

Lecture 4: Frequentist and Bayesian Estimators

Chapter 2.3

Charles A. Bouman

Purdue University

Monday, August 31, 2026

① **Frequentist estimation**

Bias, variance, and MSE; the ML estimator, its properties, and examples.

② **Bayesian estimation**

Cost functions and Bayes' risk; the MMSE, conditional median, and MAP estimators; Gaussian examples.

③ **The two compared**

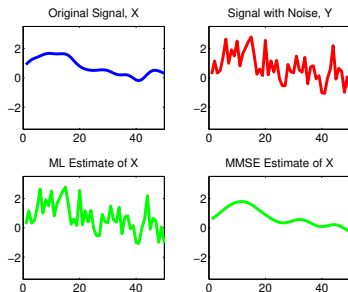
ML versus MMSE on the same data, and the bias–variance trade-off.

Reading: Section 2.3 of *Foundations of Computational Imaging*.

Today's Big Ideas

- In the Frequentist framework the unknown is a deterministic parameter, and an estimator is judged by its bias and its variance.
- The ML estimator maximizes the density of the observed data, and it is the right choice when the data far outnumber the unknowns.
- In the Bayesian framework the unknown is random with a prior distribution, and an estimator is judged by its Bayes' risk.
- The cost function picks the estimator: squared error gives the MMSE estimate, absolute error the conditional median, and zero-one cost the MAP estimate.
- For jointly Gaussian unknowns the three Bayesian estimators agree, and the estimate is an attenuated, biased version of the data.
- The MAP estimate minimizes a data term plus a prior term, the structure used throughout the course.

Two estimates from the same data



- Both estimates come from the same noisy measurements of the same signal.
- One reproduces the noise. The other recovers the smooth structure.
- By the end of class you will be able to derive both, and to say what the better one gave up.

Two frameworks for estimating from data

There are two general frameworks for estimating an unknown from observed data.

Notation used throughout the book

Frequentist: $\theta \in \Omega$ is an unknown *deterministic* parameter.

Bayesian: X is an unknown *random* vector.

- A scanner's calibration constant is typically a θ .
- The image being reconstructed is typically an X .
- Both approaches are valuable. The question is which combination fits the application.

Frequentist estimation: bias, variance, and MSE

An **estimator** is a function $T : \mathbb{R}^p \rightarrow \Omega$ of the observed data $Y \in \mathbb{R}^p$, whose density is $p_\theta(y)$:

$$\hat{\theta} = T(Y) .$$

Since $T(Y)$ is a function of Y , the estimate is itself random, with mean

$$\bar{\theta} = \mathbb{E}_\theta[\hat{\theta}] = \int_{\mathbb{R}^n} T(y) p_\theta(y) dy .$$

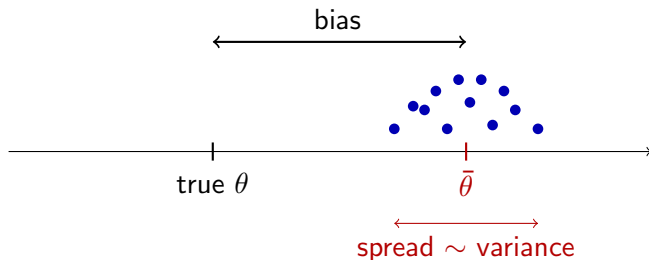
$$\text{bias}_\theta = \bar{\theta} - \theta$$

$$\text{var}_\theta = \mathbb{E}_\theta\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right]$$

$$\text{MSE}_\theta = \mathbb{E}_\theta\left[\left(\hat{\theta} - \theta\right)^2\right] = \text{var}_\theta + (\text{bias}_\theta)^2 .$$

Bias and variance, pictured

Repeat the whole experiment many times. Each repetition gives one estimate $\hat{\theta}$, so we get a cloud of values.



- **Bias** is how far the center of the cloud sits from the true value.
- **Variance** is how wide the cloud is.
- MSE combines the two. But bias is usually considered less desirable than variance, so MSE is often not a good metric of quality.

Comparing estimators is not straightforward

Bias, variance, and MSE all depend on the value of θ , so it is often unclear which estimator is better.

- Even a poor estimator can have zero error for some θ . The estimator fixed at $\hat{\theta} = 1$, which ignores the data, has MSE zero when $\theta = 1$.

Unbiased and UMVU

An **unbiased estimator** has zero bias for every θ , so $\bar{\theta} = \theta$.

