STA 360/601: Bayesian and Modern Statistics

Lecture 10: Multivariate Gaussian models

Jeff Miller

Department of Statistical Science, Duke University

Friday, September 26, 2014

Bivariate Gaussian distribution

- $Y = (Y_1, Y_2)' \sim N_2(\mu, C)$ (*bivariate Gaussian*) has pdf

$$f(y) = (2\pi)^{-1}|C|^{-1/2} \exp\left\{ -\frac{1}{2}(y-\mu)'C^{-1}(y-\mu) \right\},$$

  where $|A| = |\det A|$
- Gaussian distribution is parameterized by mean μ and covariance matrix C
- $\mu = \mathrm{E}(Y) = (\mathrm{E}(Y_1), \mathrm{E}(Y_2))' = (\mu_1, \mu_2)'$ is the mean (first moment)
- $C = \mathrm{cov}(Y) =$ covariance matrix - characterizes co-variability in the different elements of Y
- "Normal" = "Gaussian"

Covariance matrix

- Note that $Y = (Y_1, Y_2)'$ is a bivariate random variable, and its components may be linearly dependent
- The bivariate normal distribution allows this dependence to range from perfectly negatively correlated to zero (independent) to perfectly positively correlated
- The covariance matrix C encodes this dependence and the marginal variances of Y_1 and Y_2
- Writing $C = (C_{jk})$ with C_{jk} denoting the (j, k)th element of the covariance matrix, we have

$$C_{11} = \text{var}(Y_1), \quad C_{22} = \text{var}(Y_2), \quad C_{12} = C_{21} = \text{cov}(Y_1, Y_2)$$

Covariance matrix

- The covariance between Y_1 and Y_2 is defined as

$$\mathrm{cov}(Y_1, Y_2) = \mathrm{E}(Y_1 Y_2) - \mathrm{E}(Y_1)\mathrm{E}(Y_2).$$

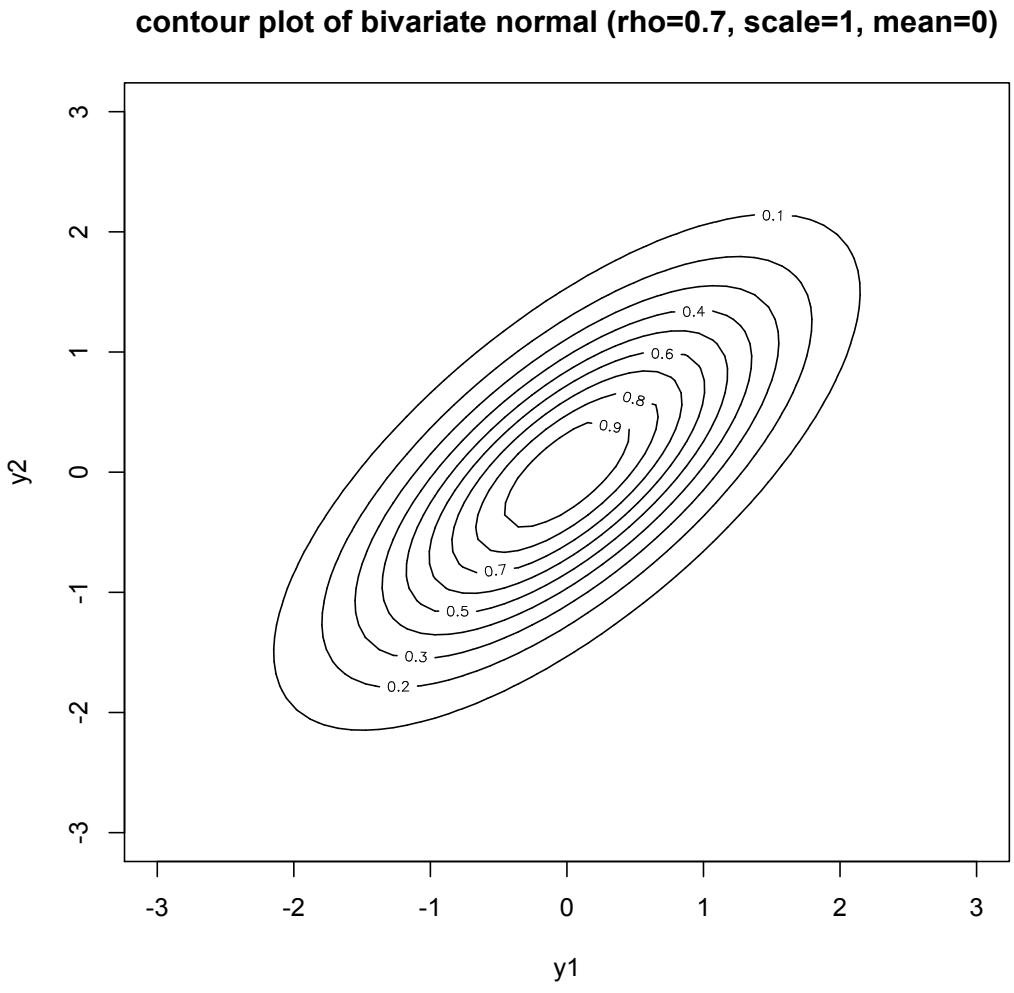
- The correlation coefficient between Y_1 and Y_2 is defined as

