STAT 221 NOTES - STATISTICAL COMPUTING AND LEARNING

GREG TAM

Contents

1. September 8th, 2014 2
1.1. Modeling Basics 2
1.2. Data Generating Process (DGP) 4
2. September 10th, 2014 4
2.1. Random Effect Models 4
2.2. Mixture Models 4
2.3. DGP (Other Method) 5
2.4. Mixed Membership Models/Admixtures 5
2.5. Homework 6
3. September 15th, 2014 6
3.1. Panos' Part of Lecture 6
3.2. Solving $Ax = b$ 7
3.3. Solving Eigenvalue 8
4. September 17th, 2014 8
4.1. Optimization 8
4.2. Termination 8
4.3. Newton-Raphson 9
4.4. Quasi-Newton 9
5. September 22nd, 2014 9
5.1. LDA ("Latent Dirichlet Allocation") 9
5.2. DGP for LDA (for 1 document) 10
5.3. Mixed Membership Stochastic Block Model 10
6. September 24th, 2014 11
6.1. Mixture of Isoform 11
6.2. Latent Space Model 11
6.3. Principal Component Analysis 11
6.4. State Space Models 12
7. September 29th, 2014 13
7.1. Odyssey 13
8. October 1st, 2014 13
8.1. Procedure 13
8.2. Multinomial 14
8.3. Dirichlet Distribution 14
9. October 6th, 2014 15
9.1. Stochastic Approximation (Robbins& Monro, 51) 15
9.2. Statistical Estimation (Sakerson?, 65) 16
10. October 15th, 2014 17
10.1. Stochastic Gradient Descent 17
10.2. GLM Model 17
10.3. Problem 2 in Assignment 3 18
11. October 20th, 2014 18
11.1. Monte Carlo Integration 18
11.2. Importance Sampling 19
12. October 22nd, 2014 21
12.1. MCMC 21
12.2. Metropolis-Hastings Monte Carlo Markov Chain (MH-MCMC) 22
12.3. Independence Metropolis-Hastings 23
12.4. Gibbs Sampler 23
13. October 27th, 2014 24

13.1. Expectation-Maximization 24
14. October 29th, 2014 26
14.1. EM 26
14.2. Statistics EM 26
14.3. E-step at iteration t 27
14.4. M-step at iteration t 27
15. November 3rd, 2014 28
15.1. EM Example (mixtures) 28
15.2. Variational EM 29
16. November 5th, 2014 29
16.1. EM (Latent Dirichlet Allocation by Bla et al. JMLB 2003) 29
17. November 10th, 2014 32
17.1. State Space Models 32
17.2. FA/Normal-Normal 32
18. November 12th, 2014 34
18.1. EM for Linear Gaussian State Space Models 34
18.2. General State Space Model 36
19. November 17th, 2014 36
19.1. Particle Filters 36
20. November 19th, 2014 38
21. December 1st, 2014 38

1. September 8th, 2014

There is no distinction between the notation of density and mass functions.

- f, g, h are deterministic functions
- $\mathbb{P}(y|\theta)$, where $y \in \mathcal{Y}, \theta \in \Theta$.
- $\mathbb{P}_{Y|\Theta}(y|\theta) = \mathbb{P}_{Y|\Theta}(Y = y|\theta)$
  Y is a random variable or a latent variable and θ is a constant
- $\mathbb{P}_{Y|\Theta}(Y = y|\Theta = \theta)$, Y is a random variable, θ is a constant.
- $\mathbb{E}_{X|\Theta}[g(X, Y)] = \mathbb{E}[g(X, Y)|\theta]$
  - $\mathbb{P}(y, x_1, \ldots, x_n, \theta)$ is the joint distribution
  - $\mathbb{P}(x_1|x_2, \ldots, x_n, \theta)$ is the full conditional distribution
  - $\mathbb{P}(x_1, x_2|x_3, \ldots, x_n, \theta)$ is the conditional distribution
  - $\mathbb{P}(y|\theta)$ is the marginal distribution

$\mathbb{P}(y, x, \theta)$, $x = (x_1, \ldots, x_n)$.

- Y is a random variable
- X is a latent variable
- θ is a constant

The distinction between a variable and a constant is that the variables (X and Y) are such that $\text{Var}(X) > 0, \text{Var}(Y) > 0$.

1.1. **Modeling Basics.**

1.1.1. *Setup.* We are given $\vec{x} = (x_1, \ldots, x_n), \vec{y} = (y_1, \ldots, y_n)$. We want to estimate some quantity $\mathbb{Q}(\vec{x}, \vec{y})$, which is the estimand. There are assumptions made on this, which is called the **model**. For example, we can have

$$x_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$
$$y_i \overset{iid}{\sim} \mathcal{N}(\tau, 1)$$

where $x_i \perp\!\!\!\perp y_i$.

	Variable	Constant
Obs	$x_1, \ldots, x_n, y_1, \ldots, y_n$	
Latent		μ, τ, σ^2

1.1.2. *Class Notation.* We will typically use Y as the random variable, X as the latent variable, and θ as a constant.

- $\mathbb{P}(Y|\theta)$ is the likelihood
- $\mathbb{P}(Y, X|\theta)$ is the complete likelihood
- $\mathbb{P}(X|Y, \theta)$ is the posterior distribution

Example 1.

$$\mathbb{P}\big(x_1,\dots,x_n,y_1,\dots,y_n|\mu,\tau,\sigma^2\big)=\prod_{i=1}^{n}\mathcal{N}(x_i,\mu,\sigma^2)\cdot\mathcal{N}(y_i|\tau_i)$$

The maximum likelihood estimator and the method of moments are methods to estimate constants.

$$\hat{\theta}_{MLE}=\arg\max_{\theta}\mathbb{P}(Y|\theta)$$

To do the method of moments, we have

$$\mathbb{E}[Y|\theta]=\frac{1}{n}\sum_{i=1}^{n}y_i$$
$$\mathbb{E}\big[Y^2|\theta\big]=\frac{1}{n}\sum_{i=1}^{n}y_i^2$$

which give us $\hat{\mu}_{MOME}$ and $\hat{\sigma}^2_{MOME}$.

In our setup, we can compute $\mathbb{P}(X|Y,\theta)$, but we cannot compute

$$\begin{aligned}\mathbb{P}(Y|\theta)&=\int_X\mathbb{P}(Y,X|\theta)\ \mathrm{d}x\\&=\int_X\mathbb{P}(Y|X,\theta)\,\mathbb{P}(X|\theta)\ \mathrm{d}x\end{aligned}$$

Definition 1 (EM Algorithm)**.** *The EM (Expectation-Maximization) algorithm is a strategy to estimate θ when $\mathbb{P}(Y|\theta)$ is not invertible, but $\mathbb{P}(X|Y,\theta)$ is.*

In the first step (E step), we find $\mathbb{E}[X|Y,\theta]$ to get $\hat{x}$. In the M step, we find $\hat{\theta}_{MLE}$ with

$$\mathbb{P}\Big(Y,\hat{X}|\theta\Big)=\int\mathbb{P}(Y,X|\theta)\ \mathrm{d}x$$

Definition 2 (Monte Carlo Integration)**.** *Suppose we have $\mathbb{P}(X|Y,\theta)$. Assume we have $\tilde{x}_1,\dots,\tilde{x}_B$. Then*

$$\mathbb{P}(Y|\theta)\approx\frac{1}{B}\sum_{i=1}^{B}\mathbb{P}(Y_i,X_i|\theta)$$

Definition 3 (MCMC: Markov Chain Monte Carlo)**.** *MCMC is a strategy to explore $\mathbb{P}(X|Y,\theta)$.*

Example 2. *1. Random effects*
2. MIxtures
3. Mixed membership

Suppose $y_1,\dots,y_n$ are random variables, θ is a constant/estimand, and $x_1,\dots,x_n$ are latent variables. Our model is

$$\begin{aligned}y_i&=\alpha+\beta x_i+\varepsilon_i\\\varepsilon_i&\overset{iid}{\sim}\mathcal{N}(0,\sigma_\varepsilon^2)\\x_i&\overset{iid}{\sim}\mathcal{N}(0,\sigma_x^2)\end{aligned}$$

where $\varepsilon_i\perp\!\!\!\perp x_i$.

	$V>0$	const
obs	$y_1,\dots,y_n$	x
not obs	$\varepsilon_1,\dots,\varepsilon_n,x_1,\dots,x_n$	$\alpha,\beta,\sigma_x^2,\sigma_\varepsilon^2$

$$\begin{aligned}\mathrm{Var}(y_i)&=\mathrm{Var}(\alpha+\beta x_i+\varepsilon_i)\\&=\beta^2\,\mathrm{Var}(x_i)+\mathrm{Var}(\varepsilon_i)\\&=\beta^2\sigma_x^2+\sigma_\varepsilon^2\\&>0\end{aligned}$$

In this case, $\theta \equiv (\alpha, \beta, \sigma_X^2, \sigma_\varepsilon^2)$.

$$\mathbb{P}\big(y_1, \ldots, y_n | \alpha, \beta, \sigma_X^2, \sigma_\varepsilon^2\big) = \iint \mathbb{P}\big(y_1, \ldots, y_n, \varepsilon_1, \ldots, \varepsilon_n, x_1, \ldots, x_n | \alpha, \beta, \sigma_X^2, \sigma_\varepsilon^2\big) \, \mathrm{d}\varepsilon \mathrm{d}x$$

It can be convenient to rewrite $y_i = \alpha + \beta x_i + \varepsilon_i$ as

$$y_i - \alpha - \beta x_i = \varepsilon_i$$

so each side has equal variability. Then we can rewrite

$$\mathbb{P}\big(y_1, \ldots, y_n, \varepsilon_1, \ldots, \varepsilon_n, x_1, \ldots, x_n | \alpha, \beta, \sigma_X^2, \sigma_\varepsilon^2\big) = \mathbb{P}\big(y_1 - \alpha - \beta x_1, \ldots, y_n - \alpha - \beta x_n, \varepsilon_1, \ldots, \varepsilon_n | \alpha, \beta, \sigma_x^2, \sigma_\varepsilon^2\big)$$

$$\mathbb{1}(\varepsilon_i = y_i - \alpha - \beta x_i) i = 1, \ldots, n$$
$$\mathbb{1}(y_i = \alpha + \beta x_i + \varepsilon_i) i = 1, \ldots, n$$

1.2. **Data Generating Process (DGP).** The data generating process is a roadmap to the complete likelihood Model: $(\alpha, \beta, \sigma_x^2, \sigma_\varepsilon^2) \longrightarrow \mathbb{P}(y_1, \ldots, y_n | \bullet)$. For $i = 1, \ldots, n$,

$$\begin{aligned} \varepsilon_i &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ x_i &\sim \mathcal{N}(0, \sigma_x^2) \\ y_i &= \alpha + \beta x_i + \varepsilon_i \end{aligned}$$

Then

$$\begin{aligned} \prod_{i=1}^n \mathbb{P}\big(\varepsilon_i | 0, \sigma_\varepsilon^2\big) \, \mathbb{P}\big(x_i | 0, \sigma_x^2\big) \, \mathbb{1}(y_i = \alpha + \beta x_i + \varepsilon_i) &= \prod_{i=1}^n \mathcal{N}(\varepsilon_i | 0, \sigma_\varepsilon^2) \cdot \mathcal{N}(x_i | 0, \sigma_x^2) \mathbb{1}(y_i = \alpha + \beta x_i + \varepsilon_i) \\ &= \prod_{i=1}^n \mathcal{N}(\varepsilon_i | 0, \sigma_\varepsilon^2) \cdot \mathcal{N}(X_i, 0, \sigma_x^2) \mathbb{1}(\varepsilon_i = y_i - \alpha - \beta x_i) \\ &= \prod_{i=1}^n \mathcal{N}(y_i - \alpha - \beta x_i | 0, \sigma_\varepsilon^2) \cdot \mathcal{N}(x_i | 0, \sigma_x^2) \\ &= \mathbb{P}\big(y_1, \ldots, y_n, x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_n | \alpha, \beta, \sigma_x^2, \sigma_\varepsilon^2\big) \end{aligned}$$

2. September 10th, 2014

2.1. **Random Effect Models.** For $I = 1, \ldots, n$, we have

$$\begin{aligned} \varepsilon_i &\sim \mathcal{N}(0, \sigma_\varepsilon^2) x_i & &\sim \mathcal{N}(0, \sigma_x^2) \\ y_i &= \alpha + \beta x_i + \varepsilon_i \end{aligned}$$

Then we have

$$\begin{aligned} \mathbb{P}\big(y_1, \ldots, y_n | \alpha, \beta, \sigma_\varepsilon^2, \sigma_x^2\big) &= \int_\varepsilon \int_{X^n} \mathbb{P}\big(y, x, \varepsilon | \alpha, \beta, \sigma_\varepsilon^2, \sigma_x^2\big) \, \mathrm{d}x \mathrm{d}\varepsilon \\ &= \prod_{i=1}^n \int_{X^n} \underbrace{\mathbb{P}\big(\varepsilon_i | \sigma_\varepsilon^2\big) \, \mathbb{P}\big(x_i | \sigma_\varepsilon^2\big) \, \mathbb{P}(y_i | \varepsilon_i, x_i, \alpha, \beta)}_{\mathbb{P}(\varepsilon_i | \sigma_\varepsilon^2) \mathbb{P}(x_i | \sigma_x^2) \mathbb{1}(y_i) \mathbb{1}(\alpha + \beta x_i + \varepsilon_i)} \, \mathrm{d}x_i \end{aligned}$$

Next, rearranging terms, we get

$$\begin{aligned} \varepsilon_i &= y_i - \alpha_i - \beta_i \\ &= \mathbb{P}\big(x_i | \sigma_x^2\big) \, \mathbb{P}\big(y_i - \alpha - \beta x_i | x_i, \alpha, \beta, \sigma_\varepsilon^2\big) \end{aligned}$$

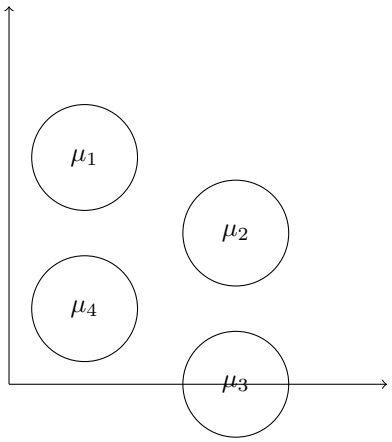
2.2. **Mixture Models.**

2.2.1. *Applications.*

- Econ: Social Strata
- Marketing: Customer Segmentation
- Biology: Functional Modules
- Statistics: Clustering 101

2.2.2. *DGP.* For $y_1, \ldots, y_n$, where $y_i \in \mathbb{R}^2$, the data generating process is for $i = 1, \ldots, n$,

$$\begin{aligned} x_i &\sim \mathcal{U}(1, \ldots, k | p_1, \ldots, p_k) \\ y_i &\sim \mathcal{N}_2(\mu_{x_i}, \sigma_{x_i}^2 I_2) \end{aligned}$$



We have variables $y_1, \ldots, y_n, x_1, \ldots, x_n, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, p_1, \ldots, p_k$.

