Lecture 5: Multiparameter Models

UNIVERSITY OF
SAN FRANCISCO

James D. Wilson

MSAN 628

Computational Statistics

Multiparameter Models

- So far, we have focused on *single parameter* models:
  - the parameter θ is one-dimensional
  - only a *marginal* prior distribution, and hence posterior distribution needs to be constructed / calculated
- In this lecture, we'll talk a bit about *multiparameter* models, where θ has two or more dimensions
- In particular, we'll look at multiparameter Gaussian models and logistic models

Multiparameter Models

- Gaussian models:
  - $y \mid (u, \sigma^2) \sim N(\mu, \sigma^2)$
  - In general, we can also talk about *multivariate* versions where $\mathbf{y} \mid (\mu, \Sigma) \sim N(\mu, \Sigma)$
- Logistic models:
  - For multi-class classification with k classes
  - $f(\mathbf{y} \mid \theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$
- Hierarchical models:
  - For parameters (θ, ϕ)
    - Hierarchical structure - $p(\theta, \phi) = p(\theta)p(\theta \mid \phi)$ and we observe $y \mid \theta$
    - Both θ and ϕ are treated as random

Nuisance Parameter Regime

Nuisance Parameter Situation: when we are only interested in one parameter of many provided in the model.

Example: $y \mid (\mu, \sigma^2) \sim N(\mu, \sigma^2)$ and we are concerned only with μ.

- First, calculate the joint posterior density:

$$p(\mu, \sigma^2 \mid y) \propto f(y \mid \mu, \sigma^2) p(\mu, \sigma^2)$$

- Calculate $p(\mu \mid y)$ by "averaging out" the nuisance parameter σ^2:

$$p(\mu \mid y) = \int p(\mu, \sigma^2 \mid y) d\sigma^2$$

Nuisance Parameter

- Alternatively, we can write $p(\mu \mid y)$ as

$$p(\mu \mid y) = \int p(\mu \mid y, \sigma^2) p(\sigma^2 \mid y) d\sigma^2$$

- The above shows that $p(\mu \mid y)$ is a *mixture* of the conditional posterior distributions given the nuisance parameter, σ^2 and $p(\sigma^2 \mid y)$ is a **weighting** function for different values of σ^2
- **Issue:** the joint density $p(\mu, \sigma^2 \mid y)$ is hard to calculate, but...

Simulating Joint posterior distributions

... in general, the previous slide shows that there is a natural strategy for computation of the joint posterior $p(\theta_1, \theta_2 \mid y)$ in two parameter models:

Practical strategy to compute $p(\theta_1, \theta_2 \mid y)$

1. for (r in 1:N) loop
   - simulate θ_2 from $p(\theta_2 \mid y) \rightarrow \theta_2^{(r)}$
   - given $\theta_2^{(r)}$, simulate θ_1 from $p(\theta_1 \mid y, \theta_2^{(r)}) \rightarrow \theta_1^{(r)}$
2. return joint density - $(\theta_1^{(1)}, \theta_2^{(1)}), \ldots (\theta_1^{(N)}, \theta_2^{(N)})$

Conclusion: We don't need the joint posterior if we can calculate the marginal posterior densities!

Example: simulating from a joint posterior

Let's see this in action with a two-parameter Gaussian model.

Setting:

- **Data**: $y \mid (\mu, \sigma^2) \sim N(\mu, \sigma^2)$
- **Noninformative prior**: uniform on $(\mu, \log \sigma)$ (see page 64):

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

Example: simulating from a joint posterior

Posterior Calculations:

- $\mu \mid (\sigma^2, y) \sim N(\overline{y}, \sigma^2/n)$
- $\sigma^2 \mid y \sim Inv - \chi^2(n-1, s^2)$

Note: This is really an illustrative example, as we can actually analytically calculate the joint density (if we have the willpower):

$$p(\mu, \sigma^2 \mid y) \propto \sigma^{-n-2} \exp\left(1 \frac{1}{2\sigma^2}[(n-1)s^2 + n(\overline{y} - \mu)^2]\right)$$

Example: simulating from a joint posterior

To make draws from the posterior of $(\mu, \sigma^2 \mid y)$, we can follow our strategy:

Computing $p(\mu, \sigma^2 \mid y)$

1. for (r in 1:N) loop
   - simulate σ^2 from $Inv - \chi^2(n-1, s^2) \rightarrow \sigma^{2(r)}$
   - given $\sigma^2(r)$, simulate μ from $N(\overline{y}, \sigma^2(r)/n) \rightarrow \mu^{(r)}$
2. return joint density - $(\mu^{(1)}, \sigma^{2(1)}), \ldots (\mu^{(N)}, \sigma^{2(N)})$

Multiparameter Conjugate Gaussian Models

One can do the math to figure out that the conjugate family for the Gaussian model is given by the following specification:

- $\mu \mid \sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$
- $\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$

Above κ_0, ν_0, and σ_0^2 are hyperparameters and assumed to be unknown and fixed.

