Statistics for Chemical Engineers:
From Data to Models to Decisions

Victor M. Zavala

Department of Chemical and Biological Engineering
University of Wisconsin-Madison

victor.zavala@wisc.edu

Chapter 5: Estimation for Structural Models

Estimation Techniques for Structural Models

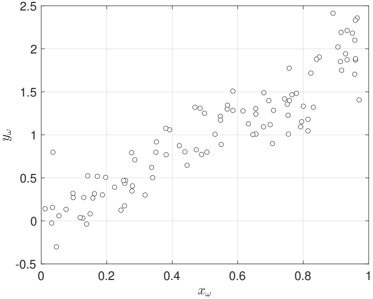
- Our discussion so far has centered on using data to characterize random phenomena.
- Our approach has been to determine an RV model that best explains the data and we have used estimation techniques to determine the parameters for such models.
- From discussion on multivariate RVs we observed that correlations between RVs uncover connections between variables (e.g., pressure and conversion).
- We now turn our attention into developing techniques that use data to develop models that capture *structural* relationships between variables.
- Structural models embed *knowledge* that allow us to make predictions for a set of variables from another set of variables.
- Structural models might originate from fundamental laws (mechanistic models), from empirical observations (empirical models), or from combinations.
- As with RV models, we will postulate a structural model (with parameters) and we will use estimation techniques to determine parameters that best fit the data.

Scalar Linear Models

- We begin our discussion by considering univariate RVs Y and X.

- Consider experiments $\omega \in \mathcal{S}$ and observation pairs (y_ω, x_ω).

- Typically, input variable X is a variable that we can control when designing an experiment while output variable Y is the result of the experiment.



- From the data above, it is clear that pairs (y_ω, x_ω) follow a *systematic trend*; however, we note that the linear trend is not perfect (contains variability).

Scalar Linear Models

- We postulate that behavior in Y follows systematic dependence on X of the form:

$$Y = \theta X + \epsilon$$

- Here, θ is a parameter that captures *linear* trend between X and Y.
- For now, assume that X and θ are deterministic variables and therefore any variability observed in Y is the result of a hidden RV ϵ.
- Postulated model contains deterministic component θX. This component will be called the *structural model* and captures the mechanism by which Y depends on X.
- Structural model θX captures knowledge about Y and allows us to make predictions.
- Postulated model also obtains a random component, which is modeled as RV ϵ. Here, we assume that RV model is $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- Random component is called random model (or noise model) and captures what we do not know about Y. Random model contains parameter σ.

Scalar Linear Models

- Assume observation pairs (y_ω, x_ω) are related as:

$$y_\omega = \theta x_\omega + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

- Here, ϵ_ω is random sample of ϵ (it is an RV); as such, y_ω is an RV.
- Input X is deterministic variable and with known values x_ω.
- Parameter θ is a deterministic variable but we do not know its value.
- Can write model as $\epsilon_\omega = y_\omega - \theta x_\omega$; noise variable captures discrepancy between model and observed output (it is a residual).
- Goal is to find θ, σ that maximize knowledge extracted (θx_ω) from data x_ω, y_ω.
- Aspects of the data that cannot be explained with model (ϵ_ω) represent the unknown; consequently, our goal is equivalent to minimize residual.
- An interesting analogy of this process is that of a separation system.

Scalar Linear Models

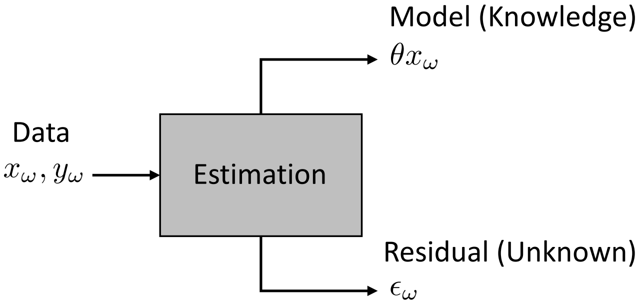


Figure: Extraction of knowledge from data. The structural model represents our knowledge while the random model represents the unknown (what cannot be explained with the structural model).

Scalar Linear Models

- Since x_ω and θ are deterministic, y_ω is linear transformation of Gaussian RV ϵ_ω.
- Consequently, y_ω is Gaussian RV $\mathcal{N}(\theta x_\omega, \sigma^2)$.
- Expected behavior of output is deterministic model.
- Variance of its behavior results from variance of noise.

Maximum Likelihood Estimation

- Most popular technique to estimate parameters is MLE.
- Define $f(y_\omega | x_\omega, \theta, \sigma)$ as conditional prob that $Y = y_\omega$ given values of x_ω, θ, and σ.
- Observations are *random* samples with joint likelihood:

$$f(\mathbf{y}|\mathbf{x}, \theta, \sigma) = \prod_{\omega \in \mathcal{S}} f(y_\omega | x_\omega, \theta, \sigma).$$

- Goal is to find θ that maximizes log likelihood function:

$$\max_{\theta} \ \log L(\theta) = \sum_{\omega \in \mathcal{S}} \log f(y_\omega | x_\omega, \theta, \sigma)$$

- We know that $y_\omega \sim \mathcal{N}(\theta x_\omega, \sigma^2)$ and thus:

$$\log f(y_\omega | x_\omega, \theta, \sigma) = -\log \sqrt{2\pi\sigma^2} - \frac{(y_\omega - \theta x_\omega)^2}{2\sigma^2}$$

- MLE problem can be written as:

$$\max_{\theta, \sigma^2} \ -\frac{S}{2} \log 2\pi - \frac{S}{2} \log \sigma^2 - \sum_{\omega \in \mathcal{S}} \frac{(y_\omega - \theta x_\omega)^2}{2\sigma^2}$$

Maximum Likelihood Estimation

- Assume now that σ^2 is a known quantity, first and second terms are fixed constants that do not affect the solution of optimization problem.

- We can thus express problem in the equivalent form:

$$\min_{\theta} \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - \theta x_\omega)^2$$

- This is a least-squares problem that estimates θ by minimizing discrepancy between observed output y_ω and deterministic model prediction θx_θ.

- Function to be minimized in this problem is known as the sum of squared errors (SSE) and will be denoted as $SSE(\theta)$.

- Value of θ that solves the problem is known as *maximum likelihood estimate* $\hat{\theta}$.

- Because observations used to estimate $\hat{\theta}$ are random, estimate is an RV.

Maximum Likelihood Estimation

- Estimate $\hat{\theta}$ that minimizes $SSE(\theta) = \frac{1}{2}\sum_{\omega\in\mathcal{S}}(y_\omega - \theta x_\omega)^2$ must satisfy:

$$\frac{\partial SSE(\hat{\theta})}{\partial\theta} = -\sum_{\omega\in\mathcal{S}} x_\omega(y_\omega - \hat{\theta}x_\omega) = 0$$

- Satisfying this guarantees that $\hat{\theta}$ is a stationary point but does not guarantee that this is a minimum point for $SSE(\theta)$ (i.e., max point also satisfies this).
- We thus say this condition is *necessary but not sufficient*.
- Sufficient condition guaranteeing that $SSE(\hat{\theta})$ is a unique minimum is:

$$\frac{\partial^2 SSE(\hat{\theta})}{\partial\theta^2} > 0.$$

- Necessary condition tells us that the estimate is given by $\hat{\theta} = \frac{\sum_{\omega\in\mathcal{S}} x_\omega y_\omega}{\sum_{\omega\in\mathcal{S}} x_\omega^2}$
- Sufficient condition tells us that $\sum_{\omega\in\mathcal{S}} x_\omega^2 > 0$.
- Estimate $\hat{\theta}$ is *unique* and becomes better defined as we add data (min is sharper).

Maximum Likelihood Estimation

- Estimate $\hat{\theta}$ gives residuals $\hat{\epsilon}_\omega = y_\omega - \hat{\theta} x_\omega, \; \omega \in \mathcal{S}$.

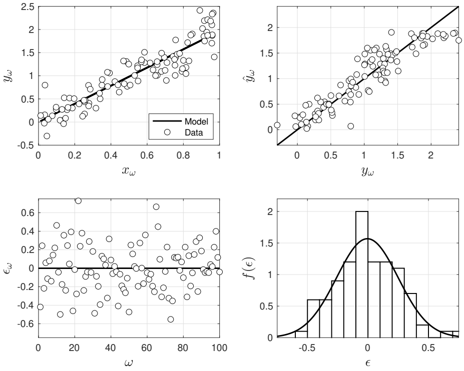
- If residual estimates follow our assumption that $\hat{\epsilon}_\omega \sim \mathcal{N}(0, \sigma^2)$ then available data and postulated model are satisfactory.