- Random Variables: $y_1, \ldots, y_k$
- Latent Variables: $x_1, \ldots, x_k$
- Constants: $\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2, p_1, \ldots, p_k$

The likelihood is

$$\begin{aligned} \mathbb{P}\big(y_1, \ldots, y_k | \mu, \sigma^2, p, k\big) &= \int \mathbb{P}\big(y_1, \ldots, y_n, x_1, \ldots, x_n | \mu, \sigma^2, p, k\big) \, \mathrm{d}x \\ &= \prod_{i=1}^{n} \int \underbrace{\mathbb{P}\big(y_i, x_i | \mu, \sigma^2\big) \, \mathbb{P}(x_i | p)}_{\prod_{j=1}^{k} p_j^{\mathbb{1}(x_i = j)} = p_{x_i}} \, \mathrm{d}x_i \end{aligned}$$

Then,

$$p_{x_i} \mathcal{N}(y_i | \mu_{x_i}, \sigma_{x_i}^2 I_2) = \prod_{j=1}^{k} p_j^{\mathbb{1}(x_i = j} \mathcal{N}(y_i, \mu_j, \sigma_j^2 I_2)^{\mathbb{1}(x_i = j)}$$

Now we have

$$\begin{aligned} \mathbb{P}\big(y | \mu, \sigma^2, p, k\big) &= \prod_{i=1}^{n} \int_X \prod_{j=1}^{k} p_j^{\mathbb{1}(x_I = j)} \mathcal{N}(y_j | \mu_j, \sigma_j^2 I_2)^{\mathbb{1}(x_i = j)} \, \mathrm{d}x_i \\ &= \prod_{i=1}^{n} \left(\sum_{j=1}^{k} p_j \mathcal{N}(y_i | \mu_j, \sigma_j^2 I_2) \right) \end{aligned}$$

2.3. **DGP (Other Method).** For $i = 1, \ldots, n$,

$$\begin{aligned} x_i &\sim \text{Multinomial}((p_1, \ldots, p_k)', 1) \\ y_i &\sim \mathcal{N}(\underbrace{(\mu_1, \ldots, \mu_k)}_{2 \times k} \cdot \underbrace{x_i}_{k \times 1}, \underbrace{(\sigma_1^2, \ldots, \sigma_k^2)}_{1 \times k} \cdot \underbrace{x_i}_{k \times 1} I_2) \\ &= \prod_{i=1}^{n} \left(\prod_{j=1}^{k} p_j^{x_{ij}} \right) \left(\prod_{j=1}^{k} \mathcal{N}(y_i | \mu_j, \sigma_j^2 I_2)^{x_{ij}} \right) \\ &= \mathbb{P}(y_1, \ldots, y_n, x_1, \ldots, x_n | \bullet) \end{aligned}$$

2.4. **Mixed Membership Models/Admixtures.** Suppose we have observations $y_i, \ldots, y_n$, $y_i \in \mathbb{R}^2$. As a motivating example, imagine if we are assigning topics to different documents. Any given document can match multiple topics, so we call it an admixture.

For $i = 1, \ldots, n$,

$$\theta_i \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_n)$$

where

$$\theta_i = \begin{pmatrix} \theta_{i1} \\ \vdots \\ \theta_{ik} \end{pmatrix}$$

such that $\theta_j \in [0,1]$ and for $j = 1, \ldots, k$, we have

$$\sum_{k=1}^{k} \theta_{ij} = 1$$

Then for $d = 1, 2$,

$$x_{id} \sim \text{Multinomial}(\theta_i, 1)$$
$$y_{id} \sim \mathcal{N}(\underbrace{(\mu_1^{(d)}, \ldots, \mu_k^{(d)})}_{1\times k} \cdot \underbrace{x_{id}}_{k\times 1}, \underbrace{(\sigma_1^2, \ldots, \sigma_k^2)}_{1\times k} x_{id})$$

A difference between mixture models and mixed membership models is illustrated in the table below:

	Mixture	Mixed Membership
Weights	$p_1, \ldots, p_k$	$\frac{\alpha_1}{\sum \alpha_j}, \ldots, \frac{\alpha_k}{\sum \alpha_j}$

2.5. **Homework.** First problem: simple transformation theorem. Casella and Berger section 2.1 and chapter 4.6 (page 185)

3. September 15th, 2014

How do we posit a probability distribution on $(p_1, \ldots, p_k)'$ such that $p_i \in [0,1]$ and $\sum_{i=1}^{k} p_i = 1$.

$$\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \sim \mathcal{N}_k \left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \begin{pmatrix} \alpha & -\beta & -\beta \\ -\beta & \ddots & -\beta \\ -\beta & -\beta & \alpha \end{pmatrix} \right), \qquad \alpha, \beta > 0$$

The homework simply asks what is the distribution $(p_1, \ldots, p_k)'$ where

$$p_i = \frac{1}{1 + e^{-x_i}}$$

which is simply a projection into $[0,1]$.

In the second problem, we have 2 sub-populations in a $2 \times n$ matrix

$$\begin{bmatrix} g_{1H} & \cdots & g_{nH} \\ g_{1L} & \cdots & g_{nL} \end{bmatrix}$$

where $g_{iL} = 1 - g_{iH}$.

We should expect that figure 2 in CastroSinger2005.pdf should have upward sloping curves, but the model is fitting the data poorly, so we do not see this.

3.1. **Panos' Part of Lecture.**

$$\max \sum_{i=1}^{n} \ell(\theta; y_i)$$

Typical optimization problems have small p, small n and are "nice". Here, n is the number of data points, p is the number of parameters, and ℓ is the log-likelihood or "loss" function.

This becomes a lot more difficult when we have a large n or if we have a very large n and large p ($p > n$). There are also situations where we have

$$\max \sum_{i=1}^{n} \ell(\theta; y_i) + g(\beta)$$

In this case, we have very large n, large p, and a not so nice function. More commonly nowadays, we have

$$\max \sum_{i=1}^{n} \tilde{\ell}(\theta; y_i) + g(\beta)$$

where $\tilde{\ell}$ is "less nice", convex, and non-smooth.

- $F(x) = 0$
- $$\mathbb{E}[\nabla \ell(\theta^*, y)] = 0$$

  where θ^* is the true parameter value.

- $$\mathbb{E}\Big[\nabla \ell(\hat{\theta}, X, Y)\Big| Y^{obs}\Big] = \nabla \ell(\hat{\theta}|Y^{obs})$$
  where X is latent.
- The idea is to solve for $x - g(x) = 0$. We can do $X_{n+1} = g(X_n)$ iteratively, but this is not unique.
  As an example, suppose we want to solve for $5x = 100$, but we cannot divide. We can set $4x + x = 100$ and try to solve
  $$x_{n+1} = -4x_n + 100$$
  but then this would diverge. Instead, we can use $4x = -x + 100$ and then solve
  $$x_{n+1} = -\frac{1}{4}x_n + 25$$
  which will converge to the right answer.

3.2. **Solving $Ax = b$.**

- **Splitting:**
  $A = A_1 + A_2$. Then we can solve for $A_1x + A_2x = b$. Suppose that A_1 dominates A_2 and is easier to invert. Then we have that $A_1x = -A_2x + b$ and so we can iterate
  $$x_{n+1} = -A_1^{-1}A_2x_n + A_1^{-1}b$$
  When A has a diagonal that dominates, we can use the Jacobi Iteration, which sets $A_1 = \text{diag}(A)$, which is the diagonal of A and $A_2 = L(A) + U(A)$, which is the sum of the upper and lower parts of A.
- **Richardson Iteration:**
  we have that
  $$Ax = b \iff x + Ax = x + b \iff x = (I - A)x + b$$
  so we have
  $$x_{n+1} = (I - A)x_n + b$$
  If $\|I - A\|$, then
  $$x_n \longrightarrow A^{-1}b$$
  If $x_{n+1} = (I - \omega A)x + \omega b$, then
  $$x_n \longrightarrow (\omega A)^{-1}\omega b = X^{-1}b$$
  where $\|I - \omega A\| < 1$. In this way, this still converges to the same thing, but we can choose ω so that it is more stable.
- For any matrix A, there is a condition number $k(A) \geq 1$ such that
  $$k(A) = \|A\| \cdot \|A^{-1}\|$$
  If we perturb b, then the perturbation in the solution is δx and
  $$\frac{\|\delta x\|}{\|x\|} = \underbrace{k(A)}_{\text{condition number}} \times \underbrace{\frac{\|\delta b\|}{\|b\|}}_{\text{perturbation}}$$
  An alternate way to do this is similar to what we did with ω. Instead, we can have
  $$x_{n+1} = (I - CA)x_n + Cb$$
  where C is now a matrix. Then we have
  $$\begin{aligned} X_n &\longrightarrow (CA)^{-1}Cb \\ &= A^{-1}C^{-1}Cb \\ &= A^{-1}b \end{aligned}$$
- It is also possible to do this backwards. Suppose we have $Ax = b$. Then $x + Ax = (I + A)x = x + b$ and so
  $$x_{n+1} = (I + A)^{-1}(x + b)$$
- Conjugate Gradient (CG)
  $$x_d = x_0 + \alpha_1 p_1 + \alpha_2 p_2 + \cdots + \alpha_d p_d$$
  such that
  $$p_i A p_j = 0$$
  At every iteration, we have
  $$x_{n+1} = \arg\min_{x \in K} \frac{1}{2}x'Ax - b'x$$
  where K is the Krylov subspace. We don't want to find a global minimum, but a minimum on a subspace.

3.3. **Solving Eigenvalue.** We want to solve $Ax = \lambda x$. The solutions are still the same even if we have $cAx = c\lambda x$, so we restrict x to $\|x\| = 1$. The most common method of solving this is using the power method, where we have

$$x_{n+1} = \frac{Ax_n}{\|Ax_n\|}$$

which yields the largest eigenvalue. We can use shift and inversion to find the other eigenvalues.

4. September 17th, 2014

4.1. **Optimization.**

- Convergence (linear, superlinear, quadratic) What this means it the relationship

$$\frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} = O(1)$$

when $n \longrightarrow \infty$. This is linear (Nelder-Mead, uses no derivatives, and is robust). If we have

$$\frac{\|x_{n+1} - x^*\|}{\|x_{n+1} - x^*\|^2} = O(1)$$

for $1 < \alpha < 2$, it is superlinear (Quasi-Newton, uses approximation to Hessian). If we have

$$\frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|^2} = O(1)$$

then we it is quadratic (Newton Raphson, uses second order Hessian).
- Linear approximation usually means solving equations of the form $Ax = b$. Sometimes there will be constraints such as $c'x = 0, y = 0$.
- Nonlinear functions solve $\min_\Omega f(x)$, which is optimization, or $F(x) = 0$, which is root finding.
- Open Domain vs. Closed Domain: A classic example of closed domain is the bisection method.
- Newton Raphson:

$$x_{n+1} = x_n - \underbrace{\nabla^2 f(x_n)}_{\text{matrix}} \cdot \underbrace{\nabla f(x_n)}_{\text{gradient}}$$

(1) Line Search:
   1. Pick a direction
   2. How long to wait

$$x_{n+1} = x_n + \alpha_n \vec{p}_n$$

(2) Trust-Region
   1. Pick a region to walk in
   2. Optimize within the region.

$$x_{n+1} = \underset{x \in \text{Region}(x_n)}{\arg\min}\ \tilde{f}$$

Example 3. *Say we want to have an iteration of the form*

$$x_{n+1} = x_n - g(x_n)$$

for some g. Clearly, the first requirement is that $g(x^*) = 0$. We also have $x_n - x^* = e_n$. Then

$$e_{n+1} = e_n - (g(x_n) - g(x^*))$$
$$e_{n+1} \approx e_n - \left(e_n g'(x^*) + \frac{1}{2} g''(x^*) e_n^2\right)$$
$$e_{n+1} \approx (1 - g'(x^*))e_n + \frac{1}{2} g''(x^*) e_n^2$$

This implies that $g(x^*) = 0$ and $g'(x^*) = 1$.

4.2. **Termination.** When do we stop?

(1) $F(x_{n+1})$ close to $F(x_n)$. In the case of a multivariate function, we can check

$$\frac{\|F(x_{n+1})\|}{\|F(x_n)\|}$$

(2) x_{n+1} close to x_n

$$\|x_{n+1} - x_n\|$$

We can also use

$$\underbrace{\|x+n=x^*\|}_{\text{distance to solution}} \leq \underbrace{K(\nabla F(x^*))}_{\text{condition number}} \cdot \underbrace{\|x_0-x^*\|}_{\text{initial bias}} \underbrace{\frac{\|F(x_n)\|}{\|F(x_0)\|}}_{\text{function value around } x^*}$$

or

$$\|F(x_{n+1})\| \leq \underbrace{T_L}_{\text{relative error}} \cdot \|F(x_0)\| + \underbrace{T_O}_{\text{tolerance}}$$

(this is how `optim` terminates)

4.3. **Newton-Raphson.** Newton-Raphson is Quadratic.

$$x_{n+1} = x_n - \nabla F(x_n)^{-1} F(x_n)$$

This is expensive since matrix inversion is on squared or cubic order

4.4. **Quasi-Newton.** Approximate $\nabla^2 f$ or ∇F, by B_n, which is simpler and inverts faster.

$$x_{n+1} = x_n - B_n^{-1} F(x_n)$$

For smooth (twice differentiable), continuous functions we have

$$\nabla^2 f(x_n)(x_{n+1}-x_n) \approx \nabla f(x_{n+1}) - \nabla f(x_n)$$

This implies that

$$B_n(x_{n+1}-x_n) = Df(x_{n+1} - Df(x_n)$$

We have

$$B_{n+1} = \arg\min_B \|B - B_n\|$$

such that $B = B'$ and B satisfies secant and $(x_2, \ldots, x_{n+1})$ (BFGS).

$$B_{n+1} = B_n - \frac{B_n s_n s_n' B_n}{s_n' B_n s_n} + \frac{y_n y_n'}{y_n' s_n}$$

where $s_n = x_{n+1} - x_n$ and $y_n = Df(x_{n+1}) - Df(x_n)$. Note that

$$B_{n+1} s_n = y_n$$

as

$$\begin{aligned} B_{n+1} s_n &= B_n s_n - \frac{B_n s_n s_n' B_n s_n}{s_n' B_n s_n} + \frac{y_n y_n' s_n}{y_n' s_n} \\ &= y_n \end{aligned}$$

Using Woodbury's formula, we can get B_n^{-1}.

5. September 22nd, 2014

5.1. **LDA ("Latent Dirichlet Allocation").** Applications: text analysis, survey data, population genetics. Python (NLTK) has a bunch of packages that does pre-processing on documents.

Given $D = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$, a collection of documents, and assuming k topics (sub-populations), focusing on a particular vocabulary (that is performing the analysis on a certain number of words) of size V (around 5000 to 30000. We want to estimate $\beta_{B\times k}$ with the constraint

$$\sum_{v=1}^{V} \beta_{vk} = 1$$

Definition 4 (Term)**.** *Distinct words in the document*

Definition 5 (Word)**.** *An instance of a term: (e.g. if there are three instances of 'the' in the document, then we have three words)*

$$w_m = \text{document} = \{\text{words } w_{mi} : i = 1, \ldots, N_m\}$$

This takes a form similar to

$$\left\{ \underbrace{\text{the}}_{1} \underbrace{\text{class}}_{2} \cdots \underbrace{\text{cool}}_{N_m} \right\}$$

Word is a one-of-many vector

$$\begin{matrix}1\\2\\\vdots\\36\\\vdots\\V\end{matrix}\begin{pmatrix}0\\0\\\vdots\\1\\\vdots\\0\end{pmatrix}$$

Example 4 (NYT, LDA)**.**

1	1003	36 : 2, 77 : 1
2	578	$\cdots$
3	2017	$\cdots$
$\vdots$	$\vdots$	

5.2. **DGP for LDA (for 1 document).**

$$\begin{aligned} N &\sim \text{Pois}(\xi) \\ \theta &\sim \text{Dirichlet}_K(\alpha_1, \ldots, \alpha_K) \qquad \alpha_i = \alpha, \quad \forall i = 1, \ldots, K, \quad \alpha_i > 0 \end{aligned}$$

For $i = 1, \ldots, N$,

$$\begin{aligned} z_i &\sim \text{Mult}_K(\theta, 1) \\ w_i &\sim \mathbb{P}(w_i | z_i, \beta) \\ &= \text{Mult}_V(\beta z_i, 1) \end{aligned}$$

and so

$$\beta_{V\times K} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbb{P}(w|z_{ii}=1,\beta) & \cdots & \mathbb{P}(w|z_{ik}=1,\beta) \\ \vdots & \vdots & \vdots \end{bmatrix}$$

- Our random variables are $w_{11}, \ldots, w_{1N_1}, \ldots, w_{m1}, \ldots, w_{mN_m}$.
- The known constants are $K, N_1, \ldots, N_m$.
- The latent variables are $\theta_1, \ldots, \theta_m, z_{11}, \ldots, z_{1N}, \ldots, z_{m1}, \ldots, z_{mN_m}$.

$$\mathbb{P}(D|\alpha, \beta) = \prod_{m=1}^{M} \int_{\Theta_m} \mathbb{P}(\theta_m|\alpha) \prod_{i=1}^{N_i} \sum_{z_{m_i}} \mathbb{P}(z_{m_i}|\theta_m)\, \mathbb{P}(w_{mi}|z_{mi}, \beta)\, \mathrm{d}\theta_m$$

5.3. **Mixed Membership Stochastic Block Model.** $Y_{N\times N}$ such that $y(p,q) \in \{0,1\}$

5.3.1. *DGP (for whole data matrix Y).* for $p = 1, \ldots, N$.

$$\pi_p \sim \text{Dirichlet}_K(\alpha_1, \ldots, \alpha_K)$$

for $p = 1, \ldots, N$, for $q = 1, \ldots, N$

$$\begin{aligned} Z_{p\to q} &\sim \text{Mult}(\pi_p, 1) \\ Z_{p\leftarrow q} &\sim \text{Mult}(\pi_q, 1) \\ y_{pq} &\sim \text{Bern}(Z'_{p\to q} B Z_{p\leftarrow q}) \end{aligned}$$

The unknown constants are $\alpha_1, \ldots, \alpha_K$

for $p = 1, \ldots, N$ **do**
  for $Q = 1, \ldots, N$ **do**
    $Z_{p\to q} \leftarrow \text{Mult}(\pi_p, 1)$
    $Z_{p\leftarrow q} \leftarrow \text{Mult}(\pi_q, 1)$
    $y_{pq} \leftarrow \text{Bern}(Z'_{p\to q} B Z_{p\leftarrow q})$
  end for
end for

6.1. **Mixture of Isoform.**

6.2. **Latent Space Model.** This model is from a paper by Hoff et al. 2001 JASA. In the model, we observe an adjacency matrix $Y_{N\times N}$. The diagonals of the marix are all zeros and it is not necessarily symmetric. $Y_{ij} \in \{0,1\}$.

