9. Bayesian estimation, Part II

ECE 830 & CS 761, Spring 2016

Overview

From last time...

- Bayesian methods assume the unknown parameter θ is stochastic and we have a prior probabilistic model for θ
- Given the prior $p(\theta)$ and data $y \sim p(y|\theta)$, we can compute the posterior

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta).$$

- Given the posterior, we can estimate θ multiple different ways. Popular examples:
  - Posterior mean: $\hat{\theta} = \mathbb{E}[\theta|y] = \int \theta p(\theta|y) d\theta$.
  - Maximum a posteriori (MAP): $\hat{\theta} = \arg\max_\theta p(\theta|y)$.
- Choosing conjugate priors ensures the posterior has a nice functional form.

Today

The multivariate Gaussian linear model...

- ... with a multivariate Gaussian prior $\Longrightarrow$ ridge regression
- ... with a multivariate Laplace prior $\Longrightarrow$ LASSO (least absolute shrinkage and selection operator) regression

These models and methods appear in a wide variety of modern machine learning and signal processing settings.

The Multivariate Gaussian Model

Today we consider the following linear statistical model

$$y = X\theta + \nu$$

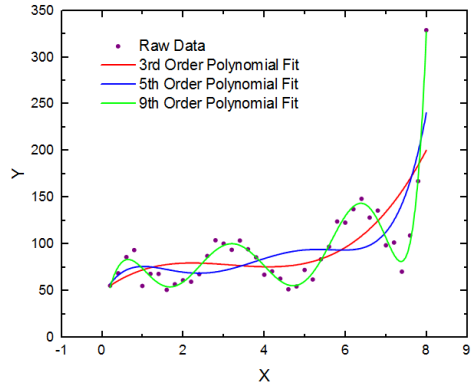
where

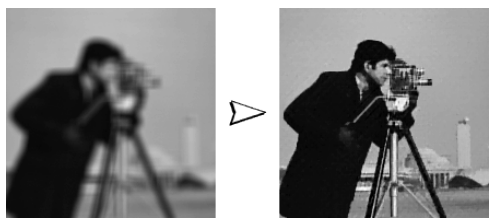
y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known

Examples

This linear model appears throughout science, engineering, and machine learning.

- y measures the US census count over n years. X is a Vandermonde matrix representing an order-p polynomial approximation, and θ contains the p polynomial coefficients.
- y is an n-pixel blurry image we take with our camera, X models the blurring process, and θ is the desired blur-free image (here $p = n$).
- Each element of y is your heart rate at n different times of the day. Each of the p columns of X measures one of your activities (e.g. alcohol consumption, nap time(s), exercise, proximity to attractive people) at the same times in the day. θ characterizes how much each of these activities contributes to your heart rate.





A Gaussian prior

Consider the following Bayesian linear statistical model

$$y = X\theta + \nu$$

where

y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known and full-rank
θ	$\sim$	$\mathcal{N}(\mu_\theta, \Sigma_\theta)$
θ	is	unknown, $p \times 1$ (p unknown parameters)
μ_θ	is	known, $p \times 1$
Σ_θ	$\in$	$\mathcal{S}_p$ is known and full-rank
θ and ν	are	independent

This model amounts to a Gaussian prior on θ and a Gaussian conditional distribution of y given θ.

What is the posterior?

First, note the y and θ are jointly Gaussian:

$$\begin{bmatrix} \theta \\ y \end{bmatrix} = \begin{bmatrix} 0 & I_p \\ I_p & X \end{bmatrix} \begin{bmatrix} \nu \\ \theta \end{bmatrix}.$$

Since

$$\begin{bmatrix} \nu \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} \Sigma_\nu & 0 \\ 0 & \Sigma_\theta \end{bmatrix} \right),$$

we have

$$\begin{bmatrix} \theta \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_\theta \\ X\mu_\theta \end{bmatrix}, \begin{bmatrix} \Sigma_\theta & \Sigma_\theta X^\top \\ X\Sigma_\theta & X\Sigma_\theta X^\top + \Sigma_\nu \end{bmatrix} \right).$$

Lemma

If

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

then

$$Z_1 | Z_2 = z_2 \sim \mathcal{N}\left(\mu', \Sigma'\right)$$

where

$$\mu' := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)$$
$$\Sigma' := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

We next apply this lemma to $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \theta \\ y \end{bmatrix}$.