In summary, this specification gives the following *joint prior density*:

$$p(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]\right)$$

Multiparameter Conjugate Gaussian Models

One nice feature about this family of Gaussian models is that in fact, it is conjugate across **all** parameters. That is,

- $\mu \mid (\sigma^2, y)$ has the same distributional form as $\mu \mid \sigma^2$
- $\sigma^2 \mid y$ has the same distributional form as σ^2
- $(\mu, \sigma^2) \mid y$ has the same distributional form as (μ, σ^2)

Key Takeaway: At least we know (i) how to estimate the mean and variance of a posterior, and (ii) the distributional form of each parameter. That's about as close as we can get.

Multivariate Gaussian Data

Suppose that we have p- dimensional observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. **Data**:

$\mathbf{y} \mid (\mu, \Sigma) \sim N_p((\mu, \Sigma))$

- $\mu = p$-dimensional mean vector
- $\Sigma = p \times p$ *positive definite* variance-covariance matrix

Applications:

- Multivariate Linear Regression
- Linear Discrimination
- Quadratic Discrimination
- Graphical Models
- Multivariate Time Series (Gaussian Processes)

Multivariate Gaussian Data

Properties: Suppose that $\mathbf{y} \mid (\mu, \Sigma) \sim N_p((\mu, \Sigma))$, then

- $y_j \sim N(\mu_j, \Sigma_{j,j})$ for all entries j
- $y_j \mid \mathbf{y}_{-j}$ is also Gaussian
- y_j and y_k are independent if and only if $\Sigma_{j,k} = 0$
- y_j and y_k are independent if and only if $\Sigma^{-1}_{j,k} = 0$
- $\Omega = \Sigma^{-1}$ is called the Gaussian Graphical Model for $\mathbf{y}$ since it explains all dependence among the entries of $\mathbf{y}$

Bivariate Gaussian Data

Let $(X, Y) \sim N_2(\mu, \Sigma)$, where

- $\mu = (\mu_X, \mu_Y)$
- $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{pmatrix}$

Then, $Y \mid X = x \sim N(\tilde{\mu}, \tilde{\sigma}^2)$ where

- $\tilde{\mu} = \mu_Y + \frac{\sigma_{X,Y}}{\sigma_Y^2}(x - \mu_X)$
- $\tilde{\sigma}^2 = \sigma_Y^2 + \frac{\sigma_{X,Y}^2}{\sigma_X}$

This looks familiar...

Bivariate Gaussian Data

Recall simple linear regression: $Y = \beta_0 + \beta_1(x - \overline{x}) + \epsilon$ Then least squares optimization gives:

- $\widehat{\beta}_0 = \overline{y}$
- $\widehat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}(x_i - \overline{x})^2}$