Suppose we have something called a social space (think of it as a graph). The connection between α_i and α_j is a function of the distance between the two points.

As an example, think about a classroom full of students and we consider the students' (x,y) coordinates. It is more likely that two students know each other if they are closer to each other.

$$\mathbb{P}(Y_{ij} = 1 | x_i, x_j, \alpha, \beta) = p_{ij}$$

and so

$$\mathbb{P}(Y_{ij} = 0 | x_i, x_j, \alpha, \beta) = 1 - p_{ij}$$

We can model this as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha - \beta\left(\frac{x_i' x_j}{\|x_i\|}\right)$$

which is known as the projection model. We can also write this as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha - \beta|x_i - x_j|$$

Then from this, we have that

$$p_{ij} = h(\alpha, \beta, x_i, x_j)$$

for some function h. The DGP would be

for $i = 1, \dots, N$ **do**
  for $j \neq i = 1, \dots, N$ **do**
    $Y_{ij} \sim \text{Bern}(h(\alpha, \beta, x_i, x_j)$
  end for
end for

where $p_{ij} = \mathbb{P}(Y_{ij} = 1 | \alpha, \beta, x_i, x_j)$. So we can rewrite the log-odds as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha - \beta\frac{x_i' x_j}{\|x_i\|} = y_{ij}$$

and we can set

$$p_{ij} = \frac{1}{1+e^{-\eta_{ij}}}$$
$$1 - p_{ij} = \frac{e^{-\eta_{ij}}}{1+e^{-\eta_{ij}}}$$

and so our likelihood is

$$\mathcal{L}(\alpha, \beta) = \prod_{\substack{i,j \\ i\neq j}} \int \text{Bern}\left(\frac{1}{1+e^{-\eta_{ij}}}\right) \, \mathrm{d}x_1 \cdots \mathrm{d}x_N$$

Taking logs, we get

$$\ell(\alpha, \beta) = \sum_{\substack{i,j \\ i\neq j}} \int (\eta_{ij} y_{ij} - \log(1 + e^{\eta_{ij}})) \, \mathrm{d}x_1 \cdots \mathrm{d}x_N$$

6.3. **Principal Component Analysis.** In PCA, we wish to reduce data from d dimensions into k dimensions where $d \gg k$. We do this by finding the projecting the data into various orthogonal components and then choosing the ones with maximal variance. The plot of the variances is called the **scree plot**.

$$\underbrace{y}_{d\times n} = \left[\ y_1 \mid y_2 \mid \cdots \mid y_n\ \right] = \underbrace{\Lambda}_{d\times k}\left[\ x_1 \mid x_2 \mid \cdots \mid x_n\ \right] + \varepsilon$$

From this, we get that

$$y_i = \Lambda x_i + \varepsilon$$

where $\varepsilon_i, y_i \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^k$.

$$x_i \sim \mathcal{N}_k(0, I_k)$$
$$\varepsilon_i \sim \mathcal{N}_d(0, \sigma^2 I_d)$$

This gives

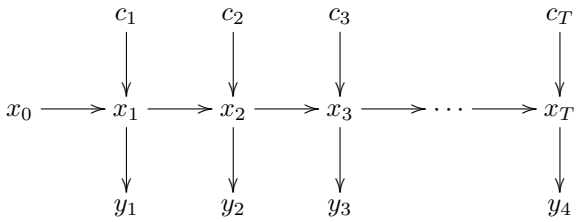
$$\begin{aligned}\mathbb{P}\big(y|\Lambda,\sigma^2\big) &= \int_X \mathbb{P}(x|0,I)\,\mathbb{P}(y|\Lambda,\sigma,x)\;\mathrm{d}x \\ &= \mathcal{N}\big(0,\underbrace{\Lambda}_{d\times k}\underbrace{\Lambda'}_{k\times d}+\sigma^2\underbrace{I}_{d\times d}\big)\end{aligned}$$

We also have

$$\lim_{\sigma\to 0}\mathbb{P}(x|y,\Lambda,\sigma)=\delta(\Lambda' y)$$

6.4. **State Space Models.**

- Observations: $y_1,\ldots,y_T$, where y_i is either a scalar of a vector. $c_1,\ldots,c_T$ which are controls
- Latent Variables: $x_0,x_1,\ldots,x_T$, which is called the state space



and so

$$\begin{cases} x_t = f_x(x_{t-1},\ldots) \\ y_t = f_y(x_t)\end{cases} \qquad \begin{cases} x_t = f_x(x_{t-1}c_t,\ldots) \\ y_t = f_y(x_t)\end{cases}$$

We need to distinguish between linear and non-linear and also need to distinguish between Gaussian vs non-Gaussian.

- Observation: $y_t = C_{x_t} + v_t$
- State: $x_t = A_{x_{t-1}} + w_t$

where $t \geq 1$ and

$$\begin{aligned} x_0 &\sim \mathcal{N}(0,V) \\ w_t &\sim \mathcal{N}(0,Q) \\ v_t &\sim \mathcal{N}(0,R)\end{aligned}$$

with $w_t \perp\!\!\!\perp v_t$.

Let $\theta=(C,A,V,R,Q)$. Making inference about

$$\mathbb{P}(x_t|y_1,\ldots,y_t,\theta)$$

is called filtering.

$$\mathbb{P}(x_t|y_1,\ldots,y_t,y_{t+1},\ldots,y_T,\theta)$$

is called smoothing.

$$\mathbb{P}(x_t|y_1,\ldots,y_{t-1},\theta)$$

is called prediction.

Using this, we have

$$\mathbb{P}(x_0,x_1,\ldots,x_T,y_1,\ldots,y_T|\theta)=\left[\prod_{t=1}^{T}\mathbb{P}(x_t|x_{t-1})\,\mathbb{P}(y_t|x_t)\right]\mathbb{P}(x_0)$$

Example 5 (GPS data)**.** *Let*

$$x_t=\begin{bmatrix}x_1\\x_2\\\dot{x}_1\\\dot{x}_2\end{bmatrix} \qquad y_t=\begin{bmatrix}x_1\\x_2\end{bmatrix}$$

where (x_1,x_2) *is the position and* $(\dot{x}_1,\dot{x}_2)$ *is the velocity. Then*

$$x_t=\begin{bmatrix}x_1\\x_2\\\dot{x}_1\\\dot{x}_2\end{bmatrix}_t=\begin{bmatrix}1&0&1&0\\0&1&0&1\\0&0&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\dot{x}_1\\\dot{x}_2\end{bmatrix}_{t-1}+w_t$$

and

$$y_t = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + v_t$$

7. September 29th, 2014

7.1. **Odyssey.** Important links:

- `https://rc.fas.harvard.edu`
- Go to `https://account.rc.fas.harvard.edu/password_reset/`
- `https://rc.fas.harvard.edu/resources/odyssey-quickstart-guide/`
- `github.com/airoldilab`
- The `.ssh/config` file allows us to login to the cluster much easier (that is without logging in and typing everything)
- Use `mkdir` in `n/regal/stats`. Do not store your files in here. Store them in your home directory.
- Slides: `http://airoldilab.github.io/odyssey/`

8. October 1st, 2014

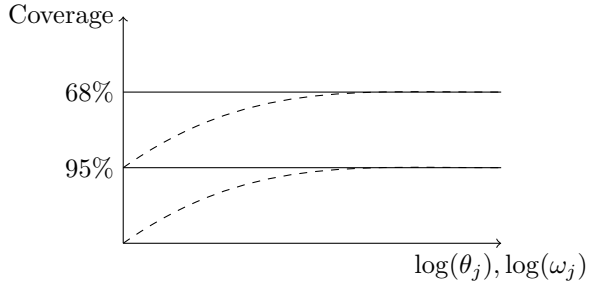
$$\begin{aligned} \mathbb{P}(\mu, \sigma^2) &\propto \frac{1}{\sigma^2} \\ \log(\theta_j) &\sim \mathcal{N}(\Gamma, \sigma^2) \qquad j = 1, \ldots, J \\ y_{nj} &\sim \text{Pois}(\omega_j \theta_j) \qquad j = 1, \ldots, J,\ n = 1, \ldots, N \end{aligned}$$

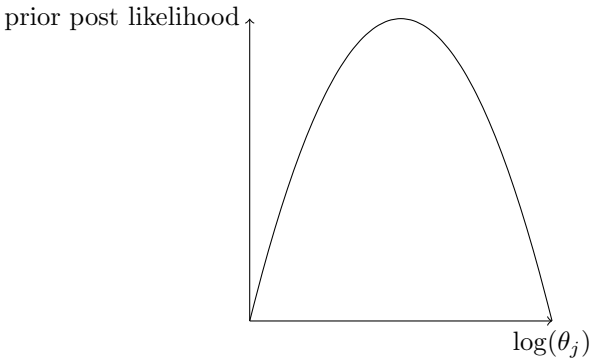
Later on in the homework, we want to graph the coverage vs $\log \theta_j$ or $\log w_j$, where the coverage is the the y-axis.

$$Y_{2\times 1000} \longrightarrow [\text{lower}, \text{upper}] \ni \log \theta_j \ j = 1, \ldots, J$$

8.1. **Procedure.**

- True Value: $\log \theta_j = 3$
- Generate $y^{(1)}, \ldots, y^{(1000)}$
- For each $y^{(i)}$, feed an MCMC chain to form a kernel density estimate





where the left hand side is the prior and the righthand side is 1/posterior.

Univariate	Multivariate
Beta-Binomial	Dirichlet_l-Multinomial_k
Posterior mean is MLE on $N + 2$ obs	Posterior mean is MLE on $N + k$ obs
Maximum a priori? is the MLE with uniform prior	Maiximum a priori? is not the MLE with the uniform prior

8.2. **Multinomial.** Suppose we have n, k, which are positive integers, and $(p_1, \dots, p_k)'$ such that $p_i \in [0,1]$ for $i = 1, \dots, k$ and $\sum_{i=1}^k p_i = 1$.

$$X \sim \text{Mult}_k\left((p_1, \dots, p_k)', n\right)$$

where $X = (X_1, \dots, X_k)'$ such that $\sum_{i=1}^k X_i = n$ and $x_i \in \{0, 1, \dots, n\}$. So our likelihood is

$$\begin{aligned}\mathcal{L}_X(x_1, \dots, x_k | p_1, \dots, p_k, n) &= \binom{n}{x_1 \cdots x_k} \prod_{i=1}^k p_i^{x_i} \\ &= n! \prod_{i=1}^k \frac{p_i^{x_i}}{x_i!}\end{aligned}$$

We say that $p \in \text{Simplex}_k$, which is a $(k-1)$-dimensional space.

8.2.1. *Note: MLE for p.* Suppose we have $x_1, \dots, x_m \overset{iid}{\sim} \text{Mult}((p_1, \dots, p_k), n)$. Then

$$\mathcal{L}(p) = \prod_{i=1}^m \left(n! \prod_{j=1}^k \frac{p_i^{x_{ij}}}{x_{ij}!} \right)$$

To get the MLE, we have two options

- Set $p_k = 1 - \sum_{j=1}^{k-1} p_j$
- $L = \mathcal{L}(p) - \lambda\left(1 - \sum_{j=1}^k p_j\right)$. This method is much easier to do.

8.3. **Dirichlet Distribution.** The dirichlet distribution is a generalization of $\text{Beta}(\alpha, \beta)$. If we have $p \sim \text{Beta}(\alpha, \beta)$, then $p \in [0,1]$. This is a very rich family of distributions that can model concave and convex functions.

$$\binom{\theta_1}{\theta_2} \sim \text{Dirichlet}(\alpha, \beta)$$

Now, if we look at $1 - p$, then we have that

$$\binom{\theta_1}{\theta_2} \in \text{Simplex}_2 \equiv \{(\theta_1, \theta_2)\theta_1 \in [0,1], \theta_2 \in [0,1], \theta_1 + \theta_2 = 1\}$$

Using this, we can in fact write $\theta_2 = 1 - \theta_1$. Then

$$p \sim \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1-p)^{\beta - 1}$$

and so if

$$\begin{aligned}\binom{\theta_1}{\theta_2} &\sim \text{Dirichlet}(\alpha_1, \alpha_2) \\ &= \frac{\Gamma\left(\sum \alpha_i\right)}{\prod \Gamma(\alpha_i)} \prod \theta_i^{\alpha_i - 1}\end{aligned}$$

In the Gamma distribution, we had

$$\begin{aligned}\mathbb{E}[p] &= \frac{\alpha}{\alpha + \beta} \\ \mathbb{E}[1 - p] &= 1 - \mathbb{E}[p] = \frac{\beta}{\alpha + \beta}\end{aligned}$$

and equivalently in the Dirichlet distribution, we have

$$\mathbb{E}[\alpha_i] = \frac{\alpha_i}{\sum_i \alpha_i}$$

Suppose we have $X_1, \dots, X_N \sim \text{Mult}(p, n)$ where $p = (p_1, \dots, p_k)'$. Then if $p \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$, then

$$\begin{aligned}f(p | x_1, \dots, X_N &\propto \left(\prod_{i=1}^k p_i^{\alpha_i - 1} \right) \prod_{j=1}^N \left(\prod_{i=1}^k p_i^{x_j} \right) \\ &= \prod_{i=1}^k p_i^{\left(\sum_j x_{ij} + \alpha_i\right) - 1} \\ &\propto \text{Dirichlet}\left(\sum_{j=1}^N x_{1j} + \alpha_1, \dots \sum_{j=1}^N x_{kj} + \alpha_k \right)\end{aligned}$$

The posterior mean for $i = 1, \dots, k$.

$$\mathbb{E}[p_i | x_1, \dots, X_N] = \frac{\alpha_i + \sum_{j=1}^N x_{ij}}{\sum_{i=1}^k \alpha_i + nN}$$

The posterior mode is

$$\arg\max f(p|x_1, \dots, x_N) = \frac{\alpha_i + \sum_{j=1}^N x_{ij} - 1}{\sum_{i=1}^k \alpha_i + nN - 2}$$

and so

$$\hat{p}_{MLE}(x_1, \dots, x_N) = \frac{\sum_{j=1}^N x_{ij}}{nN}$$

The posterior mean, if $\alpha_1 = \cdots = \alpha_k = 1$ is

$$\frac{1 + \sum_{i=1}^N x_{ij}}{k + nN}$$

Suppose now we add k observations $(1, 0, 0, \dots, 0), (0, 1, 0, 0, \dots, 0), (0, 0, 1, 0, 0, \dots, 0)$ and so forth, we have the MLE for $N + k$ observations is

$$\hat{p}_{MLE}(x_1, \dots, x_N, x_{N+1}, \dots, x_{N+k}) = \frac{1 + \sum x_{ij}}{k + nN}$$

Using the fact that

$$f(p) f(x_1, \dots, x_n | p) \propto f(p | x_1, \dots, x_n)$$

If $f(p) = 1$, then $\hat{p}_{MLE} = \hat{p}_{MAP}$ since the value of p that maximizes the left hand side would maximize the right hand side.