$$\rho_{12} = \frac{\mathrm{cov}(Y_1, Y_2)}{\sqrt{\mathrm{var}(Y_1)}\sqrt{\mathrm{var}(Y_2)}} = \frac{C_{12}}{\sigma_1 \sigma_2},$$

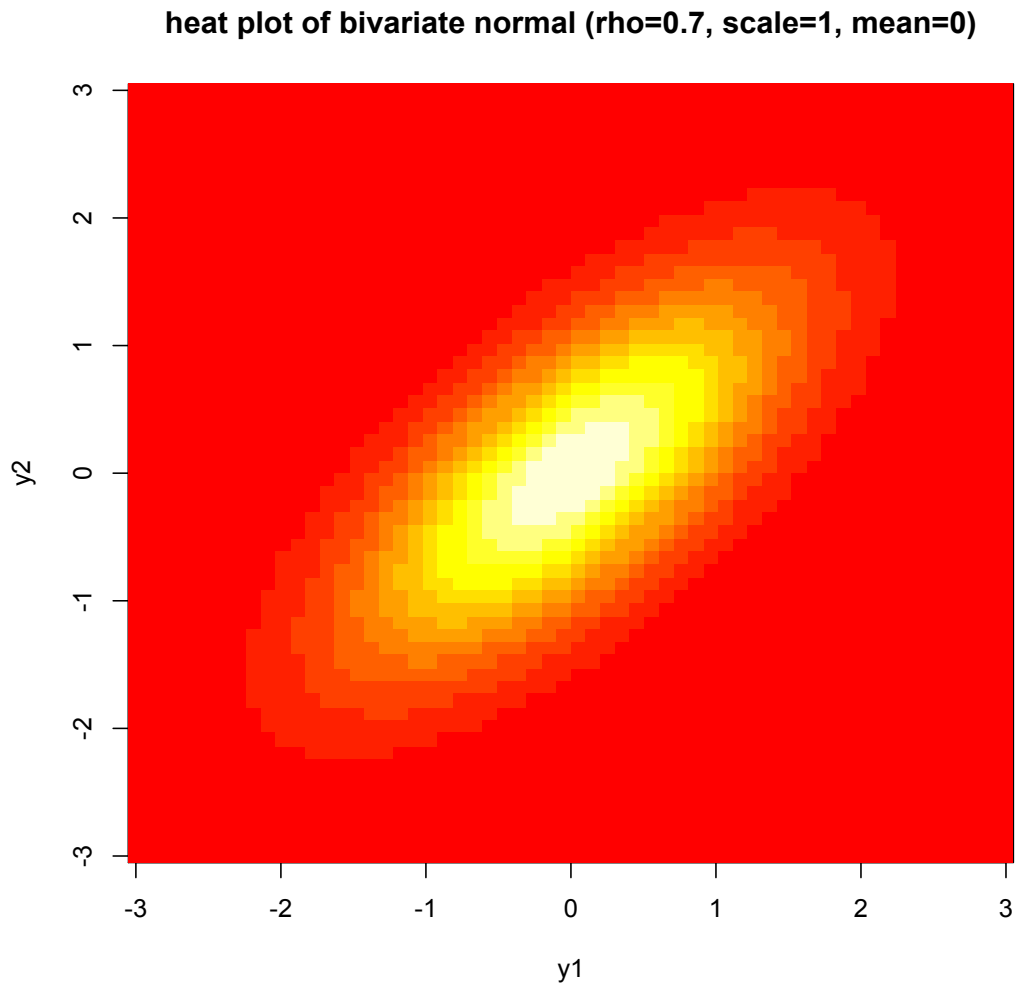
  where $\sigma_1 = \sqrt{C_{11}}$ and $\sigma_2 = \sqrt{C_{22}}$.