Gauss-Markov Theorem

The posterior distribution of $\theta|y$ is

$$\theta|y \sim \mathcal{N}(\mu_{\theta|y}, \Sigma_{\theta|y})$$

where

$$\begin{aligned}
\mu_{\theta|y} &= \mu_\theta + \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1}(y - X\mu_\theta) \\
&= \mu_\theta + \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1} X^\top \Sigma_\nu^{-1}(y - X\mu_\theta) \\
\Sigma_{\theta|y} &= \Sigma_\theta - \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1} X\Sigma_\theta \\
&= \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1}
\end{aligned}$$

The second version of each expression is a result of the following:

Matrix Inversion Lemma

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Specifically, first apply the matrix inversion lemma to $\Sigma_{\theta|y} = \Sigma_\theta - \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1} X\Sigma_\theta$ to get $\Sigma_{\theta|y} = \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1}$.

Now let $G := \Sigma_\theta X^\top \left(X\Sigma_\theta X^\top + \Sigma_\nu\right)^{-1}$, so that

$$\begin{aligned}\mu_{\theta|y} &= \mu_\theta + G(y - X\mu_\theta)\\ \Sigma_{\theta|y} &= \Sigma_\theta - GX\Sigma_\theta.\end{aligned}$$

Now using the formula above $\Sigma_{\theta|y} = \Sigma_\theta - GX\Sigma_\theta$, we have

$$\begin{aligned} GX\Sigma_\theta =& \Sigma_\theta - \Sigma_{\theta|y} = \Sigma_{\theta|y}(\Sigma_{\theta|y}^{-1}\Sigma_\theta - I) \\ =& \Sigma_{\theta|y}\left[\left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)\Sigma_\theta - I\right] \\ =& \Sigma_{\theta|y} X^\top \Sigma_\nu^{-1} X \Sigma_\theta. \end{aligned}$$

That last gives the identity

$$G = \Sigma_{\theta|y} X^\top \Sigma_\nu^{-1} = \left(X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1}\right)^{-1} X^\top \Sigma_\nu^{-1}$$

as desired.

Observations

- The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the posterior mean and MAP estimators are both

$$\begin{aligned}\widehat{\theta}(y) = \mu_{\theta|y} &= \mu_\theta + \Sigma_\theta X^\top (X\Sigma_\theta X^\top + \Sigma_\nu)^{-1}(y - X\mu_\theta) \\ &= \mu_\theta + (X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1})^{-1} X^\top \Sigma_\nu^{-1}(y - X\mu_\theta)\end{aligned}$$

- $\widehat{\theta}(y)$ is an affine function of y.
- $\widehat{\theta}(y)$ is itself multivariate Gaussian.

Special case 1: The noncommittal prior

Consider the case where $\mu_\theta = 0$, $\Sigma_\theta = \sigma^2 I_p$ and $\sigma^2 \longrightarrow \infty$. This can be thought of as a "noncommittal" prior. Then $\Sigma_\theta^{-1} \longrightarrow 0_p$ and

$$\widehat{\theta}(y) = \mu_{\theta|y} \quad = \quad (X^\top \Sigma_\nu^{-1} X)^{-1} X^\top \Sigma_\nu^{-1} y$$

Furthermore, if $\Sigma_\nu = \sigma^2 I_n$, then

$$\widehat{\theta}(y) = (X^\top X)^{-1} X^\top y$$

which is the least squares estimator and MLE!

Special case 2: Uncorrelated prior

Consider the case where $\mu_\theta = 0$, $\Sigma_\theta = \sigma_\theta^2 I_p$, and $\Sigma_\nu = \sigma_\nu^2 I_n$. Then

$$\begin{aligned}\widehat{\theta}(y) = \mu_{\theta|y} =& (X^\top \Sigma_\nu^{-1} X + \Sigma_\theta^{-1})^{-1} X^\top \Sigma_\nu^{-1} y \\ =& (\frac{1}{\sigma_\nu^2} X^\top X + \frac{1}{\sigma_\theta^2} I_p)^{-1} \frac{1}{\sigma_\nu^2} X^\top y \\ =& (X^\top X + \frac{\sigma_\nu^2}{\sigma_\theta^2} I_p)^{-1} X^\top y\end{aligned}$$

This is referred to as ridge regression.