Again, if we set $\alpha_1 = \cdots = \alpha_k = 1$, then the posterior mode is

$$\frac{\sum_{j=1}^N x_{ij}}{nN + (k-2)}$$

9. October 6th, 2014

Facebook thing on Wednesday. No Section on Friday.

9.1. **Stochastic Approximation (Robbins& Monro, 51).** Goal: Solve $M(\theta^*) = 0$ (M might be unknown) when only y_{θ_n} random variables are available such that $\mathbb{E}[Y_{\theta_n}] = M(\theta_n)$. Fix θ. Then $\mathbb{E}[Y] = M(\theta)$. If we set

$$\theta_{n+1} = \theta_n - \alpha_n Y_{\theta_n}$$

such that

(1) $\sum_i \alpha_i = \infty$
(2) $\sum_i \alpha_i^2 < \infty$
(3) Bounded $Y_{\theta_n}^2$, increasing $M(\theta)$

then $\theta_n \overset{\text{quadratic mean}}{\longrightarrow} \theta^*$, which means

$$\mathbb{E}\big[(\theta_n - \theta^*)^2\big] \longrightarrow 0$$

One such sequence that satisfies this is $\frac{\alpha}{i}$.

Sketch of Proof.

$$(\theta_{n+1} - \theta^*)^2 = (\theta_n - \theta^*)^2 - 2\alpha_n(\theta_n - \theta^*) Y_{\theta_n} + \alpha_n^2 Y_{\theta_n}^2$$

Taking expectation, we get

$$\mathbb{E}\big[(\theta_{n+1} - \theta^*)^2\big] = (\theta_n - \theta^*)^2 - 2\alpha_n(\theta_n - \theta^*) M'(\theta_n) + \alpha_n^2 \mathbb{E}\big[Y_{\theta_n}^2 \big| \theta_n\big]$$

and so

$$\begin{aligned} b_{n+1} &= b_n - 2\alpha_n \mathbb{E}[\theta_n - \theta^*]\, M'(\theta_n) + \alpha_n^2 \mathbb{E}\big[Y_{\theta_n}^2\big] \\ &\approx b_n - 2\alpha_n \mathbb{E}[\theta_n - \theta^*]\, M'(\theta_n)(\theta_n - \theta^*) + \alpha_n^2 (\mathbb{E}\big[Y^2\big] + \cdots) \\ b_{n+1} &= (1 - 2\alpha_n M'(\theta^*)) b_n + \alpha_n^2(\cdots) \qquad \text{as } b_n \longrightarrow 0 \end{aligned}$$

What was shown later (not in the paper) was that

(1) $b_n \longrightarrow 0$
(2) $\left(\frac{1}{\alpha_n}\right) b_n \longrightarrow \frac{\mathbb{E}\left[Y^2 | \theta\right]}{2\alpha M'(\theta^*) - 1}$ which is the rate of convergence

□

9.2. **Statistical Estimation (Sakerson?, 65).**

- $Y_{\theta_n} = -\ell'(y_n;\theta_n)$ where $y_n \overset{iid}{\sim} f(\theta^*)$ and θ^* is an unknown model parameter.
- $\theta_{n+1} = \theta_n + \alpha_n \ell'(y_n;\theta_n)$
- $M(\theta_n) = \mathbb{E}_Y[\ell'(y_n;\theta_n)]$. Note that $M(\theta^*) = \mathbb{E}_{Y\sim f(\theta^*)}[\ell'(y_n;\theta^*)] = 0$. This is an identity under regularity conditions.

This is called recursive estimation/stochastic gradient descent. What is key here is that once you get θ_{n+1}, you can throw out θ_n. That is, θ_n converges to the value where $\ell' = 0$, that is when $\theta_n = \theta^*$. In multiple parameters, this changes to

$$\theta_{n+1} = \theta_n + \alpha_n \nabla \ell(y_n;\theta_n)$$

Key things that make this simpler is

(1) α_n is just one number in $\mathbb{R}^+$.
(2) We compute ∇ at just one data point.

In optimization, the goal is to go from θ_0 to θ^{MLE} as fast as possible. However, statisticians care about the variance, that is it must be close across multiple datasets. There are many variables that we can change which will affect the algorithm

(1) Choice of α_n. Intuition? $\alpha_n = \frac{\alpha}{n}$ is a weight of the previous statistics.
(2) Choice of statistic. $\nabla\ell(y_n;\theta_n)$ is a function of the data.
(3) Parameterization.

Why are the weights different across observations? If we have $y_1,\dots,y_n \sim \mathcal{N}(\mu,\sigma^2)$, the MLE for the mean is an equal weight of all of the y_i's.

Example 6. *Suppose $y_i \sim \mathcal{N}(\mu,\sigma^2)$ with σ^2 fixed. Let μ_n be an estimate at n.*

$$\begin{aligned}\mu_{n+1} &= \mu_n + \alpha_n(y_n - \mu_n)\\ &= (1-\alpha_n)\mu_n + \alpha_n y_n\end{aligned}$$

sample	weight
y_n	α_n
y_{n-1}	$(1-\alpha_n)\alpha_{n-1}$
y_{n-2}	$(1-\alpha_n)(1-\alpha_{n-1})\alpha_{n-2}$
$\vdots$	$\vdots$

If $\alpha_n = \frac{\alpha}{n}$ and we set $\alpha = 1$, then all the weights are equal ($\mu_n = \bar{y}_n$). This is important, since if we don't have uniform weights, we have information loss. In 1950, Sacks showed that

$$\mathrm{Var}(\mu_n) = \underbrace{\frac{\sigma^2}{n}}_{MLE} \times \overbrace{\frac{\alpha^2}{2\alpha-1}}^{loss\ from\ learning\ rate}$$

It is important to note that the α must be finely tuned so that the learning rate is not too inaccurate.

Example 7. *Let $\phi_n = \mu_n^3$. Then we have*

$$\begin{aligned}\phi_{n+1} &= \phi_n + \alpha_n(y_n - \phi_n^{4/3})\\ &= \phi_n + \alpha_n(y_n - f(\phi_n))\end{aligned}$$

Then again by Sacks,

$$\mathrm{Var}(\phi_n) = \frac{\alpha^2}{2\alpha f'(\phi^*) - 1}\frac{\sigma^2}{n}$$

If we revert back using the transformation theorem, then

$$\mathrm{Var}(\mu_n) = \frac{\alpha^2 f'(\phi^*)^2}{2\alpha f'(\phi^*) - 1} \times \frac{\sigma^2}{n}$$

which is greater than 1, so we must use the natural parameterization.

One of the main reasons why statisticians stay away from the aforementioned methods is the lack of stability. Consider the following example:

Example 8. *Suppose* $y_i \sim \text{Pois}(e^\lambda)$. *Let*

$$\lambda_{n+1} = \lambda_n + \alpha_n(y_n - e^{\lambda_n})$$

Start with $\lambda_1 = 0, \alpha_1 = 1, y_1 = 100$. *Then* $\lambda_2 \approx 100$. *Suppose* $y_2 = 221, \alpha_2 = \frac{1}{2}$. *Then*

$$\lambda_3 = 100 + \frac{1}{2}(221 - e^{100})$$

9.2.1. *Proximal methods/Implicit SCSD.*

$$\theta_{n+1} = \theta_n + \alpha_n \nabla \ell(y_n, \theta_{n+1})$$

Note the θ_{n+1} in the log-likelihood. That can be solved using

$$\min_\theta \left\{ \frac{1}{2}\|\theta_n - \theta\|^2 - \ell(y_n; \theta) \right\}$$

Going back the poisson example, we will have

$$\lambda_1 = (100 - e^{\lambda_1})$$

and so $\lambda_1 \approx \log 100$.

$$\lambda_2 = \lambda_1 + \frac{1}{2}(221 - e^{\lambda_2})$$

In this case, the convergence is a bit slower, but it guarantees θ_{n+1} is in the parameter space.

10. October 15th, 2014

10.1. **Stochastic Gradient Descent.** If we have $\nabla \ell(\theta; y)$, and $y_n \overset{iid}{\sim} f(\theta^*)$, then fitting is "easy".

$$\theta_{n+1} = \theta_n + \alpha_n \nabla \ell(\theta_n; y_n)$$

Typically, we set $\alpha \propto \frac{1}{n}$, however the proportionality constant is very important for two reasons:

(1) Stability (variance)
(2) Convergence (bias)

10.2. **GLM Model.** Our setup:

$$\begin{aligned} X_n &\sim G \qquad \text{(covariates) "random effects"} \\ \lambda_n &= X_n'\theta^* \qquad \text{linear predictor} \\ y_n|\lambda_n &\sim \exp\{\lambda_n y_n - b(\lambda_n) + \cdots\} \end{aligned}$$

We have that

$$\ell(\theta; y_n) = \lambda_n y_n - b(\lambda_n) + \cdots$$

which is sometimes called the canonical representation. The gradient of this function is then

$$\nabla \ell(\theta; y_n) = \{y_n - b'(\lambda_n)\}\vec{x}_n$$

Note that this has expected value of 0. We can then plug this into our iteration:

$$\theta_{n+1} = \theta_n + \alpha_n \{y_n - b'(x_n'\theta_n)\}\, \vec{x}_n$$

where y_n is the observed data point and $b'(x_n'\theta_n)$ is the expected y_n if $\theta^* = \theta_n$.

10.2.1. *Logistic Regression.* We have$Y_n \in \{0, 1\}$. Then we model

$$\mathbb{P}(Y_n = 1|X_n, \theta) = \frac{\exp(x_n'\theta)}{1 + \exp(x_n'\theta)} \in [0, 1]$$

10.2.2. *Poisson Regression.* The setup here is $Y_n \in \mathbb{Z}^+$. It can be used to model bookings, events, etc.

$$Y_n|X_n, \theta \sim \text{Pois}(e^{x_n'\theta})$$

10.3. **Problem 2 in Assignment 3.** Assume

$$\begin{aligned} X_n &\sim \mathcal{N}(0, A) \\ \lambda_n &= x_n'\theta^* \\ y_n|\lambda_n &\sim \mathcal{N}(\lambda_n, \sigma^2) \end{aligned}$$

Then $\vec{y} = X\theta^* + \vec{\varepsilon}$. Other assumptions are that $p = 100$, which is the number of parameters, $\theta^* = \vec{1}$, and $\sigma^2 = 1$.

The evaluation is $(\theta_n - \theta^*)'A(\theta_n - \theta^*) \geq 0$, which is the risk.

10.3.1. *Model.*

$$\begin{aligned} \mathbb{P}(y_n|x_n, \theta) &\propto \exp\left\{-\frac{1}{2}(y_n - x_n'\theta)^2\right\} \\ \ell(\theta, y_n, y_n) &= -\frac{1}{2}(y_n - x_n'\theta)^2 \\ \nabla\ell(\theta, y_n, x_n) &= (y_n - x_n'\theta)\cdot\vec{x}_n \end{aligned}$$

So, we can do the SGD as follows:

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_n(y_n - x_n'\theta)\cdot\vec{x}_n \\ &= (I - \alpha_n \underbrace{x_n x_n'}_{p\times p})\theta_n + \alpha_n y_n\vec{x}_n = O(p^2) \\ &= \{\theta_n - \alpha_n x_n(x_n'\theta_n)\} + \alpha_n y_n\vec{x}_n = O(p) \end{aligned}$$

and so this is the SGD for the Normal. If we did this implicitly, we have

$$\theta_{n+1} = \theta_n + \alpha_n(y_n - x_n'\theta_{n+1})\vec{x}_n$$

Note that we can write $x_n'\theta_{n+1}x_n$ as $(x_n x_n')\theta_{n+1}$. If a, b, c are vectors, then

$$(a'b)c = (ca')b$$

since $(a'b)$ is simply a scalar and we can shift it to the other side, so

$$\begin{aligned} (a'b)c &= c(a'b) \\ &= (ca')b \end{aligned}$$

Thus we have

$$\begin{aligned} \theta_{n+1} + \alpha_n(x_n x_n')\theta_{n+1} &= \theta_n + \alpha_n y_n x_n \\ (I + \alpha x_n x_n')\theta_{n+1} &= \theta_n + \alpha_n y_n x_n \\ \theta_{n+1} &= (I + \alpha_n x_n x_n')^{-1}(\theta_n + \alpha_n y_n x_n) \end{aligned}$$

Typically the inverse of a matrix is very computationally expensive (slightly less than $O(n^3)$). However, we can use the Sherman-Morrison formula, which is a simplified version of Woodbury's formula.

When doing the assignment, we can get the MLE by using the `lm` function. We can then pick a learning rate that is either $\alpha_n = \frac{\gamma_0}{1+\gamma_0\lambda_0 n}$ or $\alpha_n = \frac{\alpha}{\alpha+n}$.

11. October 20th, 2014

11.1. **Monte Carlo Integration.**

- Non-iterative techniques: importance sampling (IS), generalized rejection sampling, bootstrap
- Iterative methods: MCMC (Gibbs sampler, Metropolis-Hastings)
- Sequential models: particle filters
- Strategies: blocked (e.g. in a mixture model of Normals, we have the parameters for the Normals, then the Multinomial parameters), collapsed (integrate out and then sample over the remaining ones), sliced, ...

Example 9. *Suppose we have* $\int\mathbb{P}(y, x_1, x_2|\theta)\,\mathrm{d}x_1\mathrm{d}x_2$.
- *Blocked:* $\mathbb{P}(x_1|y, x_2, \theta)$*, then* $\mathbb{P}(x_2|y, x, \theta)$
- *Collapsed:* $\mathbb{P}(x_1|y, \theta)$*, then* $\mathbb{P}(x_2|y, \theta)$

Notes:

(1) $\int f(x)\,\mathrm{d}x$, $f(x)\alpha$.
(2) Can you sample from $f(x)$ easily?
(3) Integrand

Example 10. *Suppose we have an image, which has a black background and some arbitrarily shaped object which is red. We want to compute the area of the object. Let $h(x)$ be 1 if the image at x is red and 0 otherwise. This is a simple function, but it's much harder to find*

$$\int_x h(x)\,\mathrm{d}x$$

h is difficult when the following are not easily done:

- Simpson's rule
- Trapezoidal rule
- Gaussian quadrature

In that case we must do MC.