A **uniformly minimum variance unbiased (UMVU)** estimator is an unbiased estimator whose variance is less than or equal to that of every other unbiased estimator, for every θ .

Consistent estimator

Let $\hat{\theta}_n$ be computed from n independent samples. The estimator is **consistent** if it converges to the true value in probability. That is, for all $\theta \in \Omega$ and all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P_{\theta}\{|\theta - \hat{\theta}_n| > \epsilon\} = 0 .$$

- Without consistency, even unlimited data does not guarantee that the estimate approaches the true value.
- We will see later that some very intuitive estimators are not consistent.

The maximum likelihood (ML) estimator

The **ML estimator** is the value of θ that maximizes the density of the observed data:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta \in \Omega} p_{\theta}(Y) \\ &= \arg \max_{\theta \in \Omega} \log p_{\theta}(Y) ,\end{aligned}$$

where $\arg \max$ is the argument that achieves the global maximum.

- The argument of the density is the random vector Y , so $\hat{\theta}$ is a random variable.
- When $p_{\theta}(y)$ is continuously differentiable in θ , a necessary condition is

$$\nabla_{\theta} \log p_{\theta}(Y)|_{\theta=\hat{\theta}} = 0 .$$

Properties of the ML estimate

- The ML estimate is generally *not* unbiased.
- Under mild conditions it is **consistent** and **asymptotically efficient**.
- Asymptotically efficient means it attains the **Cramér–Rao bound** as the number of independent samples n goes to infinity.
- The Cramér–Rao bound is a lower bound on the variance of an unbiased estimator. So ML is about as good as an unbiased estimator can be.

When to use ML

- When there is much more data than unknowns,

$$\frac{\text{amount of data}}{\text{number of unknowns}} \gg 1 .$$

- When you have no idea which values of θ occur and need good performance in every case.

Example 2.3: ML estimate of a Gaussian mean

Let $\{Y_k\}_{k=1}^n$ be i.i.d. $N(\theta, \sigma^2)$ with unknown mean θ . Then

$$\log p_{\theta}(Y) = -\frac{1}{2\sigma^2} \sum_{k=1}^n (Y_k - \theta)^2 - \frac{n}{2} \log(2\pi\sigma^2) .$$

Differentiating the log likelihood,

$$\left. \frac{d \log p_{\theta}(Y)}{d\theta} \right|_{\theta=\hat{\theta}} = \frac{1}{\sigma^2} \sum_{k=1}^n (Y_k - \hat{\theta}) = 0 \quad \implies \quad \hat{\theta} = \frac{1}{n} \sum_{k=1}^n Y_k .$$

This estimate is unbiased:

$$\mathbb{E}_{\theta}[\hat{\theta}] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}_{\theta}[Y_k] = \theta .$$

Here the ML estimator is also the UMVU estimator.

Example 2.4: ML estimate of a multivariate Gaussian

Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, R)$ with $\mu \in \mathbb{R}^p$, and let $\theta = [\mu, R]$. The sufficient statistics are

$$b = \sum_{k=1}^n x_k, \quad S = \sum_{k=1}^n x_k x_k^t.$$

In terms of b and S , the density of X is

$$p(x|\theta) = \frac{1}{(2\pi)^{np/2}} |R|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \{ S R^{-1} \} + b^t R^{-1} \mu - \frac{n}{2} \mu^t R^{-1} \mu \right\}.$$

Maximizing over θ (see the appendix) gives the ML estimates in terms of b and S . We *define* $\hat{\mu}$ and $\hat{R}$ to be these expressions:

$$\hat{\mu} \triangleq \frac{b}{n}, \quad \hat{R} \triangleq \frac{S}{n} - \hat{\mu} \hat{\mu}^t.$$

So the ML estimates $\hat{\mu}$ and $\hat{R}$ are the sample mean and the sample covariance of the observations.

Example 2.5: ML estimate of a Bernoulli parameter

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with parameter θ , and let N_1 count the X_k equal to one. Then

$$\log p(x|\theta) = (n - N_1) \log(1 - \theta) + N_1 \log \theta ,$$

and differentiating gives

$$\frac{\partial \log p(x|\theta)}{\partial \theta} = -\frac{n - N_1}{1 - \theta} + \frac{N_1}{\theta} = 0 ,$$

so the ML estimate is

$$\hat{\theta} = \frac{N_1}{n} .$$

The ML estimate is the observed fraction of times that X_k is one, which is what intuition suggests.

