \{0\}$.
2. *A is full rank.*

3. A is invertible.

4. The system $Ax = y$ has a unique solution for every vector $y \in \mathbb{R}^n$.

Proof. We prove that the statements imply each other in the order (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (1).

(1) $\Rightarrow$ (2): If $\dim(\text{null}(A)) = 0$, by Lemma 1.5 the row space of A is the orthogonal complement of $\{0\}$, so it is equal to $\mathbb{R}^n$ and therefore the rows are all linearly independent.

(2) $\Rightarrow$ (3): If A is full rank, its rows span all of $\mathbb{R}^n$ so $\text{range}(A) = \mathbb{R}^n$ and A is invertible by Corollary 1.6.

(3) $\Rightarrow$ (4) If A is invertible there is a unique solution to $Ax = y$ which is $A^{-1}y$. If the solution is not unique then $Ax_1 = Ax_2 = y$ for some $x_1 \neq x_2$ so that 0 and $x_1 - x_2$ have the same image and A is not invertible.

(4) $\Rightarrow$ (1) If $Ax = y$ has a unique solution for 0, then $Ax = 0$ implies $x = 0$. $\square$

Recall that the inverse of a product of invertible matrices is equal to the product of the inverses,

$$(AB)^{-1} = B^{-1}A^{-1}, \tag{14}$$

and that the inverse of the transpose of a matrix is the transpose of the inverse,

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T. \tag{15}$$

Using these facts, the inverse of a matrix A can be written in terms of its singular vectors and singular values,

$$\begin{aligned} A^{-1} &= \left(USV^T\right)^{-1} && (16)\\ &= \left(V^T\right)^{-1} S^{-1}U^{-1} && (17)\\ &= V \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix} U^T && (18)\\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T. && (19) \end{aligned}$$

Note that if one of the singular values of a matrix is very small, the corresponding term in (19) becomes very large. As a result, the solution to the system of equations becomes very susceptible to noise in the data. In order to quantify the stability of the solution of a system of equations we use the **condition number** of the corresponding matrix.

Definition 1.10 (Conditioning number)**.** *The condition number of a matrix is the ratio between its largest and its smallest singular values*

$$\text{cond}\,(A) = \frac{\sigma_{\max}}{\sigma_{\min}}. \tag{20}$$

If the condition number is very large, then perturbations in the data may be dramatically amplified in the corresponding solution. This is illustrated by the following example.

Example 1.11 (Ill-conditioned system)**.** The matrix

$$\begin{bmatrix} 1.001 & 1 \\ 1 & 1 \end{bmatrix} \tag{21}$$

has a condition number equal to 401. Compare the solutions to the corresponding system of equations for two very similar vectors

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{22}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix}^{-1} \begin{bmatrix} 1.1 \\ 1 \end{bmatrix} = \begin{bmatrix} 101 \\ -100 \end{bmatrix}. \tag{23}$$

---

2 Least squares

Just like a system for which $m = n$, an overdetermined system will have a solution as long as $y \in \text{range}\,(A)$. However, if $m > n$ then $\text{range}\,(A)$ is a low-dimensional subspace of $\mathbb{R}^m$. This means that even a small perturbation in a random direction is bound to kick y out of the subspace. As a result, in most cases overdetermined systems *do not have a solution.* However, we may still compute the point in $\text{range}\,(A)$ that is closest to the data y. If we measure distance using the ℓ_2 norm, then this is denoted by the method of **least squares**.

Definition 2.1 (Least squares)**.** *The method of least squares consists of estimating x by solving the optimization problem*

$$\min_{x \in \mathbb{R}^n} ||y - Ax||_2 \tag{24}$$

Tall matrices (with more rows than columns) are said to be full rank if all their columns are linearly independent. If A is full rank, then the solution to the least-squares problem has a closed-form solution given by the following theorem.