Example 11 (Finding the area of a circle)**.** *Suppose we have a unit circle centred at $(\frac{1}{2}, \frac{1}{2})$. Then we can take $x_{11}, \ldots, x_{1,10000} \in [0,1]$ and $x_{21}, \ldots, x_{2,10000} \in [0,1]$. Then we have*

$$\begin{aligned}\int \mathbb{1}_{\text{inside circle}}\,\mathrm{d}x_s\mathrm{d}x_s &= \#\{\text{inside circle}\} \\ &= \#\left\{(x_1,x_2) : \left(x_1 - \frac{1}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^2 \le 1\right\} \\ &\approx 0.78341 \pm ? \qquad \text{(we will get the sd later)}\end{aligned}$$

Example 12 (Tail probabilities/p-values)**.** *Let*

$$T = \sum_{i,j} d_{ij}$$

Ideally, we would like to find f_T given that $(x_1, x_2) \sim \text{Uniform}(0,1)^2$. If we had the density, then we could simply find $\mathbb{P}(T > c)$. From a Monte Carlo integration perspective, we would take $x_1^{(1)}, \ldots, x_n^{(1)}$ and compute $t^{(1)}$. Then we would take $x_1^{(1)}, \ldots, x_n^{(2)}$ and compute $t^{(2)}$. We keep doing this B times when we have $t_1^{(B)} \ldots, t_n^{(B)}$ and then compute $t^{(B)}$. We can take our t samples and compute $\hat{f}_T$. This approximates

$$\mathbb{P}(T > T_{obs}) = \int_{T_{obs}}^{\infty} f_T(t)\,\mathrm{d}t$$

In the Monte Carlo case, we have

$$\sum_{i=1}^{B} \mathbb{1}_{t^{(i)} > T_{obs}} \cdot \frac{1}{B} \longrightarrow \int \mathbb{1}_{t > T_{obs}} f_T(t)\,\mathrm{d}t$$

In general, for $x_1, \ldots, x_n \sim F$, we want to esimate $\theta(F)$. We use $T(x_1, \ldots, x_n)$ as an estimator. We have two choises: we can assume a parametric model for f or a nonparametric model.

In the previous example. we have $F = \text{Uniform}(0,1)^2$. Then, givne $t^{(1)}, \ldots, t^{(B)}$, we have

$$\theta(\hat{F}) \longrightarrow \hat{\theta}(F)$$

Or, we could think of $\hat{F}$ as

$$\hat{F}(x) = \left(\frac{\#\{x_i \le x\}}{n}\right)$$

11.2. **Importance Sampling.** Suppose we wish to find

$$I = \int_a^b h(x)\,\mathrm{d}x$$

where $h(x)$ is not necessarily a pdf. We can rewrite I as

$$I = \int_a^b \left(\frac{h(x)}{f(x)}\right) f(x)\,\mathrm{d}x$$

Define $\omega(x) = \frac{h(x)}{f(x)}$. Then

$$I = \int_a^b \omega(x) f(x)\,\mathrm{d}x = \mathbb{E}_f\left[\omega(x)\right]$$

Then by CLT, we have

$$\hat{I} = \sum_{i=1}^{B} \frac{\omega(x_i)}{B} \longrightarrow \int \omega(x) f(x)\,\mathrm{d}x = I$$

if $x_1, \ldots, x_B \sim f$. The standard error is

$$\hat{\mathrm{SE}} = \frac{\hat{S}^2}{\sqrt{B}}$$

where

$$\hat{S}^2 = \sum_{i=1}^{B} \frac{(\omega(x_i) - \hat{I})^2}{B}$$

Then the $(1-\alpha)$ confidence interval for I is

$$\hat{I} \pm z_{\alpha/2}\hat{\mathrm{SE}}$$

Example 13. *Let $h(x) = x^3$. We have*

$$I = \int_0^1 x^3\,\mathrm{d}x = \frac{1}{4}$$

In Edo's simulation, he had

$$\hat{I}_{10000} = 0.248 \qquad \hat{SE} = 0.0028$$

on Uniform$(0,1)$.

$$I = \int \underbrace{h(x)}_{\text{arbitrary}}\ \underbrace{g(x)}_{\text{density}}\ \mathrm{d}x = \mathbb{E}_g\left[h\right]$$

If $x_1, \ldots, x_B \sim g(x)$, then

$$\hat{I} = \sum_{i=1}^{B} \frac{h(x_i)}{B} \longrightarrow I$$

Now if we have

$$I = \int \frac{h(x)g(x)}{f(x)} f(x)\,\mathrm{d}x$$

with $\omega(x) = \frac{h(x)g(x)}{f(x)}$ and $x_1, \ldots, X_B \sim f(x)$, then

$$\hat{I} = \sum_{i=1}^{B} \frac{\omega(x_i)}{B} \longrightarrow I$$

Now, consider

$$\mathbb{E}_f\left[\omega(w)^2\right] = \int \left(\frac{h(x)g(x)}{f(x)}\right)^2 f(x)\,\mathrm{d}x$$

We wish that $f(x) > g(x)$ since otherwise we will not have good efficiency. This is most important at the tails, where we do not want the weights to explode.

Theorem 1. *$f(x)$ that minimizes* $\mathrm{Var}(\hat{I})$ *is*

$$f(x)^* = \frac{|h(x)|g(x)}{\int |h(s)|g(s)\,\mathrm{d}s}$$

Example 14 (Tail probabilities)**.**

$$I = \mathbb{P}(Z > 3)$$

where $Z \sim \mathcal{N}(0,1)$.

$$h(x) = \begin{cases} 0 & X > 3 \\ 1 & \textit{otherwise} \end{cases} = \mathbb{1}_{x>3}$$

First, we sample $x_1, \dots, x_B \sim \mathcal{N}(0,1)$ with $B = 100$. Then, we have

$$\hat{I} = \sum_{i=1}^{100} h(x_i)$$

Repeated $B = 100$ times. Then

$$\mathbb{E}\left[\hat{I}\right] = 0.0015$$
$$\mathrm{Var}(\hat{I}) = 0.0039$$

Then, we pick $f \sim \mathcal{N}(4,1)$ and so

$$I = \int \frac{h(x)g(x)}{f(x)} f(x)\,\mathrm{d}x$$

where $\omega(x) = \frac{h(x)g(x)}{f(x)}$ as before.

$$\mathbb{E}\left[\hat{I}\right] = 0.0011$$
$$\mathrm{Var}(\hat{I}) = 0.0020$$

12. October 22nd, 2014

Example 15 (Outliers)**.** *Suppose we have $X_1, \dots, X_n$ such that $X_i = \theta + \varepsilon_i$ where θ is of interest. A simple model is*

$$X_i \sim \mathcal{N}(\theta, 1)$$
$$\varepsilon_i \sim \mathcal{N}(0,1)$$

Effectively,

$$\varepsilon_i \sim t_3$$
$$\theta \propto \textit{constant}$$

So, the likelihood is

$$\begin{aligned}\mathbb{P}(x_1, \dots, x_n|\theta) &= \prod_{i=1}^{n} t_3(x_i - \theta) \\ &= \mathcal{L}(\theta)\end{aligned}$$

We want to have

$$\begin{aligned}\mathbb{E}[\theta|x_1, \dots, x_n] &= \bar{\theta} \\ &= \frac{\int \theta \mathcal{L}(\theta)\,\mathrm{d}\theta}{\int \mathcal{L}(\theta)\,\mathrm{d}\theta}\end{aligned}$$

Recall that importance sampling, we sample $\theta_1, \dots \theta_n \sim g$. Then we have

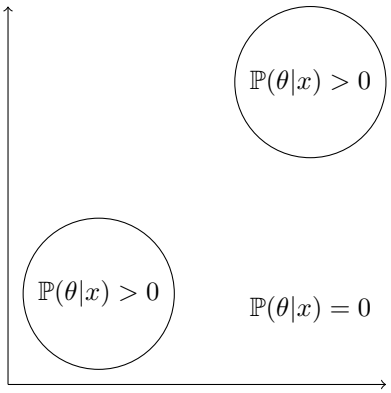
$$\bar{\theta} = \frac{\frac{1}{N}\sum_{i=1}^{n} \theta_i \frac{\mathcal{L}(\theta_i)}{g(\theta_i)}}{\frac{1}{N}\sum_{i=1}^{n} \frac{\mathcal{L}(\theta_i)}{g(\theta_i)}} \longrightarrow \mathbb{E}[\theta|x_1, \dots, x_n]$$

In Edo's code, we got that $\bar{\theta} = -0.64$ with $n = 2$. Using $g \sim \mathcal{N}(0,1)$, he got $\bar{\theta} = -0.074$. Using $g \sim t_1$, he got $\bar{\theta} = -0.53$.

12.1. **MCMC.**

- Metropolis-Hastings
- Gibbs sampler
- Metropolis-Hastings in GIbbs

Suppose we have a posterior with a really nasty support (only inside the circles)



Suppose we have

$$I = \int h(\theta_1, \theta_2)\mathbb{P}(\theta_1, \theta_2|x_1, \ldots, x_n)\ \mathrm{d}\theta_1\mathrm{d}\theta_2$$

In this case a Gibbs sampler would not be able to explore the circle opposite the one it starts from. It is important to know that the geometry of the distribution determines how well a given method works.

For

$$I = \int h(x)f(x)\,\mathrm{d}x$$

the idea is to set up a Markov chain $x^{(1)}, x^{(2)}, \ldots$ with "stationary" distribution f.

Theorem 2. *An irreducible, ergodic (recurrent and aperiodic) Markov chain has a unique stationary distribution $f(x)$. The limit distribution exists and equals $f(x)$.*

If h is bounded, then

$$\lim_{B\to\infty} \frac{1}{B}\sum_{i=1}^{B} h(x_i) \longrightarrow \mathbb{E}_f\left[h\right] = I$$

"Transition Kernel": $\mathbb{P}(X_t|X_{t-1})$.

$$\{X_t : t = \underbrace{0, 1, \ldots, m}_{\text{"burn-in"}}, m+1, \ldots, B\}$$

We will typically throw away the first m samples, which is calle the "burn-in" and we take

$$\hat{I} = \sum_{i=m+1}^{B} \frac{h(x_i)}{B-m}$$

12.2. **Metropolis-Hastings Monte Carlo Markov Chain (MH-MCMC).**

- Pick a starting point $X^{(0)}$.
- Generate a candidate $y \sim q(y|x^{(i)})$ at iteration i.
- Evaluate the rejection probability

$$r(x^{(i)}, y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}\right\}$$

- Sample $U \sim \text{Uniform}(0, 1)$ and set

$$\begin{cases} X^{(i+1)} = y & U < r \\ X^{(i+1)} = X^{(i)} & U \geq r \end{cases}$$

Usually, we have $q(y|x^{(i)}) = \mathcal{N}(y|x^{(i)}, b)$.

Example 16.

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

If we set $q(y|x) \sim \mathcal{N}(x, b^2)$, then

$$r(y, x) = \min\left\{1, \frac{1+x^2}{1+y^2}\right\}$$

Suppose we have a density which has two sections separated by a portion where the density is equal to 0. If b is very small, the Metropolis-Hastings algorithm will never be able to jump across and it gets stuck. As a good check to see if the algorithm is working well, we look at a trace plot. We plot t vs $X^{(t)}$.

- For our purposes, we should aim for an acceptance rate of around $40 - 60\%$ when using Metropolis-Hastings.
- Autocorrelation plots (good check to see if it is working well) and effective sample size
- Run multiple chains (5-10) and see how similar they are to each other
- Gelman/Rubin statistic. Or we can use the Rattery/Lewis statistic.

12.3. **Independence Metropolis-Hastings.** Suppose we have $f(x)$ and we use $g(y)$.

$$r(x,y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{g(x)}{g(y)}\right\}$$

Then we pick

$$g(y) \approx \mathcal{N}(\underbrace{\hat{\mu}}_{\text{posterior mode}}, \hat{I})$$

where $\hat{I}$ in this case is the observed Fisher information, and not the same $\hat{I}$ as defined earlier.

12.4. **Gibbs Sampler.** $(X,Y) \sim f(X,Y|\alpha,\theta)$. We assume this is "full-conditional". We have

$$f(X|Y,\alpha,\theta)$$
$$f(Y|X,\alpha,\theta)$$

and we assume that we can sample *exactly* from these distributions.

- Start with $(X^{(0)}, Y^{(0)})$.
- Iterate, $X^{(i)} \sim f(X|Y^{(i-1)},\alpha,\theta)$ and $Y^{(i)} \sim f(Y|X^{(i)},\alpha,\theta)$.

Example 17. *Suppose we have*

$$Y_i \sim \text{Bin}(n_i, p_i)$$
$$p_i \sim F$$
$$I = \int pF(\mathrm{d}p) = \mathbb{E}[p]$$

We have

$$\hat{p}_i = \frac{Y_i}{n_i} \approx \mathcal{N}(p_i, \hat{p}_i, (1-\hat{p}_i)n_i)$$
$$\psi_i = \log\left(\frac{p_i}{1-p_i}\right)$$
$$\hat{\psi} \approx \mathcal{N}\left(\psi_i, \frac{1}{n_i\hat{p}_i(1-\hat{p}_i)}\right)$$

By Wasserman, All of statistics,

$$\psi_i \sim \mathcal{N}(\mu, \tau^2)$$
$$Z_i|\psi_i \sim \mathcal{N}(\psi_i, \sigma_i^2)$$

Set $\tau = 1$ and $f(\Gamma) \propto 1$ and $\theta = (\mu, \psi_1, \ldots, \psi_k)'$. Then

$$f(\Gamma|\text{rest}) \propto \text{full joint}(Z_1,\ldots,Z_k,\psi_1,\ldots,\psi_k,\Gamma|\sigma_1,\sigma_k,\tau)$$
$$\propto \mathcal{N}\left(\frac{1}{k}\sum_i \psi_i, \frac{1}{k}\right)$$

which then gives

$$f(\psi_i|\text{rest}) \propto \mathcal{N}\left(\frac{Z_i/\sigma_i^2 + \Gamma}{1+\frac{1}{\sigma_i^2}}, \frac{1}{1+\frac{1}{\sigma_i^2}}\right)$$
$$\theta_0 = \left(\sum_i \frac{y_i}{n_i k}, \ldots\right)$$

Then, we iterate

$$\Gamma^{(i)} \sim \mathcal{N}(\psi_i^{(i-1)}, i = 1,\ldots,k)$$
$$\psi_i^{(i)} \sim \mathcal{N}(\Gamma^{(i)}, \psi_i^{(i-1)}, i = 2,\ldots,k)$$
$$\psi_2 \sim \mathcal{N}(\Gamma^{(i)}, \psi^{(i)}, \psi^{(i-1)}, i = 3,\ldots,k)$$

$$Y_i|N,\theta \sim \text{Bin}(N,\theta)$$
$$N|\mu \sim \text{Pois}(\mu)$$

We have data $y_1,\dots,y_L$, latent variable N, and constants (μ,θ). We use a fully Bayesian strategy to estimate the parameters.
$\lambda = \theta$
cdotμ. Then we have

$$\begin{aligned}\mathbb{P}(\lambda,\theta) &= \mathbb{P}(\lambda)\cdot\mathbb{P}(\theta)\\ &\propto \frac{1}{\lambda}\end{aligned}$$

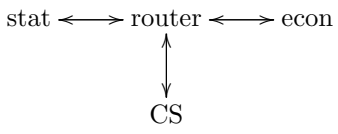
Then we have

$$\begin{aligned}\mathbb{E}\big[y_|\mu,\theta\big] &= \mathbb{E}[\mathbb{E}[y_i|N,\mu,\theta]]\\ &= \mu\cdot\theta\\ &= \lambda\end{aligned}$$

Then, when we have $\mathbb{P}(N,\lambda,\theta)$, we can get $\mathbb{P}(N,\theta)$ via

$$\mathbb{P}(N,\theta) = \int_\Lambda \mathbb{P}(N,\lambda,\theta)\ \mathrm{d}\lambda$$

Suppose we have emails between the stat, CS, and econ departments. Whenever an email is sent, it goes through a central router. If you cc yourself in an email, it will go to the router, then back towards your own department. This can be modelled as



$$\begin{pmatrix} Y_{stat,in}\\ Y_{stat,out}\\ Y_{cs,in}\\ Y_{cs,out}\\ Y_{econ,in}\\ Y_{econ,out}\end{pmatrix}_{6\times 1} = A\begin{pmatrix} X_{stat,stat}\\ X_{stat,econ}\\ X_{stat,cs}\\ X_{cs,stat}\\ X_{cs,econ}\\ X_{cs,cs}\\ X_{econ,stat}\\ X_{econ,econ}\\ X_{econ,cs}\end{pmatrix}_{9\times 1}$$

so we have $yY^t = AX^t$ for $t = 1,\dots,T$. The router has a counter, which counts emails going to certain places.

$$Y = \underbrace{\begin{pmatrix} 1&1&1&0&0&0&0&0&0\\ 1&0&0&1&0&0&1&0&0\\ 0&0&0&1&1&1&0&0&0\\ 0&1&0&0&1&0&0&1&0\\ 0&0&0&0&0&0&1&1&1\\ 0&0&1&0&0&1&0&0&1\end{pmatrix}}_{A} X$$

We can write X as $(X_1,X_2)'$ where X_1 is 5×1 and X_2 is 3×1. The rank of the matrix is, in fact, of rank 5. So,

$$\mathbb{P}(\Lambda|X,Y) \propto \mathbb{P}(X_1|X_2,\Lambda,Y)\,\mathbb{P}(X_2|Y,\Lambda)$$

13.1. **Expectation-Maximization.** Setup: Y - observations, X - latent variables, θ - constants. The goal is to get the MLE for θ. We would typically write

$$\mathbb{P}(y|\theta) = \int_X \underbrace{py,x|\theta}_{\text{DGP}}\ \mathrm{d}x$$

and then

$$\hat\theta_{MLE} = \arg\max_\theta \mathbb{P}(y|\theta)$$

Notes/challenges:

(1) If you can compute $\mathbb{P}(x|y,\theta)$, but cannot compute

$$\int \mathbb{P}(y,x|\theta)\ \mathrm{d}x$$

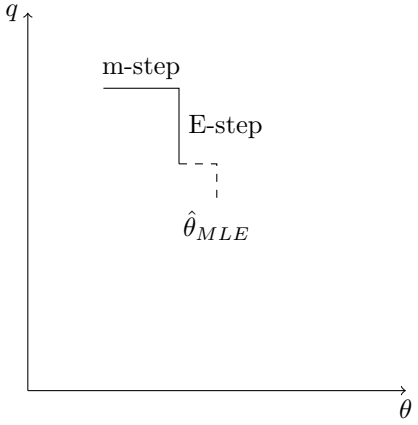
then you can use the EM.