Bayesian estimation

In Bayesian estimation the unknown itself is random.

The core assumption

The unknown is a **random** vector X with an assumed prior distribution $p_x(x)$. The estimator is again a function of the observations, $\hat{X} = T(Y)$.

- Bayesian methods aim to beat Frequentist accuracy by assuming which values of the unknown are likely.
- They help most when the data is small compared to the dimension of the unknown.
- The real cost is that Bayesian methods introduce bias. They also require choosing a distribution for X , and a poor choice can badly degrade accuracy.

The posterior distribution

The **posterior distribution** $p_{x|y}(x|y)$ is the distribution of the unknown X given the observation Y . It holds everything the data tells us about X , and it follows from Bayes' rule:

$$p_{x|y}(x|y) = \frac{p_{y|x}(y|x) p_x(x)}{p_y(y)} .$$

- The **forward model** $p_{y|x}(y|x)$ says how the observations relate to the unknown.
- The **prior model** $p_x(x)$ is the distribution assumed for X before any data is observed.

Every Bayesian estimator in this lecture is computed from the posterior. Which one you get depends on the cost you attach to an error.

Cost functions and Bayes' risk

The **cost** $C(x, \hat{x}) \geq 0$ is the penalty for choosing $\hat{x}$ when x is correct. A perfect estimate has $C(x, x) = 0$.

For a given $X = x$, the expected cost is

$$\bar{C}(x) = \mathbb{E}[C(x, T(Y)) | X = x] = \int_{\mathbb{R}^n} C(x, T(y)) p_{Y|X}(y|x) dy .$$

The **Bayes' risk** averages this over X :

$$\text{Risk} = \mathbb{E}[C(X, T(Y))] = \int \bar{C}(x) p_X(x) dx .$$

- The risk is not a function of X , unlike the Frequentist MSE, which was a function of θ .
- The price is that the risk depends heavily on the prior $p_X(x)$.

The cost function determines the estimator

- **Squared error**, $C(x, \hat{x}) = |x - \hat{x}|^2$, gives the conditional expectation,

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y] .$$

- **Absolute error**, $C(x, \hat{x}) = |x - \hat{x}|$, gives the conditional median,

$$\int_{-\infty}^{\hat{x}_{median}} p_{x|y}(x|Y) dx = \frac{1}{2} .$$

- **Zero-one cost**, $C(x, \hat{x}) = 1 - \delta(x - \hat{x})$, which treats tiny and huge errors the same, gives the **MAP** estimate,

$$\hat{X}_{MAP} = \arg \max_{x \in \Omega} p_{x|y}(x|Y) .$$

MAP in terms of the forward and prior models

Substituting Bayes' rule into the MAP estimate gives

$$\begin{aligned}\hat{X}_{MAP} &= \arg \max_{x \in \Omega} \log p_{x|Y}(x|Y) \\ &= \arg \max_{x \in \Omega} \{ \log p_{Y|x}(Y|x) + \log p_x(x) - \log p_Y(Y) \} \\ &= \arg \min_{x \in \Omega} \{ -\log p_{Y|x}(Y|x) - \log p_x(x) \} .\end{aligned}$$

- The forward model comes from the physics of how the measurements are made.
- The prior model captures the variation in the images being estimated.

MAP balances fitting the data against fitting the prior. This two term structure, a data term plus a prior term, recurs through the whole course.

Example 2.6: scalar Gaussian, $Y = X + W$

Let $X \sim N(0, \sigma_x^2)$ and $W \sim N(0, \sigma_w^2)$, with $Y = X + W$. Estimate X from Y .

Instead of evaluating Bayes' rule directly, write

$$p_{x|y}(x|y) = \frac{1}{z_y} p_{y,x}(y, x) ,$$

where the **partition function** z_y is the normalizing constant. It may depend on y but not on x . Then

$$p_{x|y}(x|y) = \frac{1}{z'_y} \exp \left\{ -\frac{1}{2\sigma_w^2} (y - x)^2 - \frac{1}{2\sigma_x^2} x^2 \right\} = \frac{1}{z''_y} \exp \{ a(x - b)^2 \} ,$$

where a and b may depend on y but not on x .

- The exponent is quadratic in x , so the posterior is Gaussian. It remains to identify a and b .