Theorem 2.2 (Least-squares solution)**.** *If $A \in \mathbb{R}^{m \times n}$ is full rank and $m \geq n$ the solution to the least-squares problem* (24) *is equal to*

$$x_{\text{LS}} = V S^{-1} U^T y \tag{25}$$

$$= \left(A^T A\right)^{-1} A^T y. \tag{26}$$

Proof. The problem (24) is equivalent to

$$\min_{z \in \text{range}(A)} ||y - z||_2 \tag{27}$$

since every $x \in \mathbb{R}^n$ corresponds to a unique $z \in \text{range}\,(A)$ (we are assuming that the matrix is full rank, so the null space only contains the zero vector). By Theorem 1.7 in Lecture Notes 9, the solution to Problem (27) is

$$\mathcal{P}_{\text{range}(A)}\, y = \sum_{i=1}^{n} \left(u_i^T y\right) u_i \tag{28}$$

$$= U U^T y. \tag{29}$$

Where $A = USV^T$ is the SVD of A, so the columns of $U \in \mathbb{R}^{m \times n}$ are an orthonormal basis for the range of A. Now, to find the solution we need to find the unique x_{LS} such that

$$A x_{\text{LS}} = USV^T x_{\text{LS}} \tag{30}$$

$$= U U^T y. \tag{31}$$

This directly implies

$$U^T USV^T x_{\text{LS}} = U^T U U^T y. \tag{32}$$

We have

$$U^T U = \text{I}, \tag{33}$$

because the columns of U are orthonormal (note that $UU^T \neq \text{I}$ if $m > n$!). As a result

$$x_{\text{LS}} = \left(SV^T\right)^{-1} U^T y \tag{34}$$

$$= \left(V^T\right)^{-1} S^{-1} U^T y \tag{35}$$

$$= V S^{-1} U^T y, \tag{36}$$

where we have used the fact that

$$V^{-1} = V^T \quad \text{and} \quad \left(V^T\right)^{-1} = V \tag{37}$$

because $V^T V = V V^T = \mathrm{I}$ (V is an $n \times n$ orthogonal matrix).

Finally,

$$\begin{aligned}
\left(A^T A\right)^{-1} A^T &= \left(V S^T U^T U S V^T\right)^{-1} V S^T U^T && (38)\\
&= \left(V \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} V^T \right)^{-1} V S^T U^T \quad \text{by (33)} \\
&= V \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} V^T V S^T U^T \quad \text{by (37)} && (39)\\
&= V \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} S U^T \quad \text{by (37)} && (40)\\
&= V S^{-1} U^T \quad \text{by (37)}, && (41)
\end{aligned}$$

where we have used that S is diagonal so $S^T = S$ and A is full rank, so that all the singular values are nonzero and S is indeed invertible. □

The matrix $\left(A^T A\right)^{-1} A^T$ is called the **pseudoinverse** of A. In the square case it reduces to the inverse of the matrix.

2.1 Linear regression

A very important application of least squares is fitting linear regression models. In linear regression, we assume that a quantity of interest can be expressed as a linear combination of other observed quantities.

$$a \approx \sum_{j=1}^{n} \theta_j c_j, \tag{42}$$

where $a \in \mathbb{R}$ is called the **response** or **dependent variable**, $c_1, c_2, \ldots, c_n \in \mathbb{R}$ are the **covariates** or **independent variables** and $\theta_1, \theta_2, \ldots, \theta_n \in \mathbb{R}$ are the parameters of the

model. Given m observations of the response and the covariates, we can place the response in a vector y and the covariates in a matrix X such that each column corresponds to a different covariate. We can then fit the parameters so that the model approximates the response as closely as possible in ℓ_2 norm. This is achieved by solving a least-squares problem

$$\min_{\theta \in \mathbb{R}^n} ||y - X\boldsymbol{\theta}||_2 \tag{43}$$

to fit the parameters.

Geometrically, the estimated parameters are those that project the response on the subspace spanned by the covariates. Alternatively, linear regression also has a probabilistic interpretation. It corresponds to computing the maximum likelihood estimator for a particular model.

Lemma 2.3. *Let $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are random vectors of dimension n such that*

$$\boldsymbol{Y} = X\boldsymbol{\theta} + \boldsymbol{Z}, \tag{44}$$

where X is a deterministic matrix (not a random variable). If Z is an iid Gaussian random vector with mean zero and unit variance then the maximum likelihood estimator of Y given Z is the solution to the least-squares problem (41)*.*

Proof. Setting $\Sigma = \mathrm{I}$ in Definition 2.20 of Lecture Notes 3, we have that the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} ||\mathbf{y} - X\boldsymbol{\theta}||_2^2\right). \tag{45}$$

Maximizing the likelihood yields

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{46}$$

$$= \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}) \tag{47}$$

$$= \arg\min_{\boldsymbol{\theta}} ||\mathbf{y} - X\boldsymbol{\theta}||_2 . \tag{48}$$