- $-1 \leq \rho_{12} \leq 1$ and the correlation coefficient is free of the measurement units

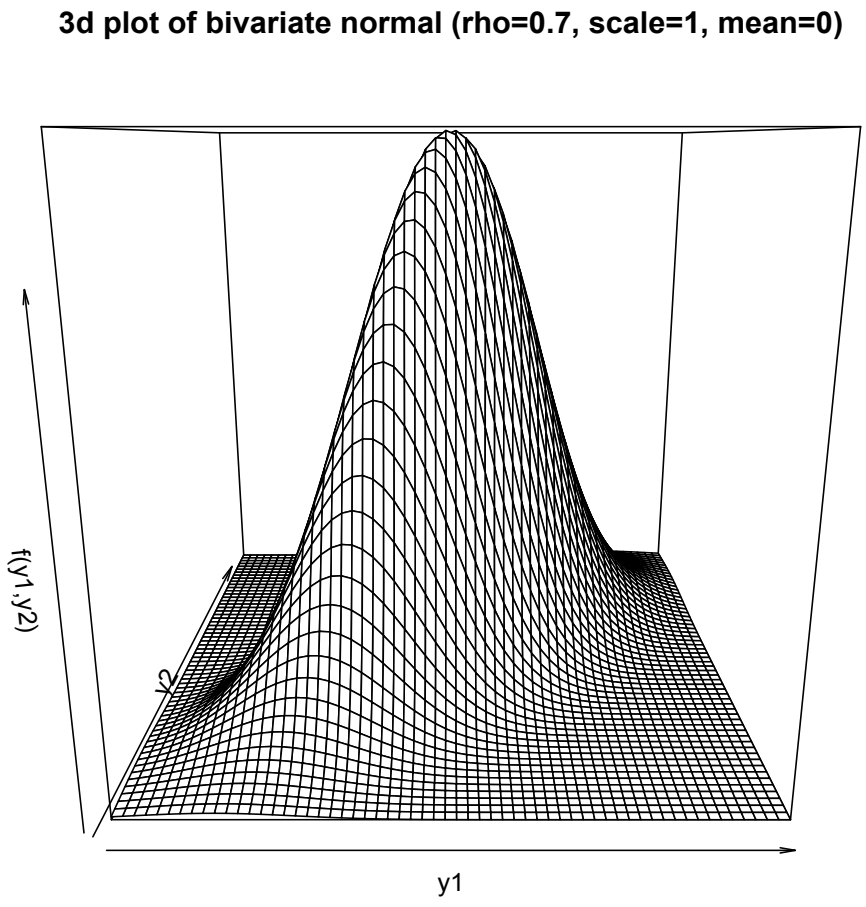
Plotting bivariate Gaussian - contour plot