Example 2.6: posterior mean and variance

Matching the first two derivatives of the two logarithms gives

$$a = -\frac{1}{2} \left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2} \right) = -\frac{1}{2\sigma_{x|y}^2} , \quad b = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} \right) y = \mu_{x|y} .$$

So the posterior is Gaussian,

$$p_{x|y}(x|y) = \frac{1}{\sqrt{2\pi\sigma_{x|y}^2}} \exp \left\{ -\frac{1}{2\sigma_{x|y}^2} (x - \mu_{x|y})^2 \right\} ,$$

with

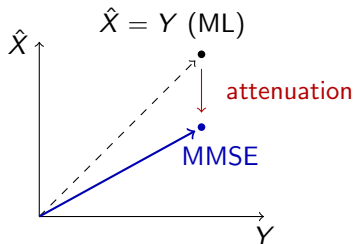
$$\mu_{x|y} = \frac{\sigma_x^2 y}{\sigma_x^2 + \sigma_w^2} , \quad \sigma_{x|y}^2 = \frac{\sigma_x^2 \sigma_w^2}{\sigma_x^2 + \sigma_w^2} .$$

- The posterior mean is proportional to y . The posterior variance does not depend on y at all.

Interpretation: attenuation and bias

The MMSE estimate is

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} Y = \frac{1}{1 + \frac{\sigma_w^2}{\sigma_x^2}} Y .$$



- The Bayesian estimate is Y pulled toward the prior mean of zero. The **noise-to-signal ratio** σ_w^2/σ_x^2 sets how strong the pull is.

Attenuation reduces MSE but biases the estimate:

$$\mathbb{E}[\hat{X}_{MMSE} | X] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2} X < X .$$

Example 2.7: jointly Gaussian random vectors

Let $X \sim N(0, R_x)$ and $W \sim N(0, R_w)$ be independent, and $Y = X + W$. The same derivation goes through with quadratic forms in place of scalar squares, giving a Gaussian posterior with

$$\begin{aligned}\mu_{x|y} &= (R_x^{-1} + R_w^{-1})^{-1} R_w^{-1} y = R_x (R_x + R_w)^{-1} y , \\ R_{x|y} &= (R_x^{-1} + R_w^{-1})^{-1} = R_x (R_x + R_w)^{-1} R_w .\end{aligned}$$

The MMSE estimate is therefore

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y] = R_x (R_x + R_w)^{-1} Y .$$

Example 2.7: properties particular to the Gaussian case

These hold for the Gaussian case, not in general

- ① The posterior mean is a *linear* function of the data y , and the posterior variance does not depend on y .
 - ② The posterior is Gaussian, so it is symmetric and unimodal. The conditional mean, conditional median, and MAP estimates are all the same: $\hat{X}_{median} = \hat{X}_{MMSE} = \hat{X}_{MAP}$.
-
- For more general distributions the three estimators differ, and the posterior mean is not linear in the data.

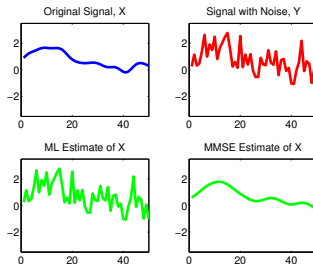
Comparison of the ML and MMSE estimators

Let $X \sim N(0, R_x)$ and $Y = X + W$ with $W \sim N(0, \sigma_w^2 I)$, where

$$[R_x]_{i,j} = \frac{1}{1 + \left| \frac{i-j}{10} \right|^2}, \quad \sigma_w = 0.75, \quad p = 50.$$

- The covariance falls off slowly with $|i - j|$, so nearby samples are highly correlated and X is probably smooth.
- The added noise destroys much of that correlation, so Y is noisy compared to X .
- For this problem the ML estimate is just the data itself, $\hat{X}_{ML} = Y$, and the MMSE estimate is $\hat{X}_{MMSE} = R_x(R_x + R_w)^{-1}Y$.

ML versus MMSE: the bias–variance trade-off



- MMSE wins here because we know a great deal about the behavior of X . ML is unbiased but uses none of that knowledge.
- ML suits problems with far more data than unknowns. Here there is one data point per unknown.
- MMSE pays for its accuracy with strong assumptions about X and a biased estimate. The estimate is smoother than the true signal.
- The bias–variance trade-off is fundamental and usually cannot be avoided.