- If this is not satisfied, more data or another model is needed (e.g., nonlinear).

- This is because noise might be absorbing aspects of trend (it is not purely random).

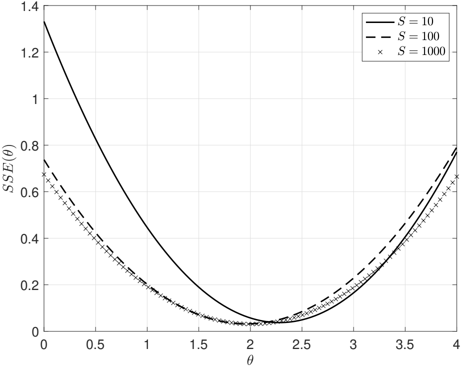
Example: Linear Estimation with Correct Model `ch5_linear_est_example.m`

- Consider X, Y with structural relationship $Y = \theta X$ and $\theta = 2$.
- Assume that observations Y are corrupted by noise $\epsilon_\omega \sim \mathcal{N}(0, 0.0625)$.
- Use model to generate $S = 100$ observations (x_ω, y_ω).
- We postulate a model $Y = \theta X + \epsilon$ with $\epsilon \in \mathcal{N}(0, \sigma^2)$ and estimate θ using MLE.



Example: Linear Estimation with Correct Model `ch5_linear_est_example.m`

- Plot $SSE(\theta)$ for different amounts of data $S = 10, 100, 1000$.
- Surface becomes better defined as we add data and reveals true value $\theta = 2$.
- If we use $S = 10$, $SSE(\theta)$ is minimized at $\hat{\theta} = 2.2$.
- If we use $S = 100$ or $S = 1000$ obs, $SSE(\theta)$ is minimized at $\hat{\theta} = 2$.



Bias, Precision, and Information

- MLE delivers an unbiased estimate $\hat{\theta}$ of true parameters θ.
- To see this, write MLE problem as:

$$\min_{\theta} \; -\log L(\theta) = -\log f(\mathbf{y}|\theta).$$

- Solution satisfies necessary condition (a.k.a. score function):

$$-\frac{\partial \log f(\mathbf{y}|\hat{\theta})}{\partial \theta} = 0.$$

- Estimate $\hat{\theta}$ is RV because it depends on random data $\mathbf{y}$.
- Imagine we use observations $\mathbf{y}$ to obtain an estimate $\hat{\theta}$ and repeat this with different observations $\mathbf{y}$ to obtain different estimates $\hat{\theta}$.
- Can we guarantee that this process delivers estimates satisfying $\mathbb{E}[\hat{\theta}] = \theta$?
- One can show that the answer is yes, because one can show that:

$$\mathbb{E}\left[-\frac{\partial \log f(\mathbf{y}|\theta)}{\partial \theta}\right] = 0 \qquad \text{(expectation wrt } \mathbf{y}\text{)}$$

Bias, Precision, and Information

- Now explore how does $\hat{\theta}$ *vary* as we vary observations $\mathbf{y}$.
- Variance tell us how certain we are that our estimate $\hat{\theta}$ is true value θ and recall also that variance is measure of estimate precision.
- To assess how much $\hat{\theta}$ varies, we need to assess variance of score function:

$$\mathcal{I}(\theta) = \mathbb{V}\left[-\frac{\partial \log f(\mathbf{y}|\theta)}{\partial \theta}\right]$$

- Function $\mathcal{I}(\theta)$ is known as the *Fisher information* and can be expressed as:

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial \theta^2}\right]$$

- FI is thus related to second derivative of log likelihood function.
- FI gives limit of best possible estimator precision. Specifically, one can show that:

$$\mathbb{V}[\hat{\theta}] \geq \frac{1}{\mathcal{I}(\theta)}.$$

- This is known as Cramer-Rao bound.

Bias, Precision, and Information

- FI in general cannot be computed (θ is not known and requires expectation wrt $\mathbf{y}$).
- In some special situations, however, FI can be computed.
- This is the case for linear models; to see this, note that log likelihood is:

$$-\log f(\mathbf{y}|\theta) = \frac{S}{2}\log 2\pi + \frac{S}{2}\log\sigma^2 + \sum_{\omega\in\mathcal{S}} \frac{(y_\omega - \theta x_\omega)^2}{2\sigma^2}.$$

- The second derivative of this function is:

$$-\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial\theta^2} = \frac{1}{\sigma^2}\sum_{\omega\in\mathcal{S}} x_\omega^2.$$

- FI is thus:

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial\theta^2}\right] = \frac{1}{\sigma^2}\sum_{\omega\in\mathcal{S}} x_\omega^2.$$

- As increase amount of data (S larger), FI increases (minimum becomes sharper).
- As variance of $\mathbf{y}$ decreases (σ^2 smaller), FI increases.

Estimation of Structural and Random Parameters

- So far, we have assumed that σ is a deterministic variable with known value.
- If not known, parameter σ can be estimated using MLE.
- To see this, consider MLE problem:

$$\min_{\theta,\sigma} \; -\log L(\theta,\sigma) = -\sum_{\omega\in\mathcal{S}} \log f(y_\omega | x_\omega, \theta, \sigma).$$

- Rearranging:

$$\min_{\theta,\sigma} \; \frac{S}{2}\log 2\pi + \frac{S}{2}\log \sigma^2 + \frac{1}{2\sigma^2}\sum_{\omega\in\mathcal{S}} (y_\omega - \theta x_\omega)^2$$

- Estimates must satisfy optimality conditions:

$$-\frac{\partial \log L(\theta,\sigma)}{\partial\theta} = 0$$
$$-\frac{\partial \log L(\theta,\sigma)}{\partial\sigma^2} = 0.$$

Estimation of Structural and Random Parameters

- First optimality condition yields:

$$\frac{1}{\sigma^2}\sum_{\omega\in\mathcal{S}} x_\omega(y_\omega-\theta x_\omega)=0.$$

- Multiplying through by σ^2 and rearranging we obtain:

$$\hat{\theta}=\frac{\sum_{\omega\in\mathcal{S}} x_\omega y_\omega}{\sum_{\omega\in\mathcal{S}} x_\omega^2}.$$

- We thus have that $\hat{\theta}$ does not depend on σ.

- Second optimality condition yields:

$$\frac{S}{2}\frac{1}{\sigma^2}-\frac{1}{2(\sigma^2)^2}\sum_{\omega\in\mathcal{S}}(y_\omega-\theta x_\omega)^2=0.$$

- Using estimate $\hat{\theta}$ and rearranging:

$$\hat{\sigma}^2=\frac{1}{S}\sum_{\omega\in\mathcal{S}}(y_\omega-\hat{\theta}x_\omega)^2.$$

- Estimate $\hat{\sigma}^2$ is thus sample variance of residuals.

Multidimensional Linear Models

- We now generalize linear model to account for multiple inputs.

$$Y = \theta_0 + \sum_{i=1}^{n} \theta_i X_i + \epsilon$$

- Observation pairs (y_ω, x_ω) satisfy:

$$y_\omega = \theta_0 + \sum_{i=1}^{n} \theta_i x_{\omega,i} + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

- This set of equations can be expressed compactly using matrix notation:

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & & & \vdots & \\ 1 & x_{S,1} & x_{S,2} & \dots & x_{S,n} \end{bmatrix},\ \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix},\ \mathbf{e} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_S \end{bmatrix}.$$

Multidimensional Linear Models

- Assume random noise is S-dimensional RV $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$.

- Here, $\Sigma \in \mathbb{R}^{S \times S}$ is noise covariance.

- Noise covariance captures variance and correlations in sensor errors.

- As in scalar case, assume$\mathbf{X}$, θ, and Σ are deterministic variables.

- Output is thus $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\theta, \Sigma)$ with joint density $f(\mathbf{y}|\mathbf{X}, \theta, \Sigma)$.

- Estimates $\hat{\theta}$ and $\hat{\Sigma}$ are found by maximizing log likelihood

$$\max_{\theta, \Sigma} \log f(\mathbf{y}|\mathbf{X}, \theta, \Sigma) = -\log\left((2\pi)^{S/2}|\Sigma|^{1/2}\right) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\theta).$$

- This is most general form of MLE problem.

Multidimensional Linear Models

- If we assume noise covariance Σ is given, MLE problem reduces to:

$$\min_{\theta} \; -\log f(\mathbf{y}|\mathbf{X}, \theta, \Sigma) = \log\left((2\pi)^{S/2}|\Sigma|^{1/2}\right) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\theta).$$

- Since Σ is a constant, first term in log likelihood is a constant and can be eliminated.