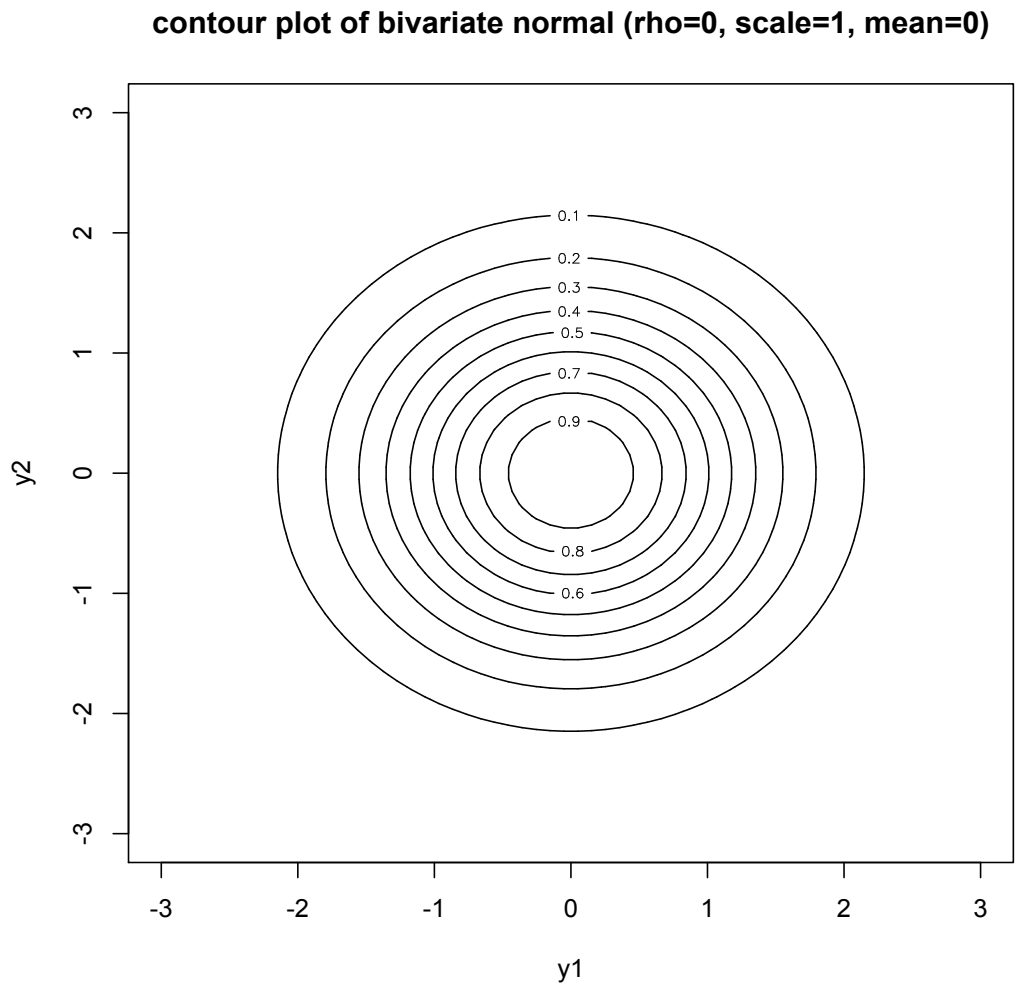
Plotting bivariate Gaussian - heat plot



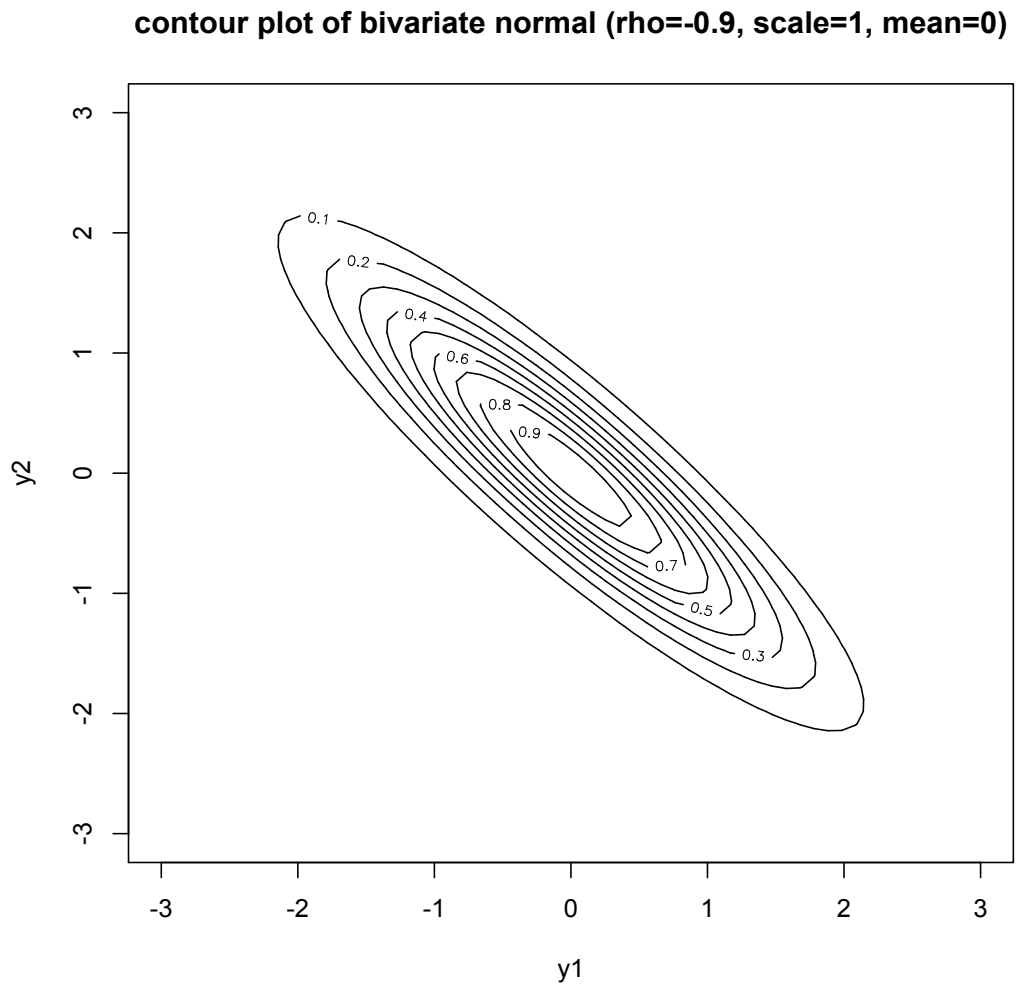
Plotting bivariate Gaussian - 3d perspective plot



Contour plot of spherical Gaussian



Contour plot - high negative correlation



General Multivariate Gaussian distribution

- Straightforward to generalize to multiple dimensions: the pdf of the p-dimensional Gaussian $N(\mu, C)$ is

$$f(y) = (2\pi)^{-p/2}|C|^{-1/2}\exp\left\{-\frac{1}{2}(y-\mu)'C^{-1}(y-\mu)\right\},$$

  where $y = (y_1, \ldots, y_p)'$
- μ is now a $p \times 1$ vector and C is a $p \times p$ matrix
- Can obtain as follows: if $Z = (Z_1, \ldots, Z_p)'$ with $Z_i \overset{iid}{\sim} N(0,1)$ then $Y = C^{1/2}Z + \mu \sim N(\mu, C)$.
- Interpretation is maintained in multivariate case: (j,k)th element of C is $\text{cov}(Y_j, Y_k)$
- Special property of multivariate normal: Y_j and Y_k are independent iff they are uncorrelated (i.e., $\rho_{jk} = 0$)