Note that even when $X^\top X$ is poorly conditioned or not invertible, the sum $X^\top X + \frac{\sigma_\nu^2}{\sigma_\theta^2} I_p$ can be well conditioned even for small values of $\frac{\sigma_\nu^2}{\sigma_\theta^2}$. As a result, this estimator, while biased, can have far less variance than the least squares estimator.

Example:

$y = \theta + \nu \in \mathbb{R}^n \;\; , \;\; \nu \sim \mathcal{N}(0, \sigma^2 I_n)$

$$
\begin{aligned}
p(\theta) &= \mathcal{N}(0, \Sigma_{\theta\theta}) \text{ indep. of } \nu \\
\mathbb{E}[y] &= \mathbb{E}[\theta] + \mathbb{E}[\nu] = 0 \\
\mathbb{E}\left[yy^\top\right] &= \mathbb{E}\left[\theta\theta^\top\right] + \mathbb{E}\left[\theta\nu^\top\right] + \mathbb{E}\left[\nu\theta^\top\right] + \mathbb{E}\left[\nu\nu^\top\right] \\
&= \Sigma_{\theta\theta} + \sigma^2 I \\
\mathbb{E}\left[y\theta^\top\right] &= \mathbb{E}\left[\theta\theta^\top\right] + \mathbb{E}\left[\nu\theta^\top\right] \\
&= \Sigma_{\theta\theta} = \mathbb{E}\left[\theta y^\top\right] \\
\begin{bmatrix} y \\ \theta \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{\theta\theta} \\ \Sigma_{\theta\theta} & \Sigma_{\theta\theta} \end{bmatrix}\right)
\end{aligned}
$$

We can invoke the Gauss-Markov theorem to get

$$\widehat{\theta} = \Sigma_{\theta\theta}(\Sigma_{\theta\theta} + \sigma^2 I_n)^{-1} y = (I + \sigma^2 \Sigma_{\theta\theta}^{-1})^{-1} y$$

Example: (cont.)

(Alternative derivation:) From our Bayesian perspective, we are interested in $p(\theta|y)$.

$$p(\theta|y) = \frac{(2\pi)^{-N/2}\,(2\pi)^{-N/2}\,|\Sigma|^{-1/2}\,\exp\left\{-\frac{1}{2}[y^\top\theta^\top]\Sigma^{-1}\begin{bmatrix} y \\ \theta \end{bmatrix}\right\}}{(2\pi)^{-N/2}\,|\Sigma_{yy}|^{-1/2}\,\exp\left\{-\frac{1}{2}y^\top\Sigma_{yy}^{-1}y\right\}}$$

In this formula we are faced with

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{yy} & \Sigma_{y\theta} \\ \Sigma_{\theta y} & \Sigma_{\theta\theta} \end{bmatrix}^{-1}$$

The inverse of this covariance matrix can be written as

$$\begin{bmatrix} \Sigma_{yy} & \Sigma_{y\theta} \\ \Sigma_{\theta y} & \Sigma_{\theta\theta} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{yy}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\Sigma_{yy}^{-1}\Sigma_{y\theta} \\ I \end{bmatrix} Q^{-1} \begin{bmatrix} -\Sigma_{\theta y}\Sigma_{yy}^{-1} & I \end{bmatrix}$$

where $Q := \Sigma_{\theta\theta} - \Sigma_{\theta y}\,\Sigma_{yy}^{-1}\,\Sigma_{y\theta}$ is the Schur complement of $\Sigma_{\theta\theta}$. (Verify this formula by applying RHS above to Σ to get I.)

Example: (cont.)

Furthermore,

$$\det \Sigma = \det \Sigma_{yy} \det Q.$$

Substituting this expression into $p(\theta|y)$ we get

$$\begin{aligned} p(\theta|y) &= (2\pi)^{-N/2} |Q|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2}(\theta - \Sigma_{\theta y}\Sigma_{yy}^{-1} y)^\top Q^{-1} (\theta - \Sigma_{\theta y}\Sigma_{yy}^{-1} y) \right\} \\ \theta|y &\sim \mathcal{N}\left(\Sigma_{\theta y}\Sigma_{yy}^{-1} y, Q\right) \end{aligned}$$

Thus the posterior mean of θ is

$$\widehat{\theta} = \Sigma_{\theta y}\Sigma_{yy}^{-1} y$$

and the posterior variance is

$$Q = \Sigma_{\theta\theta} - \Sigma_{\theta y}\Sigma_{yy}^{-1}\Sigma_{y\theta}$$