- MLE problem is thus reduced to:

$$\min_{\theta} \; \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\theta).$$

- This is a least-squares problem (covariance acts as *weighting* matrix).

Multidimensional Linear Models

- If we assume Σ is diagonal with entries $\Sigma_{\omega,\omega} = \sigma_\omega^2$ for $\omega \in \mathcal{S}$:

$$\min_\theta \ \frac{1}{2} \sum_{\omega \in \mathcal{S}} \frac{1}{\sigma_\omega^2} (y_\omega - \mathbf{x}_\omega^T \theta)^2.$$

- Reciprocal of variance $1/\sigma_\omega^2$ acts as *weight* for observation ω.

- If we further assume noise variances are equal ($\sigma_\omega^2 = \sigma^2$):

$$\min_\theta \quad \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - \mathbf{x}_\omega^T \theta)^2.$$

- This is a min problem for sum-of-squared errors (simplest form of MLE problem).

Multidimensional Linear Models

- SSE function can be written compactly as:

$$
\begin{aligned}
SSE(\theta) &= \frac{1}{2}\sum_{\omega\in\mathcal{S}}(y_\omega - \mathbf{x}_\omega^T\theta)^2 \\
&= \frac{1}{2}(\mathbf{y}-\mathbf{X}\theta)^T(\mathbf{y}-\mathbf{X}\theta) = \frac{1}{2}\|\mathbf{y}-\mathbf{X}\theta\|^2.
\end{aligned}
$$

- Solution of SSE min problem must satisfy necessary conditions:

$$\nabla_\theta SSE(\theta) = 0.$$

- Here, $\nabla_\theta SSE(\theta) \in \mathbb{R}^{n+1}$ is gradient vector:

$$\nabla_\theta SSE(\theta) = \begin{bmatrix} \frac{\partial SSE(\theta)}{\partial\theta_0} \\ \frac{\partial SSE(\theta)}{\partial\theta_1} \\ \vdots \\ \frac{\partial SSE(\theta)}{\partial\theta_n} \end{bmatrix}.$$

Multidimensional Linear Models

- Gradient vector is given by:

$$\nabla_\theta SSE(\theta) = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta).$$

- This is set of $n + 1$ linear equations that must be solved to obtain $\hat{\theta}$.

- Explicit solution can be found by using matrix manipulations:

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Recall that this estimate is only guaranteed to be a stationarity point for $SSE(\theta)$.

- A stationary point can be a min or max; moreover, such point might be non-unique.

- Non-uniqueness indicates that multiple values of θ give same min value of SSE.

Multidimensional Linear Models

- Estimate $\hat{\theta}$ is *unique min* of $SSE(\theta)$ if sufficient condition holds:

$$\nabla_\theta^2 SSE(\theta) \succ 0$$

- Here, $\nabla_\theta^2 SSE(\theta)$ is Hessian matrix:

$$\nabla_\theta^2 SSE(\theta) = \begin{bmatrix} \frac{\partial^2 SSE(\theta)}{\partial\theta_0\partial\theta_0} & \frac{\partial^2 SSE(\theta)}{\partial\theta_1\partial\theta_0} & \cdots & \frac{\partial^2 SSE(\theta)}{\partial\theta_n\partial\theta_0} \\ \frac{\partial^2 SSE(\theta)}{\partial\theta_0\partial\theta_1} & \frac{\partial^2 SSE(\theta)}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 SSE(\theta)}{\partial\theta_n\partial\theta_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 SSE(\theta)}{\partial\theta_0\partial\theta_n} & \frac{\partial^2 SSE(\theta)}{\partial\theta_1\partial\theta_n} & \cdots & \frac{\partial^2 SSE(\theta)}{\partial\theta_n\partial\theta_n} \end{bmatrix}.$$

- Hessian is given by $\nabla_\theta^2 SSE(\theta) = \mathbf{X}^T\mathbf{X}$.

- Recall that suff condition for scalar case requires second derivative to be positive.

- Suff condition in multivariate case requires that the Hessian is positive definite (all its eigenvalues are strictly positive).

Example: Estimating Heat Capacity `ch5_heat_capacity_lin_est.m`

- Calorimetry experiments used to determine heat capacity as function of temperature.

- Hypothesize a structural relationship of the form:

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \epsilon,$$

where Y is heat capacity (in cal/mol-K) and X is temperature (in K).

- Model has three parameters and relationship between Y and X is nonlinear.

- Model, however, is linear in the parameters; to see this, we write model as:

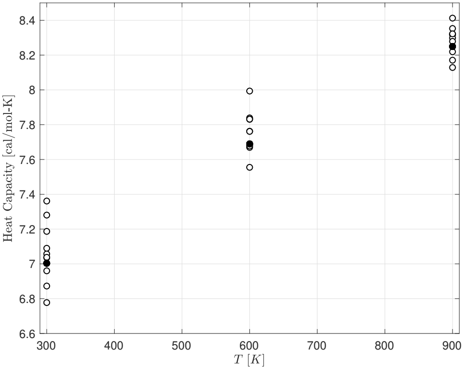
$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \epsilon$$

where $X_1 = X$ as temperature and $X_2 = X^2$ as squared temperature.

Example: Estimating Heat Capacity `ch5_heat_capacity_lin_est.m`

- Conduct a total of $S = 30$ experiments (10 replicates at each temperature).

- Why is there variability?



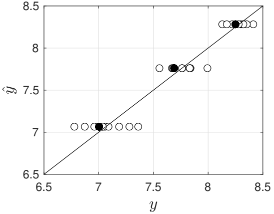
- This variability in replicates is the type of behavior that the random component of our model captures. Experiments, in general, are not perfectly reproducible.

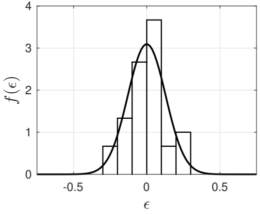
Example: Estimating Heat Capacity `ch5_heat_capacity_lin_est.m`

- We usually do not have information about sensor accuracy.
- Assume noise covariance is diagonal with entries $\Sigma_{\omega,\omega} = \sigma^2$.
- MLE thus reduces to minimization of SSE with solution:

$$\hat{\theta} = \left[\begin{array}{r} 6.19 \times 10^{+0} \\ 3.17 \times 10^{-3} \\ -9.60 \times 10^{-7} \end{array} \right].$$

- Fit and residuals are shown below





- Hessian $\nabla^2_\theta SSE(\theta) = \mathbf{X}^T\mathbf{X}$ has positive eigenvalues (estimates are unique).

Multidimensional Linear Models

- Recall that observations are random $\mathbf{y} = \mathbf{X}\theta + \mathbf{e}$ (θ is true parameter).
- Can thus express estimate $\hat{\theta}$ as:

$$\begin{aligned}\hat{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta + \mathbf{e})\\ &= \theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{e}.\end{aligned}$$

- Estimate $\hat{\theta}$ is an RV that results from a linear transformation of $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$:

$$\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

- Expected estimate is $\mathbb{E}[\hat{\theta}] = \theta$ ($\hat{\theta}$ is an *unbiased* estimator of θ).
- Covariance of estimate is $\mathrm{Cov}[\hat{\theta}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.
- Hessian matrix $\mathbf{X}^T\mathbf{X}$ plays key role (more on this later).
- For non-diagonal noise covariance we have that $\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1})$.

Example: Estimating Heat Capacity `ch5_heat_capacity_lin_est.m`

- From fit of residuals we find $\epsilon \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ with $\hat{\mu} = 5.02 \times 10^{-14}$ and $\hat{\sigma} = 0.13$.
- We confirm that noise has a mean of zero and $\hat{\theta}$ is unbiased.
- From this we also find that parameter covariance matrix is:

$$
\begin{aligned}
\mathrm{Cov}[\hat{\theta}] &= (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \\
&= \begin{bmatrix} 3.15 \times 10^{-2} & -1.16 \times 10^{-4} & 9.23 \times 10^{-8} \\ -1.16 \times 10^{-4} & 4.52 \times 10^{-7} & -3.69 \times 10^{-10} \\ 9.23 \times 10^{-8} & -3.69 \times 10^{-10} & 3.07 \times 10^{-13} \end{bmatrix}.
\end{aligned}
$$

- Parameter variances (diagonal entries) are small; estimates have a small uncertainty.