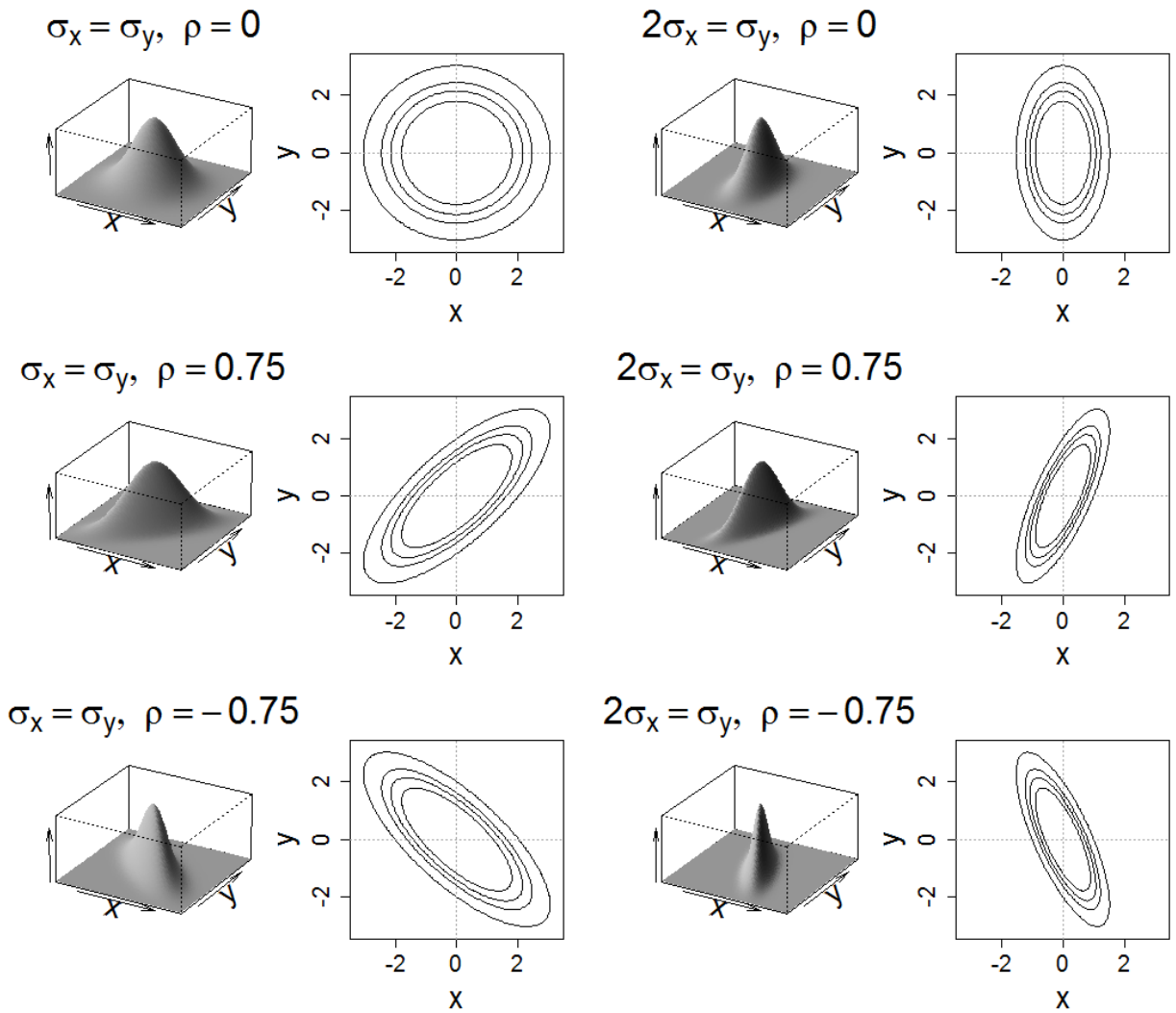
And, the best unbiased estimate of Y is

$$\underbrace{\widehat{Y}}_{\tilde{\mu}} = \underbrace{\widehat{\beta}_0}_{\widehat{\mu}_Y} + \underbrace{\overline{\beta}_1}_{\frac{\widehat{\sigma}_{X,Y}}{\widehat{\sigma}_Y}} (x - \underbrace{\overline{x}}_{\widehat{\mu}_X})$$

Conclusion: When ϵ is Gaussian, $Y \mid X = x$ acts as if (X, Y) are jointly Gaussian (even though we don't need any requirement on our observed data!)

Bivariate Gaussian Data

Data can be easily visualized:



Logistic Models

Setting: Consider the multi-class classification problem -

- $y_1, \ldots, y_n \in \{1, \ldots, k\}$
- Calculate $z_j = \sum_{i=1}^{n} \mathbb{I}(y_i = j)$ = counts of class j

Goal: classify y_j given data $\mathbf{x}_j$

- We can do this probabilistically using logistic regression models!

Logistic Regression Models

Relies on the multinomial distribution:

$$z_i \sim multinomial(\mathbf{p} = (p_1, p_2, \ldots, p_k)),$$

$$p(z_i \mid \mathbf{p}) = \frac{n!}{z_1! \cdots z_k!} p_1^{z_1} \cdots p_k^{z_k}$$

$$\propto \prod_{j=1}^{k} p_j^{z_j}$$

where $\sum_{j=1}^{k} p_j = 1$. Then, the general goal is to estimate p_j given the data x_j.

Question: is there a natural conjugate distribution for $\mathbf{p}$?

The Dirichlet Distribution

The Dirichlet (Dir(α)) distribution is a commonly used multivariate prior distribution for multivariate vectors $\mathbf{x}$ that take values on $(0,1)^K$:

- $f(\mathbf{x} \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$
- $\alpha_1, \ldots, \alpha_K > 0$
- Multivariate extension of the Beta(α, β) distribution
- Prior makes a lot of sense for probability parameters
- Has been widely used in machine learning applications:
  - Multi-armed bandit
  - Latent Dirichlet Allocation
- https://en.wikipedia.org/wiki/Dirichlet_distribution

The Dirichlet-multinomial model

The Dirichlet prior is a conjugate prior for the multinomial regression.

Dirichlet-multinomial model:

- $p(z_i \mid \mathbf{p}) \propto \prod_{j=1}^{k} p_j^{z_j}$
- $\pi(\mathbf{p}) \propto \prod_{i=1}^{K} p_i^{\alpha_i - 1}$

The resulting *joint posterior* is:

$$p(\mathbf{p} \mid \mathbf{z}) = Dir(\alpha + \mathbf{z})$$

Example

Pre-election polling (pg 69)

In late October, 1988, a survey was conducted by CBS News of 1447 adults in the United States to find out their preferences in the upcoming presidential election. Out of 1447 persons, 727 supported George Bush, 583 supported Michael Dukakis, and 137 supported other candidates. Construct a model to estimate the population difference in support for George Bush and Michael Dukakis. Provide an interpretation for your results.

Example

Comparison of two Polls

ABC News conducted a survey of 639 persons before a presidential debate between Bush and Dukakis, and an independent group of 639 persons were polled after the presidential debate. The results were recorded as follows:

Survey	Bush	Dukakis	Other
pre-debate	294	307	38
post-debate	288	332	19

Construct a model for which you can identify a distribution for the value $\alpha_2 - \alpha_1$: the difference of who favored Bush before from the proportion who favored Bush after. What is the probability that there was a shift toward Bush?

Next Time

- Ties of Bayesian Analysis with Machine Learning
- Interview Questions that can be answered with Probability