□

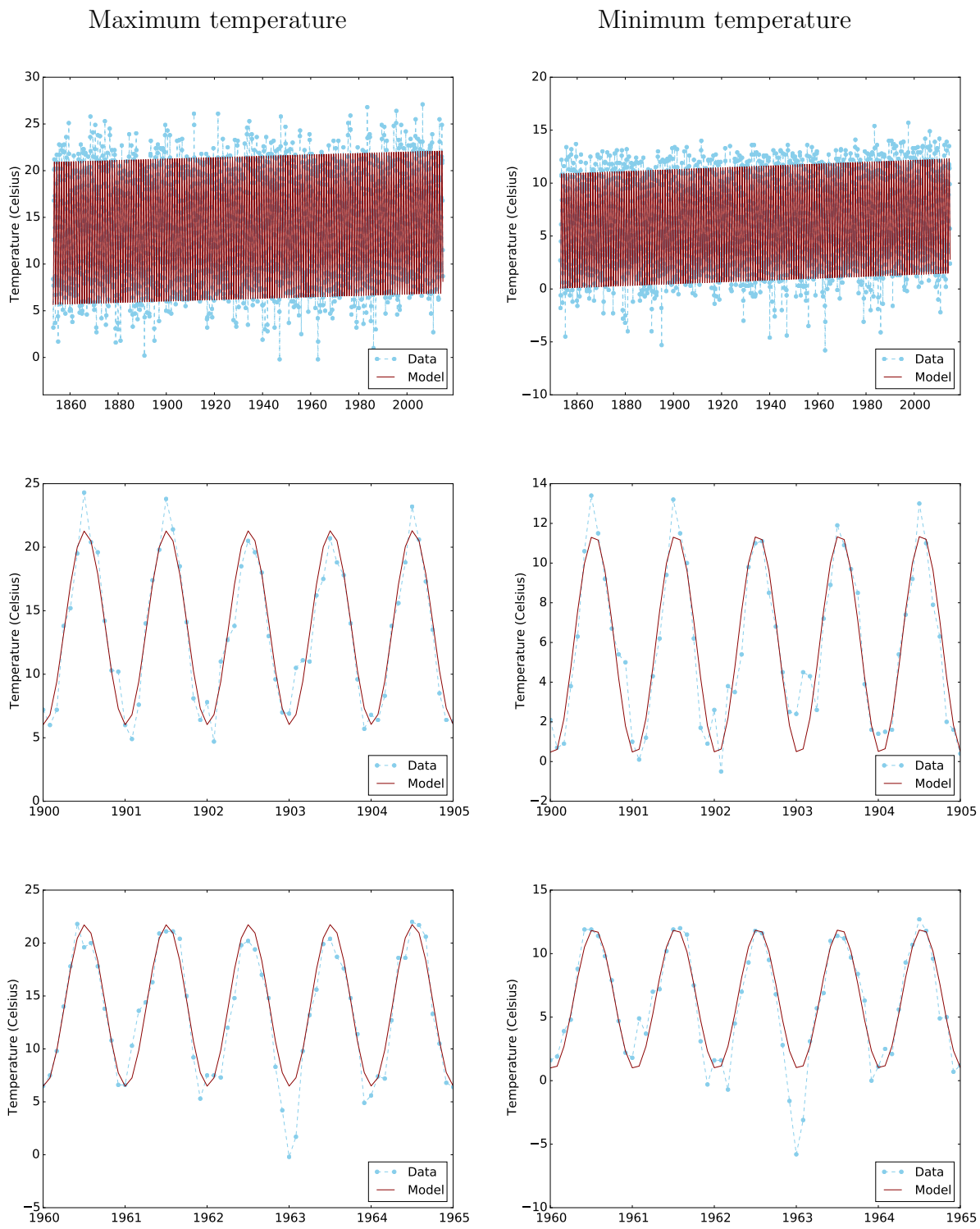


Figure 1: Data and fitted model described by Example 2.4 for maximum and minimum temperatures.