Bias, Precision, and Information

- Fisher information in multivariate case is a matrix:

$$\mathcal{I}(\theta) = \mathbb{E}[-\nabla_\theta^2 \log f(\mathbf{y}|\theta)].$$

- FI is Hessian matrix of log likelihood function.
- Use FI to generalize Cramer-Rao bound:

$$\mathrm{Cov}[\hat{\theta}] \succeq \mathcal{I}(\theta)^{-1}.$$

- From this we can conclude that:

$$\mathrm{Cov}[\hat{\theta}]_{jj} \geq \mathcal{I}(\theta)_{jj}^{-1}, \quad j = 0, ..., n$$

- For multidimensional linear model:

$$\mathcal{I}(\theta) = \mathbf{X}^T \Sigma^{-1} \mathbf{X}.$$

- This is inverse of parameter covariance $\mathrm{Cov}[\hat{\theta}] = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$.

Example: Estimating Heat Capacity `ch5_heat_capacity_lin_est.m`

- Since model is linear, FI is inverse of $\mathrm{Cov}[\hat{\theta}]$:

$$\mathcal{I}(\theta) = \mathbf{X}^T \Sigma^{-1} \mathbf{X} = \begin{bmatrix} 1.80 \times 10^3 & 1.08 \times 10^6 & 7.57 \times 10^8 \\ 1.08 \times 10^6 & 7.57 \times 10^8 & 5.84 \times 10^{11} \\ 7.57 \times 10^8 & 5.84 \times 10^{11} & 4.77 \times 10^{14}. \end{bmatrix}.$$

- FI has large eigenvalues $(4.77 \times 10^{14}, 4.19 \times 10^7, 3.16 \times 10^1)$.

- We conclude that data contains significant information.

- This also indicates that estimates lie on a sharp minimum of log likelihood.

Residual Analysis

- We now develop strategies to *quantify* quality of MLE estimate $\hat{\theta}$.
- We will analyze statistics of residuals and confidence of estimates.
- Recall that postulated model makes assumption that random model is $\epsilon \sim \mathcal{N}(0, \Sigma)$.
- After obtaining $\hat{\theta}$, it is important to verify that residuals are indeed Gaussian.
- Otherwise, random model might contain systematic errors (it is not truly random)
- A useful technique to visually confirm that residuals are Gaussian is the Q-Q plot.

Residual Analysis

How much of the variability in $\mathbf{y}$ can be explained by knowledge (model predictions $\hat{\mathbf{y}}$) and how much cannot be explained (model error ϵ)?

- Total variability is given by:

$$S_y = \sum_{\omega \in \mathcal{S}} (y_\omega - \bar{y})^2 \text{ with } \bar{y} = \frac{1}{S} \sum_{\omega \in \mathcal{S}} y_\omega$$

- Total variability can be decomposed as:

$$S_y = \underbrace{\sum_{\omega \in \mathcal{S}} (\hat{y}_\omega - \bar{y})^2}_{S_m} + \underbrace{\sum_{\omega \in \mathcal{S}} (y_\omega - \hat{y}_\omega)^2}_{S_e}$$

- S_m is *model sum-of-squares* and S_e is *sum-of-squared errors* (SSE).

- Based on these quantities we define index (coefficient of determination):

$$R^2 = \frac{S_m}{S_y} = 1 - \frac{S_e}{S_y}$$

- Index is fraction of variability captured by model; one can show that this index is directly related to the Pearson correlation of the predicted and observed outputs.

Uncertainty Quantification

How much can we trust our parameter estimate and model predictions?

- For linear model the estimate is gaussian $\hat{\theta} \sim \mathcal{N}(\theta, \mathrm{Cov}[\hat{\theta}])$ with:

$$\mathrm{Cov}[\hat{\theta}] = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}.$$

- One can use covariance matrix to obtain $1-\alpha$ confidence intervals:

$$\theta_i = \hat{\theta}_i \pm m_i,\ i = 0, ..., n \qquad \text{with} \qquad m_i = \sqrt{q_{n+1} \mathrm{Cov}[\hat{\theta}]_{ii}}$$

- Here, q_{n+1} is the $(1-\alpha)$-quantile of $\chi^2(n+1)$.
- Size of CIs (m_i) is a measure of their *precision*.
- Precision affected by number of parameters $n+1$ (q_{n+1} increases with n)
- Precision affected by variances of estimated parameters $\mathbb{V}[\hat{\theta}_i] = \mathrm{Cov}[\hat{\theta}]_{ii}$.

Uncertainty Quantification

- Can construct CIs for model predictions by noticing that the output $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ is a linear transformation of the Gaussian $\hat{\theta}$ and thus:

$$\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{y}, \mathbf{X}\mathrm{Cov}[\hat{\theta}]\mathbf{X}^T).$$

- Expected value of predictions $\mathbb{E}[\hat{\mathbf{y}}]$ is true (noise-free) output $\mathbf{y} = \mathbf{X}\theta$.

- We also have that the covariance of the predictions is:

$$\mathrm{Cov}[\hat{\mathbf{y}}] = \mathbf{X}\mathrm{Cov}[\hat{\theta}]\mathbf{X}^T$$

- From these quantities, we can obtain CIs for predictions:

$$y_\omega = \hat{y}_\omega \pm m_\omega, \qquad \text{with} \qquad m_\omega = \sqrt{q_S \,\mathrm{Cov}[\hat{\mathbf{y}}]_{\omega\omega}}.$$

- Here, q_S is the $(1-\alpha)$-quantile of $\chi^2(S)$.

Example: Estimating Heat Capacity `ch5_heat_capacity_uncertainty.m`

- Parameter estimate and covariance are:

$$\hat{\theta} = \begin{bmatrix} 6.19 \times 10^{+0} \\ 3.17 \times 10^{-3} \\ -9.60 \times 10^{-7} \end{bmatrix},$$

$$\text{Cov}[\hat{\theta}] = \begin{bmatrix} 3.15 \times 10^{-2} & -1.16 \times 10^{-4} & 9.23 \times 10^{-8} \\ -1.16 \times 10^{-4} & 4.52 \times 10^{-7} & -3.69 \times 10^{-10} \\ 9.23 \times 10^{-8} & -3.69 \times 10^{-10} & 3.07 \times 10^{-13} \end{bmatrix}.$$

- Compute $q_{n+1} = q_3 = 7.81$ (this is 0.95 quantile of $\chi^2(3)$) and thus:

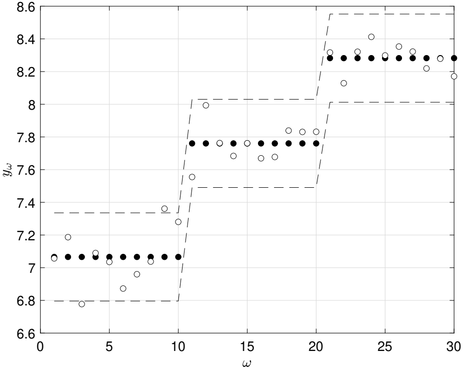
$$m_0 = \sqrt{7.81 \cdot 3.15 \times 10^{-2}} = 4.96 \times 10^{-1}$$
$$m_1 = \sqrt{7.81 \cdot 4.52 \times 10^{-7}} = 1.88 \times 10^{-3}$$
$$m_2 = \sqrt{7.81 \cdot 3.07 \times 10^{-13}} = 1.55 \times 10^{-6}.$$

- 95% confidence intervals for true parameters are:

$$\theta_0 = 6.19 \times 10^{+0} \pm 4.96 \times 10^{-1}$$
$$\theta_1 = 3.17 \times 10^{-3} \pm 1.88 \times 10^{-3}$$
$$\theta_2 = -9.60 \times 10^{-7} \pm 1.55 \times 10^{-6}.$$

Example: Estimating Heat Capacity `ch5_heat_capacity_uncertaint.m`

- Compute CIs for predicted heat capacities.
- Compute $q_S = 43.77$, this is 0.95 quantile of $\chi^2(S)$ with $S = 30$.
- Compute $\text{Cov}[\hat{\mathbf{y}}] = \mathbf{X}\text{Cov}[\hat{\theta}]\mathbf{X}^T$ (a 30× 30 matrix).
- Confidence intervals for outputs, observations, and true values are shown below.



Data Sufficiency and Overfitting

Did we use enough data to estimate model parameters?

- Hessian $\mathcal{H}(\hat{\theta}) = -\nabla_\theta^2 \log f(\mathbf{y}|\hat{\theta})$ plays a key role in answering this question.

- We analyze Hessian in context of *linear model*: $\mathcal{H}(\hat{\theta}) = \mathbf{X}^T \Sigma^{-1} \mathbf{X}$.

- Hessian (a.k.a. kernel) contains all input data and does not depend on $\hat{\theta}$ and $\mathbf{y}$.