(2) You cannot compute $\mathbb{P}(x|y,\theta)$ and cannot compute

$$\int \mathbb{P}(y,x|\theta)\ \mathrm{d}x$$

then you can use the variational EM.

Recall that X and θ are unknown. Define $q(x)$. $F(q,\theta)$ is a lower bound for $\mathbb{P}(y,\theta)$.



$Q(\theta|\theta^{(t-1)})$ is called the "transfer function" and t is the iteration.

- M-step: $\theta^{(t)} = \arg\max_\theta Q(\theta|\theta^{(t-1)})$
- E-step: compute $Q(\theta|\theta^{(t)})$

Definition 6 (Jensen's Inequality)**.**

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$$

This is due to the fact that log *is concave.*

$$\begin{aligned}
\log \mathbb{P}(y|\theta) &= \log \int \mathbb{P}(y,x|\theta)\ \mathrm{d}x \\
&= \log \int \frac{\mathbb{P}(y,x|\theta)}{q(x)} q(x)\,\mathrm{d}x \qquad \text{where } q \text{ is arbitrary} \\
&= \log \mathbb{E}_q\left[\frac{\mathbb{P}(y,x|\theta)}{q(x)}\right] \\
&\geq \mathbb{E}_q\left[\log \frac{\mathbb{P}(y,x|\theta)}{q(x)}\right] \\
&= \int \log\left(\frac{\mathbb{P}(y,x|\theta)}{q(x)}\right) q(x)\,\mathrm{d}x \\
&= F(q,\theta)
\end{aligned}$$

so we have a lower bound. We can rewrite

$$\begin{aligned}
F(q,\theta) &= \int \log \mathbb{P}(y,x|\theta)\, q(x)\,\mathrm{d}x - \int \log(q(x)) q(x)\,\mathrm{d}x \\
&= \mathbb{E}_q\left[\mathbb{P}(y,x|\theta)\right] + H(q)
\end{aligned}$$

where $H(q)$ is called the entropy of q.

13.1.1. *E-step.* Start with $(q^{(0)},\theta^{(0)})$.

$$\log \mathbb{P}(y|\theta) \geq F(q,\theta)$$

Then at step t, we have

$$q^{(t)}(x) = \arg\max_q F(q,\theta^{(t-1)})$$

given $\theta^{(t-1)}$.

13.1.2. *M-step.* Find

$$\begin{aligned}\theta^{(t)} &= \arg\max_\theta F(q^{(t)}, \theta) \\ &= \arg\max_\theta \mathbb{E}_q\left[\log \mathbb{P}(y, x|\theta)\right] \qquad \text{since } H(q) \text{ does not depend on } \theta\end{aligned}$$

If we specify $q^{(t)}(x) = \mathbb{P}\big(x|y, \theta^{(t)}\big)$, then

$$\mathbb{E}_q\left[\log \mathbb{P}(y, x|\theta)\right] = Q(\theta, \theta^{(t)})$$

14. October 29th, 2014

14.1. **EM.** We have

- y - observations
- x - latent variable
- θ - constant

$$\log \mathbb{P}(y|\theta) \geq F(q, x) = \mathbb{E}_q\left[\log \mathbb{P}(y; x|\theta)\right] - \mathbb{E}_q\left[\log q(x)\right]$$

14.2. **Statistics EM.** At iteration t, we have steps

$$\begin{aligned}q(x) &= \mathbb{P}\Big(x|y, \theta^{(t)}\Big) \\ \theta^{(t+1)} &= \arg\max_\theta Q(\theta, \theta^{(t)}) \\ &= \arg\max_\theta \mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)}\left[\log \mathbb{P}(y, x|\theta)\right]\end{aligned}$$

14.2.1. *E-step.* We have

$$F(q, \theta^{(t)}) = \log \mathbb{P}\Big(y|\theta^{(t)}\Big) + \underbrace{\mathrm{KL}(q(x) \| \mathbb{P}\Big(x|y, \theta^{(t)}\Big))}_{\int \frac{\log \mathbb{P}\left(x|y,\theta^{(t)}\right)}{q(x)} q(x)\,\mathrm{d}x}$$

We wish to find $\arg\max_q F(q, \theta)$. In practice we only need to maximize the bound. We can't easily minimize KL in practice. Then, we have

$$\begin{aligned}F(q^{(t)}, \theta) &= \mathbb{E}_{q^{(t)}}\left[\log \mathbb{P}(y, x|\theta)\right] + \underbrace{H(q^{(t)}}_{-\int \log q^{(t)}(x) q^{(t)} x\,\mathrm{d}x} \\ q^{(t)}(x) &= \mathbb{P}\Big(x|y, \theta^{(t)}\Big) \\ \theta^{(t+1)} &= \arg\max_\theta F(\mathbb{P}\Big(x, y, \theta^{(t)}\Big), \theta) \\ &= \arg\max_\theta \mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)}\left[\log \mathbb{P}(y, x|\theta)\right] \\ &= \arg\max_\theta Q(\theta, \theta^{(t)})\end{aligned}$$

Example 18. *Suppose* $z_1, \ldots, z_n \sim \mathcal{N}_k(\underbrace{\mu}_{k\times 1}, \underbrace{\Sigma}_{k\times k})$. *Partition* z *into*

$$z = \begin{pmatrix} y \\ x \end{pmatrix}$$

where y *is* k_1 *dimensional and* x *is* k_2 *dimensional, where* $k = k_1 + k_2$. *Similarly, we have*

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{pmatrix}$$

The likelihood is given by

$$\mathbb{P}(z|\mu, \Sigma) = \frac{e^{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)}}{(2\pi)^{k/2}|\Sigma|^{1/2}}$$

$\theta = (\mu, \Sigma)$. *What is* $\hat{\theta}_{MLE}$*?*

Lemma.

$$\mathbb{P}(x|y,\theta) = \mathcal{N}_{k_2}(\mu_x - \Sigma_{xy}\Sigma_y^{-1}(\mu_y - y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})$$

Let $\theta^{(0)}$ be our initial value, where

$$\theta^{(0)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

This implies

$$\mathbb{P}\Big(x|y,\theta^{(0)}\Big) = \mathcal{N}\left(0, \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}\right)$$

14.3. **E-step at iteration t.** We have

$$\mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)}\left[\log \mathbb{P}(y,x|\theta)\right]$$

Recall that $y, x|\theta \sim z|\theta$. So we have

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\begin{pmatrix} y \\ x \end{pmatrix}\right] \\ &= \begin{pmatrix} y \\ \mathbb{E}\big[x|y,\theta^{(t)}\big] \end{pmatrix} \\ \mathbb{P}(z|\theta) &= \frac{1}{\cdot} e^{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)} \end{aligned}$$

This gives us

$$\mathbb{E}[zz'] = \begin{pmatrix} yy' & y\mathbb{E}\big[x|y,\theta^{(t)}\big]' \\ \mathbb{E}\big[x|y,\theta^{(t)}\big]\, y' & \mathbb{E}\big[xx'|y,\theta^{(t)}\big] \end{pmatrix}$$

We have that

$$\begin{aligned} \mathbb{E}\Big[x|y,\theta^{(t)}\Big] &= \mu_x^{(t)} - \Sigma_{xy}^{(t)}\left(\Sigma_y^{(t)}\right)^{-1}(\mu^{(t)} - y) \\ \mathbb{E}\Big[xx'|y,\theta^{(t)}\Big] &= \mathbb{E}\Big[x|y,\theta^{(t)}\Big]\,\mathbb{E}\Big[x|y,\theta^{(t)}\Big]' + \Sigma_x^{(t)} - \Sigma_{xy}^{(t)}\left(\Sigma_y^{(t)}\right)^{-1}\Sigma_{yx}^{(t)} \end{aligned}$$

and finally

$$\begin{aligned} Q(\theta,\theta^{(t)}) &= \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)}\left[\log\left\{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\left(S\right)\right\}\right] \\ S &= -\frac{1}{2}\left((y_i-\mu_y)'\Sigma_y^{-1}(y_i-\mu_y) + (x_i-\mu_x)'\Sigma_x^{-1}(x_i-\mu_x) + (y_i-\mu_y)'\Sigma_{yx}^{-1}(x_i-\mu_x) + (x_i-\mu_x)\Sigma_{xy}^{-1}(y_i-\mu_y)\right) \end{aligned}$$

14.4. **M-step at iteration t.** We have

$$\begin{aligned} \frac{\partial Q}{\partial \mu} &= \sum_{i=1}^{n} \mathbb{E}\Big[\Sigma^{-1}(z-\mu)|y_1,\ldots,y_n,\theta^{(t)}\Big] \\ \mu^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_i \\ \mathbb{E}\big[x_i|y_1,\ldots,y_n,\theta^{(t)}\big] \end{pmatrix} \\ \frac{\partial Q}{\partial \Sigma} &= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\Big[\Sigma - (z-\mu)(z-\mu)'|y_1,\ldots,y_n,\theta^{(t)}\Big] \\ \Sigma_i^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_i \\ \mathbb{E}\big[x_i|y_i,\theta^{(t)}\big] \end{pmatrix}\begin{pmatrix} y_i & \mathbb{E}\big[x_i|y_i,\theta^{(t)}\big] \end{pmatrix} - \mu^{(t)}(\mu^{(t)})' \\ \frac{\partial}{\partial A} z'Az &= zz' \\ \frac{\partial}{\partial A}\log(\det(A)) &= (A^{-1})' \end{aligned}$$

15.1. **EM Example (mixtures).** Given $y_1, \dots, y_n$, k clusters, and $x_1, \dots, x_n$, which are memberships, then we can have

$$\begin{aligned}
&\mathbb{P}(y_i|\mu_k, \Sigma_j, x_i = k) \qquad \text{for } x_i = 1, \dots, k \\
&= \mathbb{P}(y_i|\mu_k, \Sigma_k, x_{ik} = 1) \qquad x_i = (1, 1, \dots, 1) \text{ a } k \times 1 \text{ vector of ones} \\
&\sim \mathcal{N}(y_i|\mu_k, \Sigma_k)
\end{aligned}$$

We have

$$\begin{aligned}
x_i &= (x_{i1}, \dots, x_{ik})' \\
x_{ij} &= \{0, 1\} \\
\sum_{j=1}^{k} x_{ij} &= 1 \\
\mathbb{P}(x_{ik} = 1) &= \pi_k
\end{aligned}$$

- Observed variables $y_1, \dots, y_k$.
- Latent variables $x_1, \dots, x_k$.
- Constants $\mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k, \pi_1, \dots, \pi_k$.

Putting these all together, we get

$$\begin{aligned}
\mathbb{P}(x_i) &= \text{Mult}((\pi_1, \dots, \pi_k)', 1) \\
&= \prod_{j=1}^{k} \pi_j^{x_{ij}}
\end{aligned}$$

And so we have

$$\mathbb{P}(y_1, \dots, y_n, x_1, \dots, x_n|\mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k, \pi_1, \dots, \pi_k) = \prod_{i=1}^{k} \left\{ \Big(\underbrace{\prod_{j=1}^{k} \pi_j^{x_{ij}}}_{\mathbb{P}(x_i|\pi_1,\dots,\pi_k)} \Big) \Big(\underbrace{\prod_{j=1}^{k} \mathcal{N}(y_i|\mu_j, \Sigma_j)^{x_{ij}}}_{\mathbb{P}(y_i|x_1,\dots,x_k)} \Big) \right\}$$

and so

$$\log \mathbb{P}(y, x|\mu, \Sigma, \pi) = \sum_i \sum_j \{x_{ij} \log \pi_j + x_{ij} \log \mathcal{N}(y_i|\mu_j, \Sigma_j)\}$$

Now recall that

$$F(q, \theta) = \mathbb{E}_q \left[\log \mathbb{P}(y, x|\theta)\right]$$

where

$$\begin{aligned}
q &= \mathbb{P}(x|y, \mu, \Sigma, \pi) \\
\theta &= (\mu, \Sigma, \pi)
\end{aligned}$$

and so

$$\log \mathbb{P}(y, x|\mu, \Sigma, \pi) = \sum_i \sum_j \mathbb{E}_q \left[x_{ij}\right] \log \pi_j + \mathbb{E}[x_{ij}] \log \mathcal{N}(y_i|\mu_j, \Sigma_j)$$

In the M-step, we maximize π_j, μ_j, Σ_j and in the q step, we maximize the expectation. Now, we compute

$$\begin{aligned}
\mathbb{P}(x_{ij} = 1|y_1, \dots, y_n, \mu, \Sigma, \pi) &= \frac{\mathbb{P}(x_{ij} = 1)\, \mathbb{P}(y_1, \dots, y_n|x_{ij} = 1, \mu, \Sigma, \pi)}{\sum_{j=1}^{k} \mathbb{P}(x_{ij} = 1)\, \mathbb{P}(y_1, \dots, y_n|x_{ij} = 1, \mu, \Sigma, \pi)} \\
&= \frac{\pi_j \prod_{i=1}^{n} \mathcal{N}(y_i|x_{ij} = 1, \mu_j, \Sigma_j)}{\sum_{j=1}^{k} \pi_j \prod_{i=1}^{n} \mathcal{N}(y_i|x_{ij} = 1, \mu_j, \Sigma_j)} \\
&= \mathbb{E}[x_{ij}|y_1, \dots, y_n, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k, \pi_1, \dots, \pi_k] \\
&= \mathbb{E}_{\mathbb{P}(x|y,\mu,\Sigma,\pi)} \left[x_{ij}\right]
\end{aligned}$$

1. Start from $(x^{(0)}, \mu^{(0)}, \Sigma^{(0)}, \pi^{(0)}) = \theta^{(0)}$.
2. At iteration t, we have
   - E-step:
     (a) Evaluate $(Q(\theta, \theta^{(t)})$, which involves computing $\mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)} \left[x_{ij}\right]$ for all i, j. This gives an expression which is a function of $(\mu^{(t)}, \Sigma^{(t)}, \pi^{(t)})$ for a $n \times k$ quantities.
     (b) Write down

$$Q(\mu, \Sigma, \pi; \mu^{(t)}, \Sigma^{(t)}, \pi^{(t)})$$

- M-step:

$$Q(\mu,\Sigma,\pi;\pi^{(t)},\Sigma^{(t)},\pi^{(t)}) = \sum_i \sum_j \left\{ \mathbb{E}\Big[x_{ij}|\theta^{(t)},y\Big] \log n_j + \mathbb{E}\Big[x_{ij}|\theta^{(t)}.y\Big] \log \mathcal{N}(y_i|\mu_j,\Sigma_j) \right\}$$