Summary

- **Frequentist:** θ is deterministic, $\hat{\theta} = T(Y)$ is random, and $\text{MSE}_{\theta} = \text{var}_{\theta} + (\text{bias}_{\theta})^2$.
- **ML:** $\hat{\theta} = \arg \max_{\theta} \log p_{\theta}(Y)$. Biased in general, but consistent and asymptotically efficient. Use it when data far outnumbers unknowns.
- **Bayesian:** X is random with prior $p_x(x)$, and we minimize the Bayes' risk $\mathbb{E}[C(X, T(Y))]$.
- The cost picks the estimator: squared error gives $\hat{X}_{MMSE} = \mathbb{E}[X|Y]$, absolute error the conditional median, and zero-one cost the MAP estimate.
- For jointly Gaussian X and Y the three agree, and $\hat{X}_{MMSE} = R_x(R_x + R_w)^{-1}Y$ is an attenuated, and so biased, version of the data.
- **MAP** is the equation to remember:
 $\arg \min_x \{-\log p_{y|x}(y|x) - \log p_x(x)\}$, a data term plus a prior term.

Reading: Chapter 2.3. **Next lecture:** Chapter 2.4, Discrete-Time Random Processes.

Appendix

Appendix: the ML estimate of a multivariate Gaussian

Theorem

Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, R)$ random vectors with $\mu \in \mathbb{R}^p$ and R symmetric positive definite, and let $\hat{\mu}$ and $\hat{R}$ be the sample mean and sample covariance defined in Example 2.4 as

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k ,$$

$$\hat{R} = \frac{1}{n} \sum_{k=1}^n x_k x_k^t - \hat{\mu} \hat{\mu}^t ,$$

where $\hat{R}$ is positive definite. Then the joint ML estimate, where the maximization is over $\mu \in \mathbb{R}^p$ and over symmetric positive definite matrices R , is given by

$$\arg \max_{(\mu, R)} \log p(x|\mu, R) = (\hat{\mu}, \hat{R}) .$$

The remaining slides give the proof.

Appendix: proof setup

The log likelihood is given by

$$\log p(x|\mu, R) = -\frac{n}{2} \log |R| - \frac{1}{2} \text{tr} \{ S R^{-1} \} + b^t R^{-1} \mu - \frac{n}{2} \mu^t R^{-1} \mu + \text{const} , \quad (\text{A.1})$$

where b and S are defined as

$$b = \sum_{k=1}^n x_k , \quad S = \sum_{k=1}^n x_k x_k^t .$$

Furthermore, we can use b and S to define the quantities

$$\hat{\mu} = \frac{b}{n} , \quad \hat{R} = \frac{S}{n} - \frac{b b^t}{n^2} = \frac{S}{n} - \hat{\mu} \hat{\mu}^t ,$$

which have the interpretation of being the sample mean and covariance.

Appendix: maximizing over μ for fixed R

Our goal is to jointly maximize (A.1) over the two parameters (μ, R) , given by

$$(\mu^*, R^*) = \arg \max_{(\mu, R)} \log p(x|\mu, R) .$$

We can do this by first maximizing over μ for a fixed value of R , and then substituting the resulting function back in and maximizing over the second parameter R .

Our first step is therefore to compute

$$\mu^*(R) = \arg \max_{\mu} \log p(x|\mu, R) .$$

Only the last two terms of (A.1) depend on μ , so setting the gradient with respect to μ to zero yields

$$R^{-1}b - nR^{-1}\mu^* = 0 \quad \implies \quad \mu^*(R) = \frac{b}{n} = \hat{\mu} ,$$

where the last equality is the definition of $\hat{\mu}$. Since R^{-1} is positive definite, the quadratic in μ is strictly concave, so this stationary point is the maximum.

Notice that the maximizer is not actually a function of R , so we can write the solution as $\mu^* = \hat{\mu}$, which is a function of the statistic b alone.