Example: DC Level in AWGN

$$y_i = A + \nu_i,\ i = 1, \ldots, n$$

where A is an unknown scalar to be estimated and

$$\begin{aligned} A &\sim \mathcal{N}(\mu_A, {\sigma_A}^2) \\ \nu_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, {\sigma_\nu}^2)\ , \text{ indep. of } y \end{aligned}$$

This problem falls within the Gaussian linear model with

$$\begin{aligned} X &= \mathbf{1} && (n \times 1) \\ \theta &= A && (1 \times 1) \\ \mu_\theta &= \mu_A && (1 \times 1) \\ \Sigma_\theta &= \sigma_A^2 && (1 \times 1) \\ \Sigma_\nu &= \sigma^2 I_n && (1 \times 1) \end{aligned}$$

Example: (cont.)

Using the second formula for $\mu_{A|x}$, we obtain

$$\begin{aligned}
\widehat{A}(y) = \mu_{A|x} &= \mu_A + \left(\frac{1}{\sigma_\nu^2}\mathbf{1}^\top\mathbf{1} + \frac{1}{\sigma_A^2}\right)^{-1}\mathbf{1}^\top\frac{1}{\sigma_\nu^2}\left(y - \mathbf{1}\mu_A\right) \\
&= \mu_A + \left(\frac{n}{\sigma_\nu^2} + \frac{1}{\sigma_A^2}\right)^{-1}\frac{1}{\sigma_\nu^2}\left(\sum_i y_i - N\mu_A\right) \\
&= \mu_A + \frac{1}{\frac{n}{\sigma_\nu^2} + \frac{1}{\sigma_A^2}}\frac{n}{\sigma_2^2}(\overline{y} - \mu_A) \\
&= \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_\nu^2}{n}}(\overline{y} - \mu_A)
\end{aligned}$$

Example: (cont.)

In other words,

$$\widehat{A}(y) = (1-\alpha)\mu_A + \alpha\overline{y}$$

where

$$\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_\nu^2}{n}}$$

controls the tradeoff between prior knowledge and data. Limiting cases:

$$\begin{aligned}
n \to \infty &\implies \alpha \to 1 &&\implies \widehat{A} \to \overline{y} \\
n = 0 &\implies \alpha = 0 &&\implies \widehat{A} = \mu_A \\
\sigma_A^2 \to \infty &\implies \alpha \to 1 &&\implies \widehat{A} \to \overline{y} \\
\sigma_A^2 = 0 &\implies \alpha = 0 &&\implies \widehat{A} = \mu_A
\end{aligned}$$

Tikhinov regularization

We can also compute the MAP estimator directly. Assume here that $\Sigma_\nu = \sigma_\nu^2 I_n$ and $\mu_\theta = 0$, and define Γ such that $\Sigma_\theta^{-1} = \Gamma^\top \Gamma^1$. Then

$$
\begin{aligned}
-\log p(y|\theta) &\propto \frac{1}{2\sigma_\nu^2}\|y - X\theta\|_2^2 \\
-\log p(\theta) &\propto \frac{1}{2}\|\theta^\top \Sigma_\theta^{-1}\theta\|_2^2 = \frac{1}{2}\|\Gamma\theta\|_2^2 \\
\hat{\theta}_{\mathrm{MAP}} &= \arg\min_\theta \{-\log p(y|\theta) - \log p(\theta)\} \\
&= \arg\min_\theta \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2}{2}\|\Gamma\theta\|_2^2 \right\} \\
&= (X^\top X + \sigma_\nu^2 \Gamma^\top \Gamma)^{-1} X^\top y
\end{aligned}
$$

(If $\Gamma = \sigma_\theta I_p$ we arrive at the ridge regression expression.)

Tikhinov regularization (cont.)

Consider again the estimate

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2}{2}\|\Gamma\theta\|_2^2$$

The first term measures how well θ fits our observed data. The second term reflects our prior knowledge – essentially we seek θ for which $\Gamma\theta$ has a small norm.

As σ_ν increases and we must cope with more noise in our data, we increase our dependence on this prior.

Furthermore, our choice of Γ (and hence Σ_θ) can determine which θs our estimator will be biased towards.