Conditional Distributions

- Let $Y = (Y_a', Y_b')'$ with Y_a the first q elements of Y and Y_b the remaining $p - q$ elements
- Suppose $Y \sim N_p(\mu, C)$ with

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix},$$

  where μ_a is $q \times 1$, μ_b is $(p-q) \times 1$, C_{aa} is $q \times q$, C_{bb} is $(p-q) \times (p-q)$
- Then the conditional distribution of Y_a given $Y_b = y_b$ is

$$(Y_a | Y_b = y_b) \sim N_q\big(\mu_a + C_{ab}C_{bb}^{-1}(y_b - \mu_b), C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}\big).$$

Bayesian inference for the mean μ

- For data $Y_i = (Y_{i1}, \ldots, Y_{ip})' \sim N(\mu, C)$ the likelihood is $\propto$

$$|C|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)' C^{-1} (y_i - \mu) \right\}$$

- Under a multivariate Gaussian prior $\mu \sim N_p(\mu_0, C_0)$, the conditional posterior distribution of μ is

$$(\mu | C, y_{1:n}) \sim N_p(\hat{\mu}, \hat{C}),$$

where

$$\hat{C} = (C_0^{-1} + nC^{-1})^{-1},$$

and

$$\hat{\mu} = \hat{C}(C_0^{-1}\mu_0 + nC^{-1}\overline{y}).$$

- $\hat{C}$ is the conditional posterior covariance of μ. Note that "learning" is more rapid when the precision (inverse covariance) C^{-1} is "big".
- $\hat{\mu}$ is the conditional posterior mean of μ.

What about inference for the covariance matrix?

- In the univariate case, $Y_i \sim N(\mu, \sigma^2)$, an inverse-gamma prior is commonly chosen for the variance σ^2
- This is equivalent to a gamma prior for the precision $\lambda = 1/\sigma^2$
- In the multivariate Gaussian case, we have a covariance matrix C instead of a scalar
- It would be nice to have a matrix-valued extension of the inverse-gamma that would be conjugate

Positive definite & symmetric

- One complication is that the covariance C must be symmetric and *positive definite* (posdef)
- i.e., $C = C'$ and $x'Cx > 0$ for all $x \in \mathbb{R}^p$ s.t. $x \neq 0$
- Note that this ensures that the diagonal elements of C (corresponding to the marginal variances σ_j^2) are positive
- Also, ensures that the correlation coefficients for each pair of variables are between -1 and 1.
- Prior for C must assign probability 1 to the set of symmetric positive definite matrices

Constructing a random posdef matrix

- It seems daunting to directly specify a prior with appropriate support
- But consider the following:
- Sample $Z_\ell \overset{iid}{\sim} N_p(0, U)$ for $\ell = 1, \ldots, \nu$, where U is a $p \times p$ symmetric posdef matrix
- Let $\Lambda = \sum_{\ell=1}^{\nu} Z_\ell Z_\ell'$, i.e.,

$$\Lambda_{jk} = \sum_{\ell=1}^{\nu} Z_{\ell j} Z_{\ell k}$$

- The distribution of Λ is called Wishart, and denoted $W_p(\nu, U)$
- $\nu =$ "degrees of freedom", $U =$ "scale matrix"