- Recall that Hessian is FI matrix and inverse of parameter covariance:

$$\mathcal{H}(\hat{\theta}) = \mathcal{I}(\hat{\theta}) = \mathrm{Cov}[\hat{\theta}]^{-1} = \mathbf{X}^T \Sigma^{-1} \mathbf{X}.$$

- This connects concepts of sharpness of minimum, information, and uncertainty.

Data Sufficiency and Overfitting

- Eigenvalues of Hessian $\mathcal{H}(\hat{\theta})$ help us *quantify* data sufficiency.

- For $\hat{\theta}$ to be minimizer, we need that all eigenvalues of are non-negative (either zero or positive). For linear model this is guaranteed.

- If all eigs of are large and positive, estimate $\hat{\theta}$ is a sharp minimum of likelihood. This indicates that the estimate $\hat{\theta}$ is well-defined by data.

- If one (or more) eigenvalues are close to zero, estimate $\hat{\theta}$ is sitting at a minimum that is not sharp. This indicates that $\hat{\theta}$ is ill-defined by data and exhibits high variability.

- If one eigenvalue of $\mathcal{H}(\hat{\theta})$ is exactly zero, estimate $\hat{\theta}$ cannot be obtained uniquely from data. This indicates that estimate is sitting at a minimum that is flat. This also indicates that estimate has infinite variability (infinite uncertainty).

Data Sufficiency and Overfitting

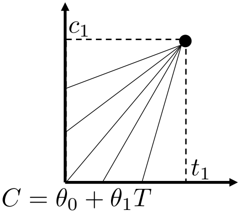
- Can improve precision by controlling $\mathbf{X}$ and with this control FI $\mathcal{I}(\theta) = \mathbf{X}^T \Sigma^{-1} \mathbf{X}$.
- Can be done by increasing amount of data; however, we also require *quality* of data.
- Data points that are *redundant* do not provide new information.
- Redundancy results in $\mathbf{X}$ that has dependent rows and to FI that has small eigs.
- If input vars X are weakly related to output Y, estimates $\hat{\theta}$ will be imprecise.
- Knowledge of application to select appropriate input vars is extremely important.
- Lack of data results in overfitting (need to validate that model predicts well for data not used for estimation). This procedure is known as *cross-validation*.
- Can design $\mathbf{X}$ to maximize measure of $\mathcal{I}(\theta)$ (e.g., sum of eigs)
- Systematic selection of input variables and data is part of *design of experiments*.

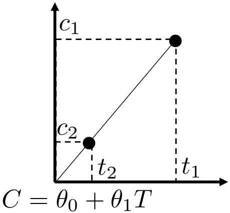
Example: Estimating Heat Capacity with Insufficient Data

- Want model of heat capacity C as a function of temperature T:

$$C = \theta_0 + \theta_1 T$$

- Assume we have one experimental data point (t_1, c_1) to estimate θ_0, θ_1.
- Can θ_0, θ_1 be estimated *uniquely* from (t_1, c_1)?
- There are multiple (infinite) combinations of (θ_0, θ_1) that can fit single data point.
- If we were to estimate these parameters using MLE, model will overfit data.





Example: Estimating Heat Capacity with Insufficient Data

- Consider SSE minimization problem:

$$\min_{\theta_0,\theta_1}\ SSE(\theta_0,\theta_1) = \frac{1}{2}(c_1-\theta_0-\theta_1 t_1)^2.$$

- First-order derivatives of $SSE(\theta)$ are:

$$\frac{\partial SSE(\theta)}{\partial \theta_0} = -(c_1-\theta_0-\theta_1 t_1)$$
$$\frac{\partial SSE(\theta)}{\partial \theta_1} = -t_1(c_1-\theta_0-\theta_1 t_1).$$

- Second derivatives of $SSE(\theta)$ are:

$$\frac{\partial^2 SSE(\theta)}{\partial \theta_0 \partial \theta_0} = 1, \quad \frac{\partial^2 SSE(\theta)}{\partial \theta_0 \partial \theta_1} = t_1$$
$$\frac{\partial^2 SSE(\theta)}{\partial \theta_1 \partial \theta_0} = t_1, \quad \frac{\partial^2 SSE(\theta)}{\partial \theta_1 \partial \theta_1} = t_1^2.$$

- We thus have that Hessian matrix is:

$$\mathcal{H}(\hat{\theta}) = \left[\begin{array}{cc} 1 & t_1 \\ t_1 & t_1^2 \end{array}\right]$$

Example: Estimating Heat Capacity with Insufficient Data

- Since model is linear, can also obtain Hessian by noticing that $\mathcal{H}(\hat{\theta}) = \mathbf{X}^T\mathbf{X}$ with:

$$\mathbf{X} = [1 \;\; t_1]$$

and thus

$$\mathbf{X}^T\mathbf{X} = \left[\begin{array}{c} 1 \\ t_1 \end{array}\right] [1 \;\; t_1] = \left[\begin{array}{cc} 1 & t_1 \\ t_1 & t_1^2 \end{array}\right].$$

- Determinant of Hessian is:

$$|\mathcal{H}(\hat{\theta})| = t_1^2 - t_1^2 = 0$$

- This tells us that one of the eigenvalues is zero and thus θ_0, θ_1 cannot be estimated uniquely from data.

Estimation for General Nonlinear Models

- We now extend discussion to general models of the form:

  $$y_\omega = m(\theta, \mathbf{x}_\omega) + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

  where $m : \mathbb{R}^n \times \mathbb{R}^S \to \mathbb{R}$ is the model (a function of parameters and inputs).
- Model captures structural relationships between inputs, params, and outputs.
- For instance, this can be a reactor model and θ are kinetic constraints.
- Model also captures advanced empirical models such as neural nets (more later).

Estimation for General Nonlinear Models

- We express model in compact form $\mathbf{y} = \mathbf{m}(\theta, \mathbf{X}) + \mathbf{e}$ and assume $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$.
- We assume θ and $\mathbf{X}$ are deterministic vars and thus $\mathbf{y} \sim \mathcal{N}(\mathbf{m}(\theta), \Sigma)$.
- We use MLE framework to estimate θ by minimizing $-\log f(\mathbf{y}|\theta)$:

$$\min_{\theta} - \log\left((2\pi)^{S/2}|\Sigma|^{1/2}\right) - \frac{1}{2}(\mathbf{y} - \mathbf{m}(\theta))^T\Sigma^{-1}(\mathbf{y} - \mathbf{m}(\theta)).$$

- As in linear case, if Σ is diagonal with entries $\Sigma_{\omega,\omega} = \sigma^2$, MLE problem reduces to: SSE minimization problem.

Estimation for General Nonlinear Models

- Solution $\hat{\theta}$ must satisfy necessary conditions (score equations):

$$-\nabla_\theta \log f(\mathbf{y}|\theta) = -\nabla_\theta \mathbf{m}(\theta)^T \Sigma^{-1}(\mathbf{y} - \mathbf{m}(\theta)) = 0,$$

- Score equations are a set of n *nonlinear* equations.
- Here, $\nabla_\theta \mathbf{m}(\theta) \in \mathbb{R}^{S \times n}$ is *Jacobian* matrix and we define this as $\mathcal{J}(\theta) = \nabla_\theta \mathbf{m}(\theta)$.
- Jacobian contains partial derivatives of model with respect to every parameter:

$$\mathcal{J}(\theta)_{\omega,j} = \frac{\partial m(\theta, \mathbf{x}_\omega)}{\partial \theta_j},\ \omega = 1, ..., S,\ \ j = 1, ..., n.$$

- If model is linear, $\mathcal{J}(\theta) = \mathbf{X}$ and score equations reduce to form previously discussed.
- Nonlinear score equations do not have an explicit solution (as in linear case) and require numerical techniques (e.g., Newton's method).
- Jacobian is also difficult to compute analytically and one must resort to numerical techniques (e.g., automatic differentiation).

Estimation for General Nonlinear Models

- To ensure that $\hat{\theta}$ minimizes $-\log f(\mathbf{y}|\theta)$, we must ensure $\mathcal{H}(\hat{\theta}) \succ 0$ holds.
- Here, $\mathcal{H}(\hat{\theta}) = -\nabla_\theta^2 \log f(\mathbf{y}|\hat{\theta}) \in \mathbb{R}^{n \times n}$.
- Hessian is matrix of second derivatives of log-likelihood function:

$$\mathcal{H}(\theta)_{j,j} = -\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial\theta_i \partial\theta_j},\ i = 1, ..., n,\ j = 1, ..., n.$$

- Hessian is a complex quantity that requires numerical techniques to be computed.
- When residuals $\mathbf{y} - \mathbf{m}(\theta)$ are small, Hessian can be approximated as:

$$\mathcal{H}(\hat{\theta}) \approx \nabla_\theta \mathbf{m}(\hat{\theta})^T \Sigma^{-1} \nabla_\theta \mathbf{m}(\hat{\theta}).$$

- This approximation is known as Gauss-Newton approximation.
- For linear models we have seen that Hessian is given by:

$$\mathcal{H}(\hat{\theta}) = \mathbf{X}^T \Sigma^{-1} \mathbf{X}.$$

- Gauss-Newton approximation is thus exact when model is linear.