Example: Smooth θ

If we think θ should vary smoothly from one element to the next, we might choose Γ so that $(\Gamma\theta)_i = \theta_i - \theta_{i-1}$.

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in AWGN:

$$y = \theta + \nu$$

where y is the observed signal, θ is the clean signal, and ν is the noise. This can be modeled using a general linear model using $\theta = \theta$ and $X = I_n$. We can adopt a Gaussian prior for θ:

$$\theta \sim \mathcal{N}(0, \Sigma_{\theta\theta}).$$

The Bayesian estimate of θ is then

$$\widehat{\theta} = \Sigma_{\theta\theta} \left(\Sigma_{\theta\theta} + \Sigma_{\nu\nu}\right)^{-1} y.$$

Now suppose that $\Sigma_{\theta\theta}$ and $\Sigma_{\nu\nu}$ are simultaneously diagonalizable, meaning there exists an orthogonal matrix U such that

$$\Sigma_{\theta\theta} = U \Lambda_\theta U^\top$$

$$\Sigma_{\nu\nu} = U \Lambda_\nu U^\top$$

with $\Lambda_\theta, \Lambda_\nu$ diagonal. For example, consider $\Sigma_{\nu\nu} = \sigma^2 I$ and $\Sigma_{\theta\theta}$ arbitrary.

Then the estimator becomes

$$\begin{aligned}
\widehat{\theta} &= \Sigma_{\theta\theta}\left(\Sigma_{\theta\theta}+\Sigma_{\nu\nu}\right)^{-1} y \\
&= U\Lambda_\theta U^\top \left(U\Lambda_\theta U^\top + U\Lambda_\nu U^\top\right)^{-1} y \\
&= U\Lambda_\theta U^\top \left(U(\Lambda_\theta+\Lambda_\nu)U^\top\right)^{-1} y \\
&= U \underbrace{\left[\Lambda_\theta(\Lambda_\theta+\Lambda_\nu)^{-1}\right]}_{\Lambda} U^\top y
\end{aligned}$$

where

$$\Lambda = \begin{bmatrix}
\frac{\lambda_1^{(\theta)}}{\lambda_1^{(\theta)}+\lambda_1^{(\nu)}} & 0 & \cdots & 0 \\
0 & \frac{\lambda_2^{(\theta)}}{\lambda_2^{(\theta)}+\lambda_2^{(\nu)}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \\
0 & 0 & \cdots & \frac{\lambda_n^{(\theta)}}{\lambda_n^{(\theta)}+\lambda_n^{(\nu)}}
\end{bmatrix}.$$

Interpretation:

- U is a change of basis matrix
- $\theta = U^\top y$ are coefficients of y in new basis
- $z = \Lambda\theta$ is a coordinate-wise rescaling of θ
- $\widehat{\theta} = Uz$ is a reconstruction of θ from z.

How should we interpret the weights

$$\lambda_i := \frac{\lambda_i^{(\theta)}}{\lambda_i^{(\theta)} + \lambda_i^{(\nu)}}?$$

Notice that

$$\begin{aligned}
U^\top y &= U^\top \theta + U^\top \nu \\
U^\top \theta &\sim \mathcal{N}(0, U^\top \Sigma_{\theta\theta} U) = \mathcal{N}(0, \Lambda_\theta) \\
U^\top \nu &\sim \mathcal{N}(0, U^\top \Sigma_{\nu\nu} U) = \mathcal{N}(0, \Lambda_\nu).
\end{aligned}$$

Writing

$$U = \left[\begin{array}{cccc} u_1 & u_2 & \cdots & u_n \end{array} \right]$$

we have

$$u_i^\top \theta \sim \mathcal{N}(0, \lambda_i^{(\theta)})$$
$$u_i^\top \nu \sim \mathcal{N}(0, \lambda_i^{(\nu)})$$

Thus, λ_i reflects the proportion of the projection onto u_i that is due to the signal.

A Laplacian prior

Consider the following Bayesian linear statistical model

$$y = X\theta + \nu$$

where

y	is	observed, $n \times 1$
X	is	known, $n \times p$
ν	$\sim$	$\mathcal{N}(0, \Sigma_\nu)$, is $n \times 1$
Σ_ν	$\in$	$\mathcal{S}_n$ (the set of all $n \times n$ positive semi-definite real-valued matrices) is known and full-rank
θ	is	unknown, $p \times 1$ (p unknown parameters)
θ	$\sim$	Laplace(λ)
$p(\theta)$	$=$	$\prod_{i=1}^{p} \frac{\lambda}{2} \exp(-\lambda \lvert\theta_i\rvert)$
λ	is	known scalar
θ and ν	are	independent

This model amounts to a Laplacian prior on θ and a Gaussian conditional distribution of y given θ.