Appendix: the ML estimate of the covariance

The joint maximization can now be done in two steps: substitute the optimum value $\mu^*(R) = \hat{\mu}$ into (A.1), and then maximize over R . This yields

$$R^* = \arg \max_R \log p(x|\mu^*(R), R) = \arg \max_R \log p(x|\hat{\mu}, R) .$$

Next, we work on simplifying the expression for $\log p(x|\hat{\mu}, R)$. To do this, we first observe that $b = n\hat{\mu}$, and we substitute this into the last two terms of (A.1), which results in

$$b^t R^{-1} \hat{\mu} - \frac{n}{2} \hat{\mu}^t R^{-1} \hat{\mu} = n \hat{\mu}^t R^{-1} \hat{\mu} - \frac{n}{2} \hat{\mu}^t R^{-1} \hat{\mu} = \frac{n}{2} \text{tr} \{ \hat{\mu} \hat{\mu}^t R^{-1} \} .$$

Then, combining the two trace terms by the linearity of the trace,

$$\begin{aligned} \log p(x|\hat{\mu}, R) &= -\frac{n}{2} \log |R| - \frac{1}{2} \text{tr} \{ S R^{-1} \} + \frac{n}{2} \text{tr} \{ \hat{\mu} \hat{\mu}^t R^{-1} \} + \text{const} \\ &= -\frac{n}{2} \log |R| - \frac{n}{2} \text{tr} \left\{ \left(\frac{S}{n} - \hat{\mu} \hat{\mu}^t \right) R^{-1} \right\} + \text{const} \\ &= -\frac{n}{2} \left[\log |R| + \text{tr} \{ \hat{R} R^{-1} \} \right] + \text{const} \triangleq g(R) + \text{const} , \end{aligned}$$

where the last step substitutes the definition $\hat{R} = S/n - \hat{\mu} \hat{\mu}^t$.

Appendix: a change of variables from R to B

By the hypothesis of the theorem, $\hat{R}$ is positive definite, so it has the eigendecomposition $\hat{R} = E\Lambda E^t$, where E is the orthonormal matrix of eigenvectors and Λ is the diagonal matrix of positive eigenvalues. The **symmetric square root** of $\hat{R}$ is defined as

$$\hat{R}^{1/2} \triangleq E\Lambda^{1/2}E^t,$$

where $\Lambda^{1/2}$ takes the square root of each diagonal entry. It satisfies $\hat{R}^{1/2}\hat{R}^{1/2} = E\Lambda E^t = \hat{R}$.

Now define

$$B \triangleq \hat{R}^{1/2}R^{-1}\hat{R}^{1/2},$$

which is symmetric positive definite with eigenvalues $\lambda_1, \dots, \lambda_p > 0$. Since $R = \hat{R}^{1/2}B^{-1}\hat{R}^{1/2}$, maximizing over R is the same as maximizing over B .

Appendix: writing both terms of $g(R)$ using the λ_i

The two terms of $g(R)$ can both be written using the λ_i . For the trace term, the property $\text{tr}\{UV\} = \text{tr}\{VU\}$ from Example 2.1, for any matrices U and V , gives

$$\text{tr}\left\{\hat{R}R^{-1}\right\} = \text{tr}\left\{\hat{R}^{1/2}R^{-1}\hat{R}^{1/2}\right\} = \text{tr}\{B\} = \sum_{i=1}^p \lambda_i .$$

For the log determinant, $|B| = |\hat{R}^{1/2}| |R^{-1}| |\hat{R}^{1/2}| = |\hat{R}|/|R|$, and $|B| = \prod_i \lambda_i$, so

$$\log |R| = \log |\hat{R}| - \log |B| = \log |\hat{R}| - \sum_{i=1}^p \log \lambda_i .$$

Appendix: maximizing over the eigenvalues

Substituting both expressions into $g(R)$ gives

$$g(R) = -\frac{n}{2} \left[\log |\hat{R}| + \sum_{i=1}^p (\lambda_i - \log \lambda_i) \right],$$

where only the eigenvalues λ_i depend on R , through $B = \hat{R}^{1/2} R^{-1} \hat{R}^{1/2}$. The statistic $\hat{R}$ does not depend on R , because $\hat{R}$ is defined from b and S . So $\log |\hat{R}|$ is a constant, and maximizing g means minimizing each term $\lambda_i - \log \lambda_i$ separately.

Setting the derivative to zero,

$$\frac{d}{d\lambda} (\lambda - \log \lambda) = 1 - \frac{1}{\lambda} = 0 \implies \lambda = 1,$$

and the second derivative $1/\lambda^2 > 0$, so $\lambda = 1$ is the minimum.

So g is maximized when every $\lambda_i = 1$. A symmetric matrix with all eigenvalues equal to one is the identity, so $B = I$. Then $\hat{R}^{1/2} R^{-1} \hat{R}^{1/2} = I$ gives $R^{-1} = \hat{R}^{-1}$, and therefore $R^* = \hat{R}$, and the theorem is proved.