Estimation for General Nonlinear Models

What is different about nonlinear estimation (compared to linear case)?

- It is significantly more difficult to obtain a pdf for $\hat{\theta}$.
- In linear case, this was obtained by noticing that estimate is a linear transformation of noise (this is no longer true in nonlinear case).
- Typically, in nonlinear case, pdf is approximated as:

$$\hat{\theta} \sim \mathcal{N}(\theta, \mathcal{H}(\hat{\theta})^{-1})$$

- This approximation is accurate if nonlinearity of $\mathbf{m}(\theta)$ is not too strong.
- Here, we make assumption that $\mathrm{Cov}[\hat{\theta}] = \mathcal{H}(\hat{\theta})^{-1}$.
- Hessian can be difficult to compute but modern modeling languages can compute this automatically (e.g., JuMP, Pyomo, Casadi).
- Surprisingly, most numerical packages do not provide Hessians. In such cases, one can use Gauss-Newton approximation.

Estimation for General Nonlinear Models

- Recall $\mathrm{Cov}[\hat{\theta}]$ tells us how $\hat{\theta}$ varies as we vary noise $\mathbf{e}$.
- We can thus obtain covariance by perturbing output data $\mathbf{y}$ with noise and by obtaining corresponding estimates.
- Consider realizations $\mathbf{e}_k,\ k = 1, ..., K$ (from $\mathcal{N}(0, \Sigma)$) and perturb data $\mathbf{y}_k = \mathbf{y} + \mathbf{e}_k$.
- We solve MLE problems for $k = 1, ..., K$:

$$\hat{\theta}_k \in \underset{\theta}{\mathrm{argmin}}\ -\log f(\mathbf{y}_k|\theta)$$

- This gives realizations for estimates $\hat{\theta}_k,\ k = 1, ..., K$ and from this we compute:

$$\hat{\mathbb{E}}[\hat{\theta}] = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k.$$

  and:

$$\hat{\mathrm{Cov}}[\hat{\theta}] = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\mathbb{E}}[\hat{\theta}])(\hat{\theta}_k - \hat{\mathbb{E}}[\hat{\theta}])^T.$$

- This is a Monte Carlo simulation approach that is easy to implement.

Example: Hougen-Watson `ch5_hougen_watson.m`

- We are interested in finding parameters for Hougen-Watson (HW) model:

$$Y = \frac{(\theta_0 X_2 - X_3/\theta_4)}{(1 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)}$$

- Model captures competitive reaction and adsorption rates in a heterogeneous catalytic reaction (gas components adsorb into solid catalyst).

- In this model, Y is reaction rate and (X_1, X_2, X_3) are species partial pressures.

- Parameters $(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ have physical meaning (these represent adsorption and reaction rates and thus need to be non-negative).

Example: Hougen-Watson `ch5_hougen_watson.m`

- We have $S = 13$ observations (y_ω, x_ω) available and $n = 5$ parameters to estimate.
- We are not given info about noise model and we thus solve SSE problem:

$$\min_{\theta} \frac{1}{2}\|\mathbf{y} - \mathbf{m}(\theta)\|^2$$

where:

$$m_\omega(\theta) = \frac{(\theta_0 x_{2,\omega} - x_{3,\omega}/\theta_4)}{(1 + \theta_1 x_{1,\omega} + \theta_2 x_{2,\omega} + \theta_3 x_{3,\omega})}.$$

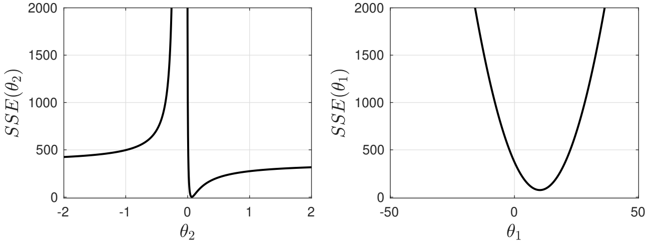
Example: Hougen-Watson `ch5_hougen_watson.m`

- HW model is highly nonlinear and this results in high nonlinearity of $SSE(\theta)$.

- Below we show behavior of $SSE(\theta)$ for different values of $\theta_1 \in [-2, 2]$ and $\theta_2 \in [-2, 2]$ (while keeping rest of parameters fixed).

- In a nonlinear model, we can find emergence of max and min points in $SSE(\theta)$ and complex behavior (e.g., SSE function "blows up" at specific parameter values).

- This behavior differs from that observed in linear model (SSE function only has minimum points) and makes nonlinear problems difficult to solve.



Example: Hougen-Watson `ch5_hougen_watson.m`

- We have intentionally evaluated SSE function at negative values of parameters.
- We know from physical understanding that these parameters cannot be negative.
- Unless this knowledge is explicitly encoded in SSE problem, we run risk that numerical solver lands at minimum point that has non-physical values.
- Used `Matlab` function `fitnlm`, which provides estimates:

$$\begin{aligned}\hat{\theta}_0 &= 1.25\\ \hat{\theta}_1 &= 0.06\\ \hat{\theta}_2 &= 0.04\\ \hat{\theta}_3 &= 0.11\\ \hat{\theta}_4 &= 1.19.\end{aligned}$$

- Minimum SSE value found was $SSE(\hat{\theta}) = 0.2989$; we use initial guess $\hat{\theta} = (1, 1, 1, 1, 1)$ to initialize the Newton algorithm used by `fitnlm`.

Example: Hougen-Watson `ch5_hougen_watson.m`

- We now solve problem by using initial guess $\hat{\theta} = (-1, -1, -1, -1, -1)$ and find that Newton algorithm does not converge after 200 iterations and delivers estimates:

$$\hat{\theta}_0 = -73.60$$
$$\hat{\theta}_1 = -3.70$$
$$\hat{\theta}_2 = -2.62$$
$$\hat{\theta}_3 = -6.36$$
$$\hat{\theta}_4 = -0.02$$

  which are clearly non-physical.

- Interestingly, these estimates give $SSE(\hat{\theta}) = 0.3834$, which is not too far from value found with the other set of estimates.

- This shows how numerical solvers might encounter estimates that give excellent fits (minimize the SSE function) but that are non-physical.

Example: Hougen-Watson `ch5_hougen_watson.m`

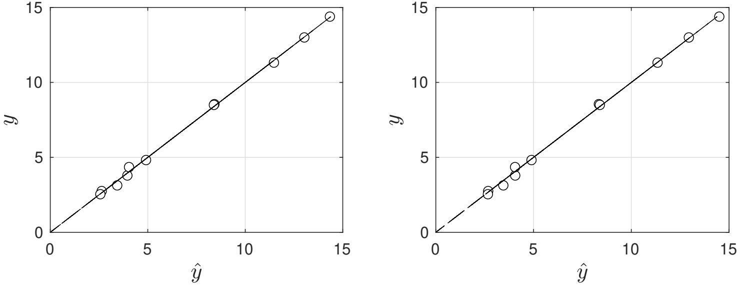


Figure: Fit of HW model to exp data with physical (left) and non-physical (right) parameters.

Prior Knowledge

- We often have some idea of what actual param values should be or if they must respect some general rules.

- For instance, our physical knowledge might tell us that a param cannot be negative or that it cannot be above a certain threshold.

- We are interested in incorporating this *prior* knowledge in our estimation procedure.

- Prior knowledge help us deal with situations in which the available data is insufficient to estimate params uniquely (e.g., Hessian has zero eigs) or reliably (e.g., Hessian has eigs that are close to zero).

- Prior knowledge helps us eliminate spurious estimates $\hat{\theta}$ with no physical meaning.

- Prior knowledge helps us narrow down space over we search on.

- Reducing number of parameters helps us address overfitting.

- Narrowing search space helps numerical solvers find estimates.

Prior Knowledge

- Can incorporate prior knowledge by adding bounds on parameters:

$$\begin{aligned} \min_{\theta} \quad & -\log f(\mathbf{y}|\theta) \\ \text{s.t.} \quad & \underline{\theta} \le \theta \le \overline{\theta} \end{aligned}$$

- Can add general constraints (e.g., sum of params must be equal to some value):

$$\begin{aligned} \min_{\theta} \quad & -\log f(\mathbf{y}|\theta) \\ \text{s.t.} \quad & \underline{\theta} \le \Pi\theta \le \overline{\theta} \end{aligned}$$

- Can add a *penalty* term to log-likelihood to control parameter behavior:

$$\min_{\theta} \; -\log f(\mathbf{y}|\theta) + \kappa \cdot \rho(\theta).$$

- Here, $\kappa \ge 0$ is a constant that trades-off model fit and allowed parameter movement.