Define

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\mu,\Sigma,\pi} Q(\mu,\Sigma,\pi;\mu^{(t)},\Sigma^{(t)},\pi^{(t)}) \\
\mu^{(t+1)} &= \frac{\sum_{i=1}^n y_i \mathbb{E}\big[x_{ij}|y_1,\dots,y_n,\theta^{(t)}\big]}{\sum_{i=1}^n \mathbb{E}\big[x_{ij}|y_1,\dots,y_n,\theta^{(t)}\big]} \\
\pi_j^{(t+1)} &= \frac{\sum_{i=1}^n \mathbb{E}\big[x_{ij}|y_1,\dots,y_n,\theta^{(t)}\big]}{n} \\
\Sigma_j^{(t+1)} &= \frac{\sum_{i=1}^n (y_i-\mu_j)(y_i-\mu_j)' \mathbb{E}\big[x_{ij}|y_1,\dots y_n,\theta^{(t)}\big]}{\sum_{i=1}^n \mathbb{E}\big[x_{ij}|y_1,\dots,y_n,\theta^{(t)}\big]}
\end{aligned}$$

15.2. **Variational EM.** We have

$$q(x_i|\delta_i) = \text{Mult}((\delta_{1i},\dots,\delta_{ki}),1)$$

and so we have

$$\underbrace{q(x_1,\dots,x_n|\delta_1,\dots,\delta_n)}_{q(x|\delta)} = \underbrace{\prod_{i=1}^n \text{Mult}(x_i|(\delta_1,\dots,\delta_k),1)}_{\text{"fully factorized"}}$$

In **regular EM**, we can compute $\mathbb{P}\big(x|\theta^{(t)}\big) = q$.

$$F(q,\theta^{(t)}) = Q(\theta,\theta^{(t)}) = \mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)}\left[\log \mathbb{P}(y,x|\theta)\right]$$

and we have that

- E-step: compute $Q(\theta,\theta^{(t)})$
- M-step: $\theta^{(t+1)} = \arg\min_\theta Q(\theta,\theta^{(t)}$

In **variational EM**, we cannot compute $\mathbb{P}\big(x|y,\theta^{(t)}\big)$. We introduce a "variational shift" $q = q(x|\delta)$ and so

$$\begin{aligned}
F(q,\theta^{(t)}) &= F(\delta,\theta^{(t)}) \\
&= \mathbb{E}_{q(x|\delta)}\left[\log \mathbb{P}(y,x|\theta)\right] - \mathbb{E}_{q(x|\delta)}\left[q(x|\delta)\right]
\end{aligned}$$

- E-step: At iteration t,

$$\begin{aligned}
\delta^* &= \arg\max_\delta F(\delta,\theta^{(t)}) \\
F(\delta^*,\theta) &= \mathbb{E}_{q(x|\delta^*)}\left[\log \mathbb{P}(y,x|\theta)\right]
\end{aligned}$$

  and so $\delta^* = \delta^*(y_1,\dots,y_n)$, some function of our y's. So we have $q(x|\delta^*) = q(x|\delta^*(y))$.
- M-step:

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{q(x|\delta^*(y))}\left[\log \mathbb{P}(y,x|\theta)\right]$$

$$\mathbb{E}_{q(x|\delta^*)}\left[\log \mathbb{P}(y,x|\theta)\right] = \mathbb{E}_{q(x|\delta^*)}\left[\sum_{i=1}^n \log \mathbb{P}(x_i|\theta) + \log \mathbb{P}(y_i|x,\theta)\right]$$

16. November 5th, 2014

16.1. **EM (Latent Dirichlet Allocation by Bla et al. JMLB 2003).** We label the dth document as $w_d = (w_{d_1},\dots,w_{dN_d})$, that is it has N_d words. The entire document collection is $(w_1,\dots,w_m)$. Beforehand, we must specify a vocabulary. Let V denote the size of the vocabulary. We can store this in a structure like the following:

$$\begin{array}{c} 1 \\ 2 \\ \vdots \\ m \end{array} \begin{pmatrix} N_1 & \text{term 1} & \cdots & \text{term 30000} \\ N_2 & & & \\ \vdots & & & \\ N_m & & & \end{pmatrix}$$

The resulting structure is a jagged array as the number of terms in each document may differ.

$$w_{dn} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

is a $V \times 1$ matrix with $\sum_{v=1}^{V} w_{dnv} = 1$ with $w_{dnv} \in \{0, 1\}$. And so, w_d is a $V \times N_d$ matrix.

for $d = 1, \ldots, m$ **do**
  $\theta_d \sim \text{Dirichlet}_k(\alpha_1, \ldots, \alpha_k),\ \alpha_i = \alpha\ \forall i$
  for $n - 1, \ldots, N_d$ **do**
    $Z_{dn} \sim \text{Mult}(\theta_d, 1)$
    $w_{dn} \sim \mathbb{P}(w_{dn}|Z_{dn}, \beta)$
  end for
end for

where Z_{dn} is $k \times 1$ with $Z_{dnk} \in \{0, 1\}$ and $\sum_k Z_{dnk} = 1$ and β is $V \times k$, equalling $[\beta_1|\beta_2|\cdots|\beta_k]$, with $\sum_{v=1}^{V} \beta_{kv} = 1$ and $\beta_{kv} \in [0, 1]$.

- Our observations are $Y = (w, N)$, where w is $m \times N_{1:k} \times V$ and N is $m \times 1$.
- Our latent variables are $X = (\theta, Z)$ where θ is $m \times k$ and Z is $m \times N_{1:m} \times k$.
- Our constants are $\theta = (\alpha, \beta, K, V)$ where α is 1×1, β is $V \times k$, K is 1×1, and V is 1×1.

The goal is to find

$$F(q_\Delta, \theta) = \mathbb{E}_{q_\Delta}\left[\log \mathbb{P}(y, x|\theta)\right] - \mathbb{E}_{q_\Delta}\left[\log q_\Delta\right]$$

At iteration $t - 1$, in the variational E-step, we find

$$\Delta^* = \arg\max_{\Delta} F(q_\Delta, \theta^{(t-1)})$$

In the variational M-step, we find

$$\theta^{(t)} = \arg\max_{\theta} F(q_{\Delta^*}, \theta)$$

where $\Delta^*(y, \theta^{(t-1)})$, that is it is a function of y and $\theta^{(t-1)}$.

We have that

$$\begin{aligned}
\mathbb{P}(w, N, \theta, Z|\alpha, \beta, K, V) &= \mathbb{P}(w, \theta, Z|\alpha, \beta) \\
&= \prod_d \left\{ \mathbb{P}(\theta_d|\alpha) \prod_n \mathbb{P}(z_{dn}|\theta_d)\, \mathbb{P}(w_{dn}|z_{dn}, \beta) \right\} \\
\mathbb{P}(w|\alpha, \beta) &= \int_\theta \prod_d \left\{ \mathbb{P}(\theta_d|\alpha) \prod_n \sum_{Z_n} \mathbb{P}(Z_{dn}|\theta_d)\, \mathbb{P}(w_{dn}|Z_{dn}, \beta) \right\} \mathrm{d}\theta \\
\mathbb{P}(\theta, Z|w, \alpha, \beta) &= \frac{\mathbb{P}(w, \theta, Z|\alpha, \beta)}{\mathbb{P}(w|\alpha, \beta)}
\end{aligned}$$

Now, we have

$$\begin{aligned}
\sum_{Z_{dn}} \mathbb{P}(Z_{dn}|\theta_d)\, \mathbb{P}(w_{dn}|Z_{dn}, \beta) &= \sum \left\{ \prod_k \theta_{dk}^{Z_{dnk}} \mathbb{P}(w_{dn}|\beta_k)^{Z_{dnk}} \right\} \\
&= \theta_{d1}\mathbb{P}(w_{d1}|\beta_1) + \theta_{d2}\mathbb{P}(w_{d2}|\beta_2) + \cdots
\end{aligned}$$

where we sum over

$$Z_{dn} \in \underbrace{\begin{matrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{matrix}}_{k \text{ elements}}$$

Then,

$$\mathbb{P}(w|\alpha, \beta) = \frac{\Gamma(k\alpha)}{k\Gamma(\alpha)} \int_\theta \prod_k \theta_k^{\alpha - 1} \left(\prod_n \sum_k \theta_{dk} \prod_v \beta_{vk}^{w_{dnk}} \right) \mathrm{d}\theta$$

where $\sum_k \theta_{dk} \prod_v \beta_{vk}^{w_{dnk}} = \mathbb{P}(w_{dn}|\beta_k)$. Now, we can compute

$$\mathbb{P}(w, \theta|\alpha, \beta)$$

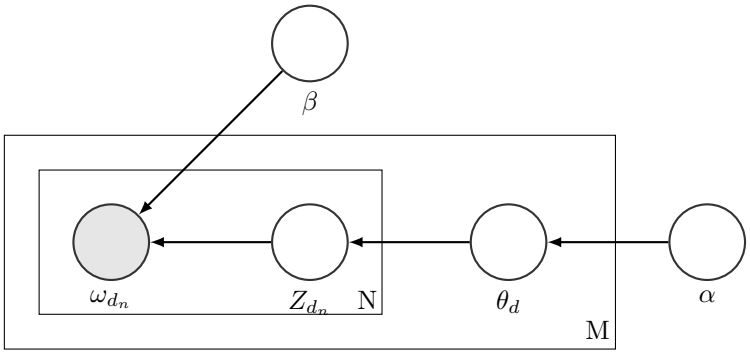
which has the same integral $\int_\theta \mathrm{d}\theta$. We can also calculate

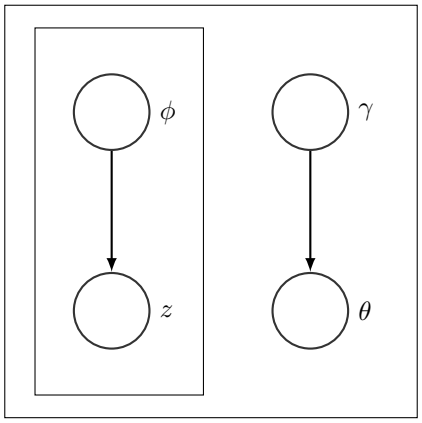
$$\mathbb{P}(w, \theta, Z|\alpha, \beta)$$

Now, in the first case, we pretend q is of the form $q_\Delta(\theta)$ and in the second case, we pretend $q_\Delta(\theta, Z)$. In the first case, we need $m \times k$ numerical optimization problems to solve. In the second, we need $m \times N_{1:m} \times k$. In this case, all of these are closed form updates.

$$\begin{aligned} q_\Delta(\theta, Z) &= \left(\prod_d q(\theta_d|\gamma_d)\right) \prod_d \prod_n q(Z_{dn}|\phi_{dn}) \\ &= \left(\prod_d \mathrm{Dirichlet}_k(\theta_d|\gamma_d)\right) \left(\prod_d \prod_n \mathrm{Mult}(Z_{dn}|\phi_{dn})\right) \end{aligned}$$

where γ_d is a matrix of size $m \times k$ of variational parameters. We can visualize this as follows:





Then,

$$\mathbb{E}_{\mathbb{P}(x|y,\theta)} [\log \mathbb{P}(y, x|\theta)] = \mathcal{L}(\theta) + \mathrm{KL}(q_\Delta(x) \| \mathbb{P}(x|y, \theta))$$

Now, we have

$$\begin{aligned} F(q(\theta, Z|\gamma, \phi), (\alpha, \beta)) &= \mathbb{E}_{q(\theta,Z|\gamma,\phi)} [\log \mathbb{P}(w, \theta, Z|\alpha, \beta)] - \mathbb{E}_{q(\theta,Z|\gamma,\phi)} [\log q(\theta, Z|\gamma, \phi)] \\ &= F((\gamma, \phi), (\alpha, \beta)) \end{aligned}$$

So in the E-step, we maximize over (γ, ϕ) and in the M-step we maximize over (α, β). We can further reduce this to

$$F(q(\theta, Z|\gamma, \phi), (\alpha, \beta)) = \mathbb{E}_q [\log \mathbb{P}(\theta|\alpha)] + \mathbb{E}_q [\log \mathbb{P}(Z|\theta)] + \mathbb{E}_q [\log \mathbb{P}(w|\alpha, \beta)] - \mathbb{E}_q [\log q(\theta|\gamma)] - \mathbb{E}_q [\log q(Z|\gamma)]$$

If we look at the $\mathbb{E}_q [\log \mathbb{P}(\theta|\alpha)]$ term, we have

$$\mathbb{E}_q [\log \mathbb{P}(\theta|\alpha)] = \mathbb{E}_q \left[\sum_d \sum_k (\alpha - 1) \log \theta\right]$$

This calculation can be found in appendix A1.

17. November 10th, 2014

17.1. **State Space Models.** The general formulation for a linear state space model is:

$$\begin{aligned} x_{t+1} &= Ax_t + \omega_t \\ y_t &= Cx_t + v_t \end{aligned}$$

where $t \geq 1$. As an example, think of y_t as a 2-dimensional vector of GPS values and x_t as a 4-dimensional vector which includes latent variables for the true position and velocity.

$$y_t = \begin{pmatrix} \text{pos}_x \\ \text{pos}_y \end{pmatrix}_t \qquad x_t = \begin{pmatrix} \text{pos}_x \\ \text{pos}_y \\ \text{vel}_x \\ \text{vel}_y \end{pmatrix}_t$$

In this particular example, A and C are known.

$$\begin{aligned} \omega_t &\sim \mathcal{N}(0, Q) \\ v_t &\sim \mathcal{N}(0, R) \\ \omega_t &\perp\!\!\!\perp v_t \end{aligned}$$

The EM for a linear, gaussian, state space model is called the Kalman filter.

- Prediction: $\mathbb{P}(x_t | y_1, \ldots, y_{t-1})$
- Filtering: $\mathbb{P}(x_t | y_1, \ldots y_t)$
- Smoothing: $\mathbb{P}(x_t | y_1, \ldots, y_t, y_{t+1}, y_n)$

Remarks:

(1) Time-varying network tomography and state space model.

$$\begin{aligned} x_{t+1} &= A_{t+1}x_t + \lambda_{t+1}\mathbf{1} + \omega_t \\ y_t &= Cx_t + v_t \end{aligned}$$

Suppose X_t is $O(n^2)$ and Y_t is $O(n)$. To estimate X_t from Y_t, we specify a window $(t - \delta, t + \delta)$ to estimate the higher dimensional object. This is called the local likelihood.

$$\begin{aligned} \begin{pmatrix} x_{t+1} \\ 1 \end{pmatrix} &= \begin{pmatrix} F_{t+1} & \lambda_{t+1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x_t \\ 1 \end{pmatrix} + \begin{pmatrix} \omega_t \\ 0 \end{pmatrix} \\ y_t &= \begin{pmatrix} C & 0 \end{pmatrix} \begin{pmatrix} x_t \\ 1 \end{pmatrix} + v_t \end{aligned}$$

The top equation is of the form $\begin{pmatrix} x_{t+1} \\ 1 \end{pmatrix}$ since this gives something of the form

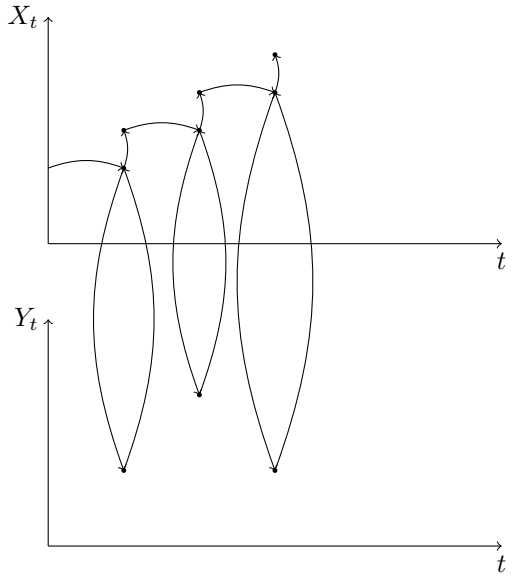
$$\begin{aligned} \tilde{x}_{t+1} &= \tilde{A}_{t+1}\tilde{x}_t + \tilde{\omega}_t \\ y_t &= \tilde{C}\tilde{x}_t + v_t \end{aligned}$$

17.2. **FA/Normal-Normal.** In the Normal-Normal model, we have

$$\begin{array}{ccc} x_1 & x_2 & \cdots \\ \downarrow & \downarrow & \\ y_1 & y_2 & \cdots \end{array}$$

whereas in the latent gaussian state space model, we have

$$\begin{array}{ccccc} x_t & \longrightarrow & x_2 & \longrightarrow & \cdots \\ \downarrow & & \downarrow & & \\ y_1 & & y_2 & & \cdots \end{array}$$



17.2.1. *E-step.*
(1) Predict
(2) Project
(3) Correct
Once we have done this, we get $x_1^1, \dots, x_T^T$ which gives us all our filtering distributions. Then, we can go back and find $x_1^T, \dots, x_{T-1}^T, x_T^T$, which are the smoothing distributions.