Example 2.4 (Global warming)**.** In this example we build a model for temperature data taken in a weather station in Oxford over 150 years.[1] The model is of the form

$$y \approx a + \tilde{b}\cos\left(\frac{2\pi t}{12} + \tilde{c}\right) + dt \tag{49}$$

$$= a + b\cos\left(\frac{2\pi t}{12}\right) + c\sin\left(\frac{2\pi t}{12}\right) + dt, \tag{50}$$

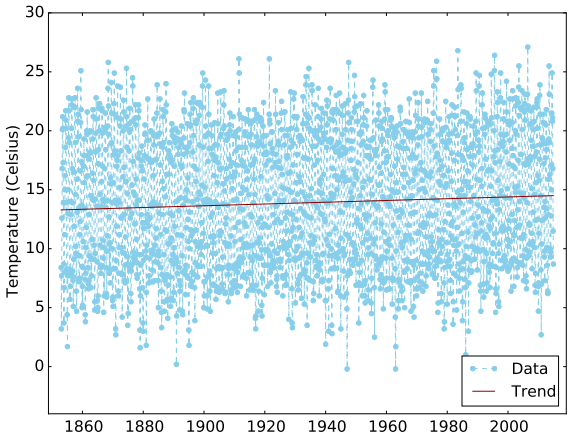
where t denotes the time in months. The parameter a represents the mean temperature, b and c account for periodic yearly fluctuations and d is the overall trend. Is d is positive then the model indicates that temperatures are increasing, whereas if it is negative then it indicates that temperatures are decreasing. To fit these parameters using the data, we build a matrix A with four columns,

$$A = \begin{bmatrix} 1 & \cos\frac{2\pi t_1}{12} & \sin\frac{2\pi t_1}{12} & dt_1 \\ 1 & \cos\frac{2\pi t_2}{12} & \sin\frac{2\pi t_2}{12} & dt_2 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \cos\frac{2\pi t_n}{12} & \sin\frac{2\pi t_n}{12} & dt_n \end{bmatrix}, \tag{51}$$

compile the temperatures in a vector y and solve a least-squares problem. The results are shown in Figures 1 and 2. The fitted model indicates that both the maximum and minimum temperatures are increasing by about 0.8 degrees Celsius (around 1.4 ° F).

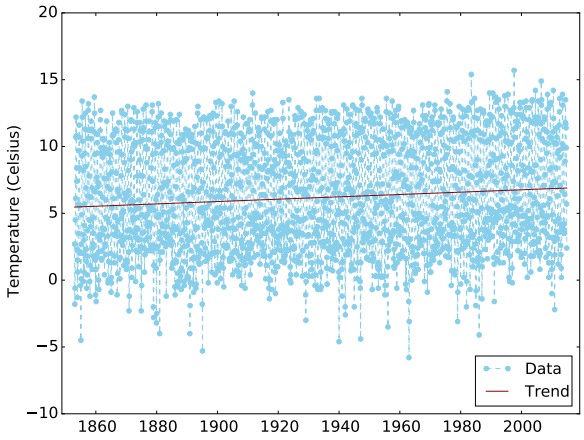
[1]The data is available at `http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt`.

Maximum temperature



+ 0.75 °C / 100 years

Minimum temperature



+ 0.88 °C / 100 years

Figure 2: Temperature trend obtained by fitting the model described by Example 2.4 for maximum and minimum temperatures.

A Proofs

A.1 Proof of Theorem 1.2

The matrix is

$$T := \begin{bmatrix} \mathcal{T}(e_1) & \mathcal{T}(e_2) & \cdots & \mathcal{T}(e_n) \end{bmatrix}, \tag{52}$$

i.e. the columns of the matrix are the result of applying $\mathcal{T}$ to the standard basis of $\mathbb{R}^n$. Indeed, for any vector $x \in \mathbb{R}^n$

$$\mathcal{T}(x) = \mathcal{T}\left(\sum_{i=1}^{n} x[i] e_i\right) \tag{53}$$

$$= \sum_{i=1}^{n} x[i] \mathcal{T}(e_i) \text{ by (1) and (2)} \tag{54}$$

$$= Tx. \tag{55}$$

A.2 Proof of Lemma 1.5

We prove (6) by showing that both sets are subsets of each other.

Any vector x in the row space of A can be written as

$$x = A^T z, \tag{56}$$

for some vector $z \in \mathbb{R}^m$. If $y \in \text{null}\,(A)$ then

$$\begin{aligned} y^T x &= y^T A^T z && (57)\\ &= (Ay)^T z && (58)\\ &= 0. && (59) \end{aligned}$$

So $\text{null}\,(A) \subseteq \text{row}\,(A)^\perp$.

If $x \in \text{row}\,(A)^\perp$ then in particular it is orthogonal to every row of A, so $Ax = 0$ and $\text{row}\,(A)^\perp \subseteq \text{null}\,(A)$.