- Balances *prior knowledge* induced by $\rho(\theta)$ with *new knowledge* contained in data.

Prior Knowledge

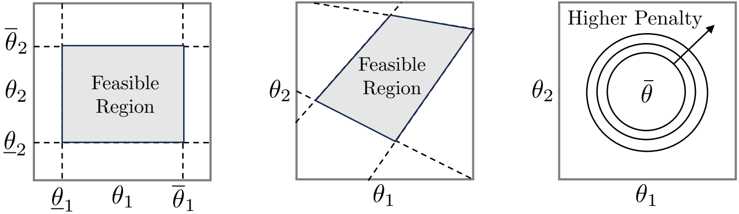


Figure: Feasible region defined by bounds (left) and general constraints (middle). Penalization induced by ridge penalty (right).

Prior Knowledge

Common choices for penalty function $\rho(\theta)$ are:

- ℓ-2 norm (a.k.a ridge or Tikhonov penalty):

$$\begin{aligned}\rho(\theta) &= \frac{1}{2}\|\theta - \bar{\theta}\|_2^2 \\ &= \frac{1}{2}\sum_{i=1}^{n}(\theta_i - \bar{\theta}_i)^2 \\ &= \frac{1}{2}(\theta - \bar{\theta})^T(\theta - \bar{\theta}).\end{aligned}$$

- ℓ-1 norm (a.k.a. lasso penalty):

$$\rho(\theta) = \|\theta - \bar{\theta}\|_1 = \sum_{i=1}^{n}|\theta_i - \bar{\theta}_i|.$$

- Bayes penalty function:

$$\rho(\theta) = \frac{1}{2}(\theta - \bar{\theta})^T\Sigma_\theta^{-1}(\theta - \bar{\theta})$$

  Similar to ridge penalty but matrix Σ_θ induces different weights on params.

Prior Knowledge

- Estimate obtained from a penalized problem depends on the value of κ chosen and thus we have a range of solutions (for different values of κ) that we denote as $\hat{\theta}(\kappa)$.

- As a result, one often forms a *trade-off* curve $-\log f(\mathbf{y}|\hat{\theta}(\kappa))$ vs. $\rho(\hat{\theta}(\kappa))$ and one selects the point that has the best trade-off.

- The smaller κ, the smaller $-\log f(\mathbf{y}|\hat{\theta}(\kappa))$ and the larger the $\rho(\hat{\theta}(\kappa))$.

- One can also obtain trade-off curve by solving constrained MLE problem:

$$\begin{aligned} \min_{\theta} \; & -\log f(\mathbf{y}|\theta) \\ \text{s.t. } & \rho(\theta) \leq \kappa \end{aligned}$$

- Here, κ is now interpreted as the allowed threshold for the penalty function $\rho(\theta)$.

- From this we can see that one can express penalty functions as constraints.

Prior Knowledge

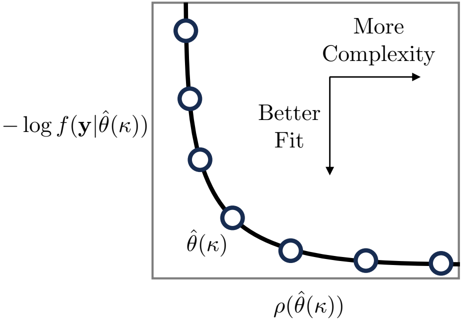


Figure: Trade-off between model flexibility and model fit.

Example: Hougen-Watson Regularized `ch5_hougen_watson_regularized.m`

- We augment SSE using a ridge regularizer:

$$SSE(\theta) + \frac{1}{2}||\theta - \bar{\theta}||_2^2.$$

where $\kappa = 1$ and $\bar{\theta} = 0$.

- We can see that regularizer helps better define minimum points of SSE unction but does not avoid presence of non-physical regions.

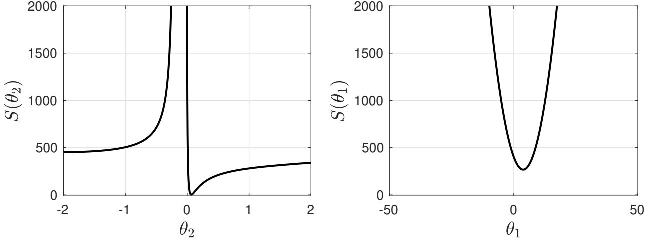


Figure: Behavior of regularized SSE function as we span values for θ_1 and θ_2.

Example: Hougen-Watson Constrained `ch5_hougen_watson_constraints.m`

- Consider now constrained MLE problem:

$$\min_{\theta} \frac{1}{2}\|\mathbf{y} - \mathbf{m}(\theta)\|^2$$
$$\text{s.t. } \underline{\theta} \leq \theta \leq \overline{\theta}$$

  with $\underline{\theta} = (0,0,0,0,0)$ and $\overline{\theta} = (2,2,2,2,2)$.

- Constrained MLE problem solved using a numerical package that implements a Newton-like algorithm.
- Routines uses Newton algorithm to find parameters but can also handle constraints.
- Solve problem with guess $\hat{\theta} = (-1,-1,-1,-1,-1)$ and without using constraints.
- We find that algorithm does not converge and obtains estimates that are negative.
- Upon incorporating constraints we find:

$$\hat{\theta} = (1.25, 0.06, 0.04, 0.11, 1.19).$$

Example: Hougen-Watson Constrained `ch5_hougen_watson_constraints.m`

- The numerical package does not provide Hessian info but provides Jacobian $\nabla_\theta \mathbf{m}(\hat{\theta})$.
- We approximate Hessian using Gauss-Newton approximation.
- Info used to determine if estimated params are unique and to obtain covariance:

$$\mathcal{H}(\hat{\theta}) = \begin{bmatrix} 619.12 & -4305.49 & -6345.59 & -940.39 & 99.98 \\ -4305.49 & 37792.30 & 44531.04 & 4721.95 & -434.18 \\ -6345.59 & 44531.04 & 75036.07 & 9325.14 & -663.07 \\ -940.39 & 4721.95 & 9325.14 & 2149.56 & -193.73 \\ 99.98 & -434.18 & -663.07 & -193.73 & 39.38 \end{bmatrix}$$

- Eigs of Hessian are all positive:

$$\lambda = (1.06 \times 10^5, 8.35 \times 10^3, 9.52 \times 10^2, 3.53 \times 10^1, 2.55 \times 10^{-2}).$$

- This indicates that $\hat{\theta}$ is sitting at a minimum point of the SSE function.
- Estimate variances are:

$$\mathbb{V}[\hat{\theta}] = (7.63 \times 10^{-1}, 1.92 \times 10^{-3}, 9.69 \times 10^{-4}, 5.73 \times 10^{-3}, 7.10 \times 10^{-1}).$$

Bayesian Estimation

- We have used MLE as main paradigm to determine parameters from data.
- Idea is to find the estimate that maximizes conditional pdf $f(\mathbf{y}|\mathbf{X}, \theta, \Sigma)$.
- Approach implicitly assumes that $\mathbf{X}$, θ, and Σ are deterministic variables.
- This is limiting because input data $\mathbf{X}$ might contain uncertainty (i.e., sensors) or because we want to convey prior knowledge on $\mathbf{X}, \theta, \Sigma$ in statistical terms.

Bayesian Estimation

- The is a more general estimation paradigm known as Bayes estimation that allows us to handle overcome the limitations of MLE.
- Key difference is that $\theta, \mathbf{X}$ and Σ are treated as RVs.
- For simplicity, we will focus on θ (generalizing arguments is relatively easy).
- In context of estimation problem of interest, Bayes' theorem states that:

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}$$

Bayesian Estimation

- $f(\theta|\mathbf{y})$ is prob that parameters take value θ given knowledge that output takes value $\mathbf{y}$. This pdf is known as posterior pdf.

- $f(\theta)$ is prob that parameters take value θ given no knowledge of output. This pdf is known as the prior pdf.

- $f(\mathbf{y}|\theta)$ is prob that output takes value $\mathbf{y}$ given knowledge that parameters take value θ. This pdf is likelihood function maximized in MLE.