Using the Laplacian prior

With the Laplacian prior we do not get the same simple expression for the posterior distribution.

However, we can still examine the MAP estimate where $\Sigma_\nu = \sigma_\nu^2 I_n$:

$$
\begin{aligned}
-\log p(y|\theta) &\propto \frac{1}{2\sigma_\nu^2}\|y - X\theta\|_2^2 \\
-\log p(\theta) &\propto \lambda \sum_{i=1}^{p} |\theta_i| \equiv \lambda\|\theta\|_1 \\
\hat{\theta}_{\text{MAP}} &= \arg\min_\theta \{-\log p(y|\theta) - \log p(\theta)\} \\
&= \arg\min_\theta \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2 \lambda}{2}\|\theta\|_1 \right\}
\end{aligned}
$$

This estimate is called the LASSO (least absolute shrinkage and selection operator) estimate.

Ridge vs. LASSO

$$\hat{\theta}_{\text{Ridge}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2}{2\sigma_\theta^2}\|\theta\|_2^2 \right\}$$

$$\hat{\theta}_{\text{LASSO}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\nu^2\lambda}{2}\|\theta\|_1 \right\}$$

In both cases, we attempt to find a θ which (a) is a good fit to our data and (b) adheres to prior information captured by either the ℓ_2 or ℓ_1 norm of θ.

When should we use one vs. the other?

In general, the LASSO estimator favors *sparser* θ – i.e., θ with more zero-valued elements. There is no closed-form expression for the LASSO estimate.

Example: Deblurring

θ is a p-pixel image. X is a blur operator. $y = X\theta + \nu$ is a p-pixel blurry, noisy image. Some of the eigenvalues of X are very close to zero, so

$$(X^\top X)^{-1}$$

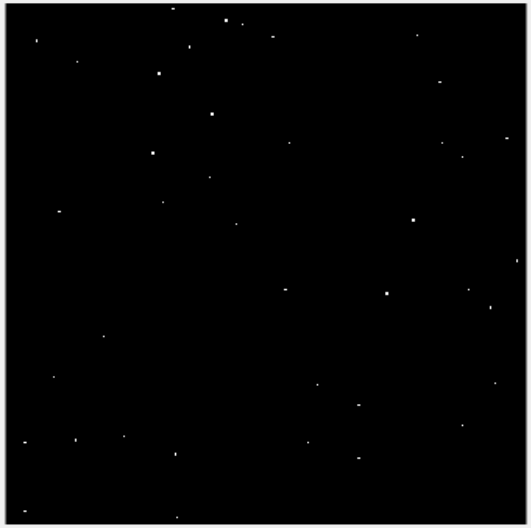
has some huge elements, meaning the least squares estimate

$$\hat{\theta} = (X^\top X)^{-1} X^\top y = (X^\top X)^{-1} X^\top (X\theta + \nu) = \theta + (X^\top X)^{-1} X^\top \nu$$

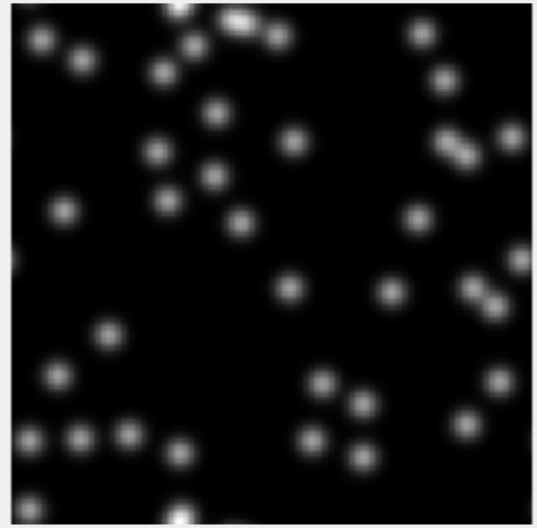
will contain θ plus amplified noise.
We will compare the least-squares estimate, the ridge estimate, and the LASSO estimate.