Wishart - some properties

- Λ is always symmetric
- If the degrees of freedom $\nu \geq p$, then Λ is posdef with probability 1
- $\mathrm{E}(\Lambda) = \nu U$
- Hence, U can be viewed as a scaled version of the mean of Λ
- The pdf of $\Lambda \sim \mathrm{W}_p(\nu, U)$ is $\propto$

$$|\Lambda|^{\frac{\nu-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(U^{-1}\Lambda)}$$

  where $\mathrm{tr}(\cdot)$ is the *trace* function (sum of diagonal elements)
- In the univariate case in which $p = 1$, this reduces to

$$\lambda^{\nu/2-1} e^{-\frac{1}{2}u^{-1}\lambda} \propto \mathrm{Ga}(\lambda|\nu/2, 1/(2u))$$

Wishart - some comments

- Wishart provides a conditionally-conjugate prior for the precision matrix $\Lambda = C^{-1}$ in a multivariate normal model
- Wishart is a multivariate generalization of the gamma prior for a scalar precision (1/variance)
- What about the covariance matrix?

Inverse Wishart

- When $\Lambda \sim \mathrm{W}_p(\nu, U)$, $\nu \geq p$, the distribution of Λ^{-1} is called *Inverse Wishart*, and denoted $\mathrm{IW}_p(\nu, S)$, where $S = U^{-1}$.
- $C = \Lambda^{-1}$ is symmetric posdef with probability 1
- The pdf of $C \sim \mathrm{IW}_p(\nu, S)$ is $\propto$

$$|C|^{-(\nu+p+1)/2} e^{-\frac{1}{2}\mathrm{tr}(SC^{-1})}$$

- $\nu =$ "degrees of freedom", $S =$ "scale matrix" (posdef)
- $\mathrm{E}(C) = \frac{1}{\nu-p-1} S$
- If we have a prior guess C_0 for C then we might choose $S = (\nu - p - 1)C_0$
- Under this choice, $\mathrm{E}(C) = C_0$, and $\nu = p + 2$ would correspond to a vague prior
- The inverse Wishart is a conditionally-conjugate prior for the multivariate normal covariance and provides a multivariate generalization of the inverse-gamma

Bayesian inference for the covariance matrix

- Suppose $C \sim \text{IW}_p(\nu, S)$ and $Y_1, \ldots, Y_n \sim N_p(\mu, C)$
- Letting $S_{y,\mu} = \sum_{i=1}^n (y_i - \mu)(y_i - \mu)'$, we have

$$\begin{aligned} f(C|\mu, y_{1:n}) &\propto |C|^{-(\nu+p+1)/2} e^{-\text{tr}(SC^{-1})/2} |C|^{-n/2} e^{-\text{tr}(S_{y,\mu}C^{-1})/2} \\ &= |C|^{-(\hat{\nu}+p+1)/2} e^{-\text{tr}(\hat{S}C^{-1})/2} \end{aligned}$$

where

$$\hat{\nu} = \nu + n, \quad \hat{S} = S + S_{y,\mu}.$$

- (Rewriting the Gaussian likelihood in this manner relies on a linear algebra trick — see Hoff pg 111)
- Hence, we get the conditional posterior $C|\mu, y_{1:n}$ as $\text{IW}(\hat{\nu}, \hat{S})$.

Homework Exercise

- Simulate data $Y_i \sim N_2(\mu, C)$, $i = 1, \ldots, 100$, with $\mu = (0, 0)'$ and C chosen so that $\rho_{12} = 0.8$, $C_{11} = 2$, $C_{22} = 1$.
- Compute the MLE of (μ, C) (you don't have to derive it, you can look it up) and plot the contours of the distribution with the MLE as parameters, compared with the contours of the true distribution.
- Assuming independent normal and inverse Wishart priors for μ and C, respectively, run a Gibbs sampler. (The hyperparameters are up to you but keep it somewhat informative.)
- Make a scatterplot of your posterior samples for μ, and indicate the MLE and the true parameters on the plot.
- Do the same thing (scatterplot, etc.) with C_{11} on the x-axis vs. C_{22} on the y-axis.
- Make a histogram of the ρ_{12} values from your posterior samples and indicate the MLE and the true value on the plot.