- $f(\mathbf{y})$ is marginal probability of outputs; this pdf is often neglected as it does not carry knowledge on parameters and thus one often expresses Bayes theorem as:

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta)$$

Bayesian Estimation

- Bayesian estimation maximizes prob of parameters $f(\theta|\mathbf{y})$.
- This makes more sense than MLE (maximizes probability of outputs $f(\mathbf{y}|\theta)$).
- We thus find Bayes estimate $\hat{\theta}$ by solving:

$$\max_{\theta} \quad \log f(\mathbf{y}|\theta) + \log f(\theta)$$

- Prior term $\log f(\theta)$ naturally carries prior knowledge.

Bayesian Estimation

- Consider now model $\mathbf{y} = \mathbf{m}(\theta) + \mathbf{e}$ with $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$.

- Moreover, consider prior knowledge that $\theta \sim \mathcal{N}(\bar{\theta}, \Sigma_\theta)$.

- This indicates that:

$$\log f(\theta) = \log\left((2\pi)^{n/2}|\Sigma_\theta|^{1/2}\right) + \frac{1}{2}(\theta - \bar{\theta})^T\Sigma_\theta^{-1}(\theta - \bar{\theta})$$
$$\log f(\mathbf{y}|\theta) = \log\left((2\pi)^{S/2}|\Sigma|^{1/2}\right) + \frac{1}{2}(\mathbf{y} - \mathbf{m}(\theta))^T\Sigma^{-1}(\mathbf{y} - \mathbf{m}(\theta)).$$

- After dropping constant terms and changing from max to min, we obtain:

$$\min_{\theta} \quad \frac{1}{2}(\mathbf{y} - \mathbf{m}(\theta))^T\Sigma^{-1}(\mathbf{y} - \mathbf{m}(\theta)) + \frac{1}{2}(\theta - \bar{\theta})^T\Sigma_\theta^{-1}(\theta - \bar{\theta}).$$

Bayesian Estimation

- Gaussian prior $f(\theta)$ achieves its maximum at reference point $\bar{\theta}$ while $f(\mathbf{y}|\theta)$ achieves its maximum when θ fits perfectly the data.
- Bayes estimation naturally seeks to strike a *balance* (trade-off) between what we previously knew about θ and the new knowledge gained through the data of $\mathbf{y}$.
- If we ignore prior knowledge, all we know about θ is through data $\mathbf{y}$ (as in MLE).
- This might lead to ambiguity if data is insufficient to estimate parameters uniquely.
- Consequently, adding prior knowledge seeks to reduce *ambiguity*.
- Reference $\bar{\theta}$ is prior estimate (best estimate prior to observing new data $\mathbf{y}$).
- How much we trust our prior estimate is dictated by covariance matrix Σ_{θ}.

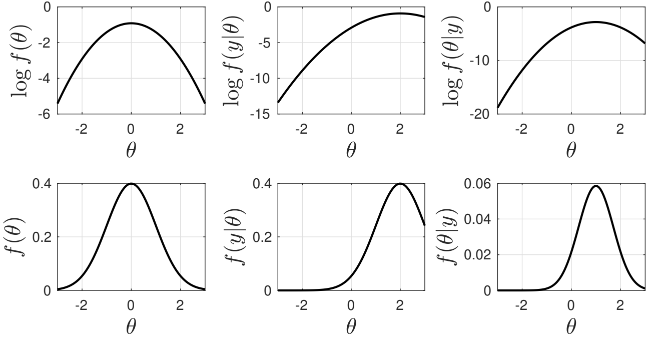
Example: Bayesian Estimation `ch5_bayes_estimation.m`

- Consider simple linear model $Y = \theta X + \epsilon$ and assume that we have a single observation that we denote as $(x = 1, y = 2)$.

- Prior knowledge is $\theta \sim \mathcal{N}(0, 1)$ and this defines marginal $f(\theta)$.

- Info provided by data (x, y) is $y \sim \mathcal{N}(\theta x, 1)$ and this defines likelihood $f(y|\theta)$.

- Combined information given by posterior pdf $f(\theta|y) \propto f(y|\theta)f(\theta)$.



Example: Bayesian Estimation ch5_bayes_estimation.m

- Location of maximum points for pdfs and log pdfs coincides.

- According to prior information (embedded in$f(\theta)$), we have that $\theta = 0$ is the most likely value.

- According to new information (embedded in $f(y|\theta)$), we have that $\theta = 2$ is the most likely value. This would be estimate obtained with MLE (ignoring prior information).

- According to combined prior and new information (embedded in $f(\theta|y)$), we have that $\theta = 1$ is the most likely value.

- Bayesian estimation naturally "blends" old and new information and finds an estimate that is a trade-off point.

Bayesian Estimation

- Bayes framework can be naturally capture prior info on input data $\mathbf{X}$.

- Assume uncertainty in input data $\mathbf{X} \sim \mathcal{N}(\bar{\mathbf{X}}, \Sigma_X)$; here, the reference point (mean) $\bar{\mathbf{X}}$ is observed value and Σ_X captures their uncertainty (e.g., precision of sensor).

- Given model $\mathbf{y} = \mathbf{m}(\theta, \mathbf{X})$, we *simultaneously* estimate params θ and inputs $\mathbf{X}$.

- Bayes' theorem tells us that:

$$f(\theta, \mathbf{X}|\mathbf{y}) \propto f(\mathbf{y}|\theta, \mathbf{X}) f(\theta, \mathbf{X})$$

- If θ and $\mathbf{X}$ are independent $f(\theta, \mathbf{X}) = f(\theta) f(\mathbf{X})$.

- Our estimates can thus be obtained by solving problem:

$$\max_{\theta, \mathbf{X}} \; \log f(\mathbf{y}|\theta, \mathbf{X}) + \log f(\theta) + \log f(\mathbf{X}).$$

- Bayes estimation problem can be expressed as:

$$\min_{\theta, \mathbf{X}} \quad \frac{1}{2}(\mathbf{y} - \mathbf{m}(\theta))^T \Sigma^{-1} (\mathbf{y} - \mathbf{m}(\theta)) + \frac{1}{2}(\theta - \bar{\theta})^T \Sigma_\theta^{-1} (\theta - \bar{\theta}) + \frac{1}{2}(\mathbf{X} - \bar{\mathbf{X}})^T \Sigma_X^{-1} (\mathbf{X} - \bar{\mathbf{X}}).$$

- Note that this approach is equivalent to defining $\mathbf{X}$ as parameters.

Example: Data Reconciliation

- We consider a system with output flow Y and input flow X.
- There is no transformation in the system and we thus know have that conservation $Y = X$ must hold.
- We have a single input-output observation (measurement) that we denote as (x_ω, y_ω). This input and output measurements are subject to Gaussian errors with variances σ_x^2 and σ_y^2.
- Because of sensor errors in input and output flow, observed input and output do not satisfy conservation (i.e., $x_\omega \neq y_\omega$).
- We use Bayesian estimation to obtain estimate of input flow $\hat{x}$ that reconciles observed input and output data to conservation equation (model).

Example: Data Reconciliation

- Bayes' theorem tells us that $f(x|y) \propto f(y|x)f(x)$.

- We obtain $\hat{x}$ by finding value x that maximizes $f(x|y)$.

- Bayes estimation problem seeks to maximize $f(x|y)$ and can be expressed as:

$$\min_{x} \quad \frac{1}{2\sigma_y^2}(y_\omega - x)^2 + \frac{1}{2\sigma_x^2}(x - x_\omega)^2.$$

- Estimate $\hat{x}$ seeks to minimize error from conservation and minimize deviation from reference value x_ω (from observation).

- How much weight is put on conservation and correction is controlled by σ_y^2 and σ_x^2.

Example: Data Reconciliation

- To find solution, we take derivative of objective function with respect to x:

$$-\frac{1}{\sigma_y^2}(y_\omega - x) + \frac{1}{\sigma_x^2}(x - x_\omega) = 0$$

- From here we can determine that the estimate $\hat{x}$ is:

$$\hat{x} = \frac{\frac{1}{\sigma_y^2} y_\omega + \frac{1}{\sigma_x^2} x_\omega}{\frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2}}.$$

- This can also be written as:

$$\hat{x} = \frac{y_\omega + \frac{\sigma_y^2}{\sigma_x^2} x_\omega}{1 + \frac{\sigma_y^2}{\sigma_x^2}}.$$

- From this expression it is clear that, if we trust output measurement much more than input measurement then $\sigma_y^2/\sigma_x^2 \to 0$ and thus $\hat{x} \to y_\omega$.

- On other other hand, if we trust input measurement much more than output measurement then $\sigma_y^2/\sigma_x^2 \to \infty$ and thus $\hat{x} \to x_\omega$.