Example: Deblurring

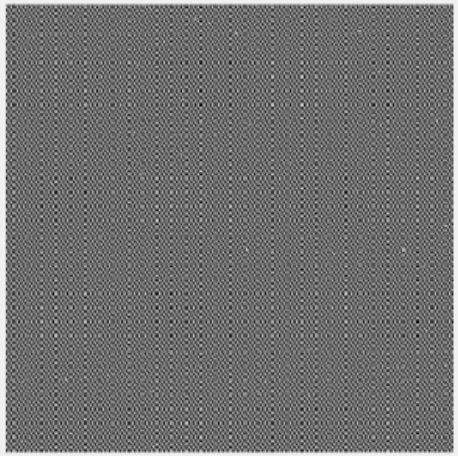
original



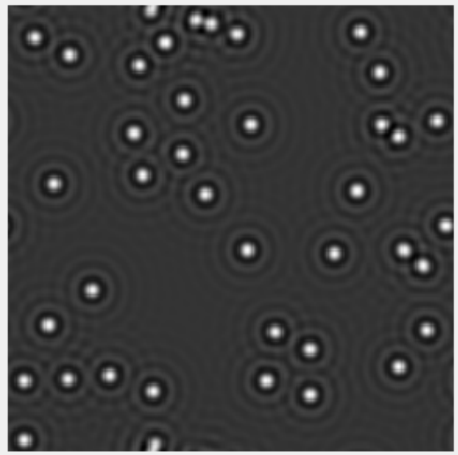
observed



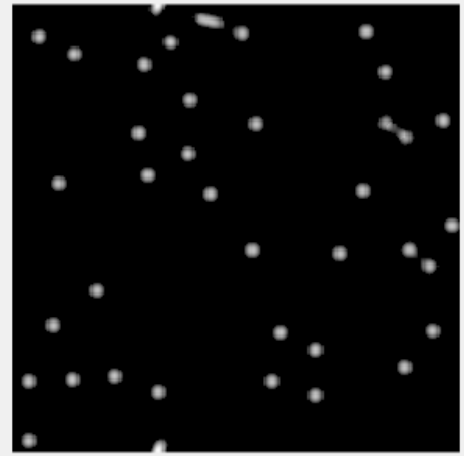
LS, MSE = 17495.1339



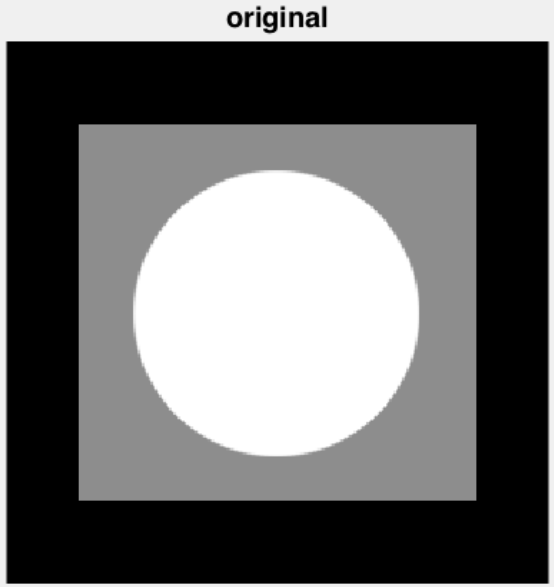
Tikhonov, MSE = 37.9969

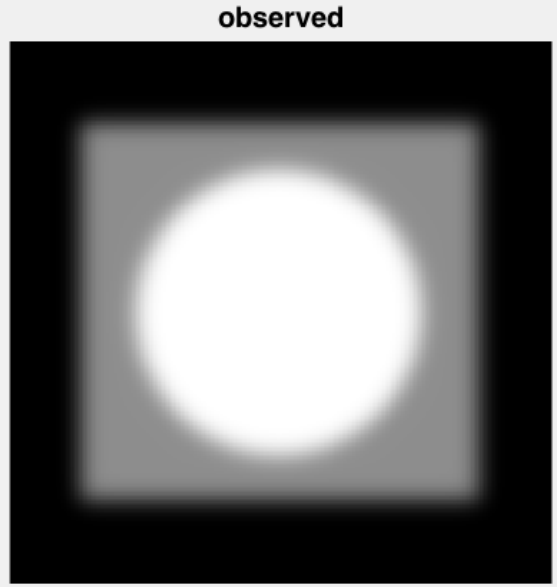


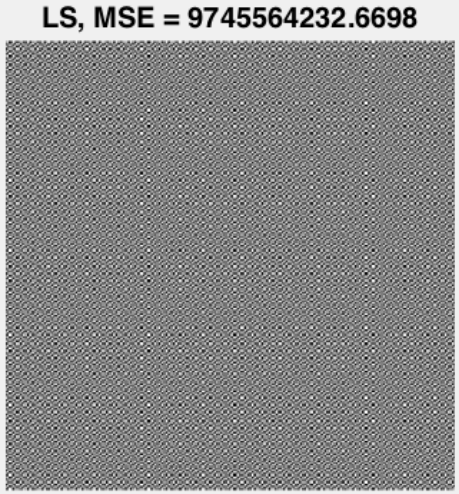
LASSO, MSE = 38.612

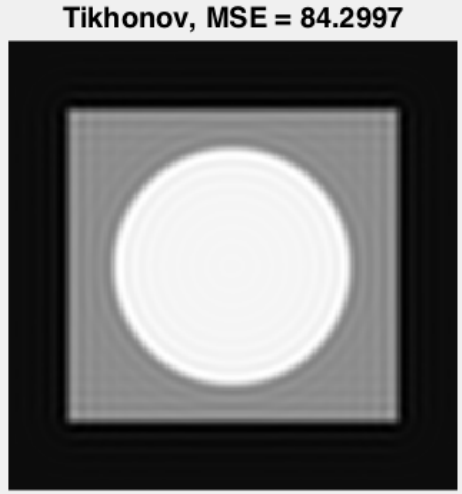


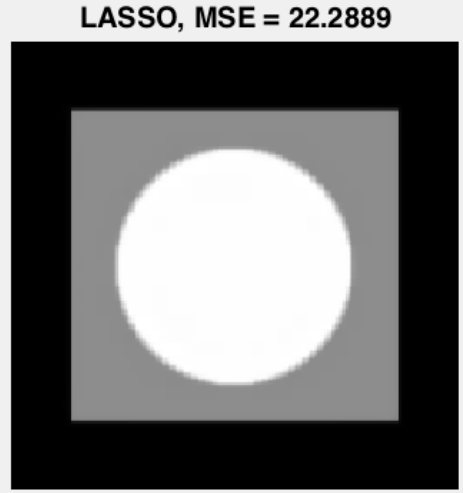
Example: Deblurring











Proof of lemma

Note that the conditional distribution must be a Gaussian, so we just need to calculate the mean and covariance of this Gaussian to fully characterize the distribution.
Let $A := -\Sigma_{12}\Sigma_{22}^{-1}$ and $t := z_1 + Az_2$. Then

$$\text{cov}[t, z_2] = \text{cov}(z_1 + Az_2, z_2) = \Sigma_{12} + A\Sigma_{22} = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$$

which means that t and z_2 are uncorrelated, and since they're Gaussian this implies that they are also independent. Thus

$$\begin{aligned}
\mathbb{E}[z_1|z_2] =& \mathbb{E}[t - Az_2|z_2] = \mathbb{E}[t|z_2] - A\mathbb{E}[z_2|z_2] \\
=& \mathbb{E}[t] - Az_2 = \mu_1 + A(\mu_2 - z_2) \\
=& \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)
\end{aligned}$$

Proof of lemma (cont.)

We now use the fact that for two random vectors x and y,

$$\mathsf{var}(x+y) = \mathsf{var}(x) + \mathsf{var}(y) + \mathsf{cov}(x,y) + \mathsf{cov}(y,x).$$

Thus

$$\begin{aligned}
\mathsf{var}(z_1|z_2) =& \mathsf{var}(t - Az_2|z_2) \\
=& \mathsf{var}(t|z_2) + A\mathsf{var}(z_2|z_2)A^\top - \mathsf{cov}(t, Az_2) - \mathsf{cov}(Az_2, t) \\
=& \mathsf{var}(t) = \mathsf{var}(z_1 + Az_2) \\
=& \mathsf{var}(z_1) + A\mathsf{var}(z_2)A^\top + A\mathsf{cov}(z_1, z_2) + \mathsf{cov}(z_2, z_1)A^\top \\
=& \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
=& \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\end{aligned}$$