17.2.2. *M-step.* Find $\theta^{(t)}$.
We have $\theta^{(0)} X^{(0)}$. At iteration i we have the E-step, where we compute

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{\mathbb{P}\left(x|y,\theta^{(t)}\right)} \left[\log \mathbb{P}(y, , x|\theta)\right]$$

where $\theta = (A, C, Q, R)$. To find this, we need to compute

$$\begin{aligned}
&\mathbb{E}[x_t|y_1, \dots, y_t] \\
&\mathbb{E}[x_t x_t'|y_1, \dots, y_t] \\
&\mathbb{E}[x_t|y_1, \dots, y_T] \\
&\mathbb{E}[x_t x_t'|y_1, \dots, y_T]
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(x_1, \dots, x_T, y_1, \dots, y_T) &= \mathbb{P}(x_1) \prod_{t=2}^{T} \mathbb{P}(x_{t+1}|x_t) \prod_{t=1}^{T} \mathbb{P}(y_t|x_t) \\
\mathbb{P}(x_{t+1}|x_t) &= \exp\left\{-\frac{1}{2}(x_{t-1} - Ax_t)' Q^{-1} (x_{t+1} - Ax_t)\right\} (2\pi)^{-k/2} |Q|^{-1/2} \\
\mathbb{P}(y_t|x_t) &= \exp\left\{-\frac{1}{2}(y_t - Cx_t)' R^{-1} (y_t - Cx_t)\right\} (2\pi)^{-p/2} |R|^{-1/2} \\
\mathbb{P}(x_1) &= \mathcal{N}_k(x_1; \pi_1, v_1)
\end{aligned}$$

Now, we have

$$\begin{aligned}
Q &= \mathbb{E}_{\mathbb{P}(x|y,\theta)} \left[\log \mathbb{P}(x, y|\theta)\right] \\
\hat{x}_t &= \mathbb{E}_{\mathbb{P}(x|y)} \left[x_t\right] \\
P_t &= \mathbb{E}_{\mathbb{P}(x|y)} \left[x_t x_t'\right] \\
P_{t,t-1} &= \mathbb{E}\left[x_t x_{t-1}'\right]
\end{aligned}$$

Now,

$$x_t^T = \mathbb{E}_{\mathbb{P}(x_t|y_1,\dots,y_T)} \left[x_t\right]$$

$$V_t^T = \text{Var}_{\mathbb{P}(x_t|y_1,\dots,y_T)}(x_t)$$
$$x_t^{t-1} = Ax_{t-1}^{t-1}$$
$$V_t^{t-1} = AV_{t-1}^{t-1}A' + Q$$

κ_t, which is known as our Kalman gain, is

$$\kappa_t = V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1}$$

Filtering:

$$x_t^t = \underbrace{x_t^{t-1}}_{\text{start position}} + \underbrace{\kappa_t}_{\text{allocation}}(y_t - \underbrace{Cx_t^{t-1}}_{\text{error}})$$

Another correction term is:

$$V_t^t = V_t^{t-1} - \kappa_t C V_t^{t-1}$$

Iterating, we start with

$$x_1^0 = \pi_1$$
$$V_1^0 = V_1$$

Then, compute

$$\hat{x}_t = x_t^T$$
$$P_t = V_t^T + x_t^T(x_t^T)'$$
$$P_{t,t-1} = V_{t,t-1}^T + x_t x_{t-1}'$$

To go backwards, we have

$$J_{t-1} = V_{t-1}^{t-1}A'(V_t^{t-1})^{-1}$$
$$x_{t-1}^T = x_{t-1}^{t-1} + J_{t-1}(x_t^T - Ax_{t-1}^{t-1})$$
$$V_{t-1}^T = V_{t-!}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1})J_{t-1}'$$

Combining this all, we get

$$V_{t-1,t-2}^T = V_{t-1}^{t-1}J_{t-2}' + J_{t-1}(V_{t,t-1}^T - AV_{t-1}^{t-1})J_{t-2}'$$

Then, we have

$$V_{T,T-1}^T = (I - \kappa_T C)AV_{T-1}^{T-1}$$

because we are starting from the variance-covariance matrix of X_T^T, X_T^T.

18. November 12th, 2014

18.1. **EM for Linear Gaussian State Space Models.**

$$x_t = Ax_t + \omega_t$$
$$y_t = Cx_t + v_t$$
$$\omega_t \sim \mathcal{N}(0, Q)$$
$$v_t \sim \mathcal{N}(0, R)$$
$$x_1 \sim \mathcal{N}(\pi_1, V_1)$$

A variant of this is to have

$$x_{t+1} = Ax_t + Bv_t + \omega_t$$

where v_t is a vector of observable covariates "controls". Then to compute $Q(\theta, \theta^{(t)})$, we need

$$\hat{x}_t = \mathbb{E}[x_t|y_1, \dots, y_T]$$
$$P_t = \mathbb{E}[x_t x_t'|y_1, \dots, y_T]$$
$$P_{t.t-1} = \mathbb{E}\big[x_t x_{t-1}'|y_1, \dots, y_T\big]$$

At iteration t, we have $C^{(t-1)}, A^{(t-1)}, R^{(t-1)}, Q^{(t-1)}, \pi_1^{(t-1)}, V_1^{(t-1)}$. Then,

$$\frac{\partial Q}{\partial C} = -\sum_{t=1}^T R^{-1}y_t\hat{x}_t' + \sum_{t=1}^T R^{-1}CP_t = 0$$
$$C^{(t)} = \left(\sum_t y_t\hat{x}_t'\right)\left(\sum_t P_t\right)^{-1}$$

$$
\begin{aligned}
\frac{\partial Q}{\partial R^{-1}} &= \frac{T}{2} - \sum_t \left(\frac{1}{2} y_t y_t' - C\hat{x}_t y_t' + \frac{1}{2} C P_t C' \right) = 0 \\
R^{(t)} &= \frac{1}{T} \sum_t (y_t y_t' - C^{(t)} \hat{x}_t y_t') \\
\frac{\partial Q}{\partial A} &= -\sum_{t=2}^{T} Q^{-1} P_{t,t-1} + \sum_{t=2}^{T} Q^{-1} A P_{t-1} = 0 \\
A^{(t)} &= \left(\sum_{t=1}^{T} P_{t,t-1} \right) \left(\sum_{t=2}^{T} P_t \right)^{-1} \\
\frac{\partial Q}{\partial Q^{-1}} &= \frac{T-1}{2} Q - \frac{1}{2} \sum_{t=1}^{T} (P_t - A P_{t-1,t} - P_{t,t-1} A' + A P_{t-1} A') \\
&= \frac{T-1}{2} Q - \frac{1}{2} \left(\sum_{t=2}^{T} P_t - A^{(t)} \sum_{t=2}^{T} P_{t-1,t} \right) = 0 \\
Q^{(t)} &= \frac{1}{T-1} \left(\sum_{t=2}^{T} P_t - A^{(t)} \sum_{t=2}^{T} P_{t-1,t} \right) \\
\frac{\partial Q}{\partial \pi_1} &= (\hat{x}_1 - \pi_1) V_1^{-1} = 0 \\
\frac{\partial Q}{\partial V_1^{-1}} &= \frac{1}{2} V_1 - \frac{1}{2} (P_1 - \hat{x}_1 \pi_1' - \pi_1 \hat{x}_1' + \pi_1 \pi_1') = 0
\end{aligned}
$$

Now,

$$
\begin{aligned}
x_1^0 &= \pi \\
V_1^0 &= V_1 \\
\kappa_1 &= \frac{V_1}{R + V} \\
x_1^1 &= \frac{R}{R + V_1} \pi + \frac{V_1}{R + V_1} y_1 \\
&= (1 - \kappa_1) \pi + \kappa_1 y_1
\end{aligned}
$$

And for our iterates, we have

$$
\begin{aligned}
V_1^1 &= \frac{R V_1}{R + V_1} \\
&= \kappa_1 \cdot R \\
x_2^1 &= A x_1^1 \\
V_2^1 &= A^2 V_1^1 + Q \\
\kappa_2 &= \frac{A^2 V_1^1 + Q}{R + (A^2 V_1^1 + Q)} \\
&= \frac{V_2^1}{R + V_2^1}
\end{aligned}
$$

The value

$$
A \left\{ (1 - \kappa_1) \pi + \kappa_1 y_1 \right\} (1 - \kappa_2) + \kappa_2 y_2
$$

is some weights of something...

$$
\begin{aligned}
x_1^0 &= \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} \\
V_1^0 &= \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \\
\kappa_1 &= \begin{pmatrix} \frac{S_1}{R + S_1 + S_2} \\ \frac{S_2}{R + S_1 + S_2} \end{pmatrix}
\end{aligned}
$$

This gives

$$x_1^1 = \begin{pmatrix} \frac{R+S_2}{R+S_1+S_2}\pi_1 - \frac{S_1}{R+S_1+S_2}\pi_2 + \frac{S_1}{R+S_1+S_2}y_1 \\ \frac{R+S_1}{R+S_1+S_2}\pi_2 - \frac{S_2}{R+S_1+S_2}\pi_1 + \frac{S_2}{R+S_1+S_2}y_1 \end{pmatrix}$$

$$V_1^1 = \begin{pmatrix} \frac{S_1(R+S_2)}{R+S_1+S_2} & -\frac{S_1S_2}{R+S_1+S_2} \\ -\frac{S_1S_2}{R+S_1+S_2} & \frac{S_2(R+S_1)}{R+S_1+S_2} \end{pmatrix}$$

18.2. **General State Space Model.** Suppose we have

- $\mathbb{P}(X_0)$
- $\mathbb{P}(X_t|X_{t-1})$ for $t \geq 1$
- $\mathbb{P}(Y_t|X_t)$ for $t \geq 1$.

From this, we get a latent trajectory

$$x_{0:t} = \{x_0, \ldots, x_t\}$$

and

$$y_{1:t} = \{y_1, \ldots, y_t\}$$

We have that

$$\begin{aligned} I(f_t) &= \mathbb{E}_{\mathbb{P}(x_{0:t}|y_{1:t})}\left[f_t(x_{0:t})\right] \\ &= \int f(x_{0:t})\mathbb{P}(x_{0:t}|y_{1:t}) \, \mathrm{d}x_{0:t} \end{aligned}$$

Recall that $x_0, x_1, \ldots, x_t, y_1, \ldots, y_t$ is Multivariate Normal.

Example 19. $f_t(x_{0:t}) = x_t x_t' - \mathbb{E}_{\mathbb{P}(x_t|y_{1:t})}\left[x_t\right]\mathbb{E}_{\mathbb{P}(x_t|y_{1:t})}\left[x_t\right]'$

At time t, we can write

$$\mathbb{P}(x_{0:t}|y_{1:t}) = \frac{\mathbb{P}(x_{0:t})\,\mathbb{P}(y_{1:t}|x_{0:t})}{\int \mathbb{P}(x_{0:t})\,\mathbb{P}(y_{1:t}|x_{0:t}) \, \mathrm{d}x_{0:t}}$$

Then, we have

$$\mathbb{P}(x_{0:t+1}|y_{1:t+1}) = \mathbb{P}(x_{0:t}|y_{1:t}) \, \frac{\mathbb{P}(x_{t+1}|x_t)\,\mathbb{P}(y_{t+1}|x_{t+1})}{\mathbb{P}(y_{t+1}|y_{1:t})}$$

For $\mathbb{P}(x_t|y_{1:t})$ and $\mathbb{P}(x_{t+1}|y_{1:t+1})$ we have

(i) $\mathbb{P}(x_{t+1}|y_{1:t}) = \int \mathbb{P}(x_{t+1}|x_t)\,\mathbb{P}(x_t|y_{1:t}) \, \mathrm{d}t$

(ii) $\mathbb{P}(x_{t+1}|y_{1:t+1}) = \mathbb{P}(x_{t+1}|y_{1:t}) \, \frac{\mathbb{P}(y_{t+1}|x_{t+1})}{\int \mathbb{P}(x_{t+1}|y_{1:t})\mathbb{P}(y_{t+1}|x_{t+1}) \, \mathrm{d}x_{t+1}}$

19. November 17th, 2014

19.1. **Particle Filters.** Recall that we have

(1) $\mathbb{P}(x_0)$

(2) $\mathbb{P}(x_t|x_{t-1})$, $t \geq 1$

(3) $\mathbb{P}(y_t|x_t)$, $t \geq 1$.

$$I(f_t) = \int f_t(x_{0:t})\mathbb{P}(x_{0:t}|y_{1:t}) \, \mathrm{d}x_{0:t}$$

We can sample multiple trajectories $x_{0:t}^{(i)}$ which will gives

$$\frac{1}{B}\sum_{i=1}^{B} f(x_{0:t}^{(i)}) \longrightarrow I(f_t)$$

Suppose we have samples $x_{0:t}^{(i)}$ and want to get it for $t+1$. Instead of resampling $x_{0:t+1}^{(i)}$, for each i, we can condition on $x_{0:t}^{(i)}$ for each i and then extend them using the recursions.

If we start at $B = 1000$, say, and then sample each of them, then eventually after some t, they will converge. This is very bad, and so we implement a strategy to separate them at some point. Recall from last time, that we had

$$\mathbb{P}(x_{0:t+1}|y_{1:t+1}) = \mathbb{P}(x_{0:t}|y_{1:t}) \, \frac{\mathbb{P}(x_{t+1}|y_t)\,\mathbb{P}(y_{t+1}|x_{t+1})}{\mathbb{P}(y_{1:t+1}|y_{1:t})}$$

$$\mathbb{P}(x_t|y_{1:t}) \longrightarrow \mathbb{P}(x_{t+1}|y_{1:t}) \longrightarrow \mathbb{P}(x_{t+1}|y_{1:t+1})$$

(a) Importance Sampling:

$$I(f_t) = \frac{\int f_t(x_{0:t})\omega(x_{0:t})\pi(x_{0:t}|y_{1:t})\,\mathrm{d}x_{0:t}}{\int \omega(x_{0:t})\pi(x_{0:t}|y_{1:t})\,\mathrm{d}x_{0:t}}$$
$$\omega(x_{0:t}) = \frac{\mathbb{P}(x_{0:t}|y_{1:t})}{\pi(x_{0:t}|y_{1:t}}$$

Given B "particles", $x_{0:t}^{(i)}$,

$$\begin{aligned}\hat{I}_B(f_t) &= \sum_{i=1}^{B} \frac{f(x_{0:t}^{(i)})\omega(x_{0:t}^{(i)})\cancel{\frac{1}{B}}}{\sum_{i=1}^{B}\omega(x_{0:t}^{(i)}\cancel{\frac{1}{B}}} \\ &= \sum_{i=1}^{B} f_t(x_{0:t}^{(i)})\tilde{\omega}(x_{0:t}^{(i)})\end{aligned}$$

where

$$x_{0:t}^{(i)} \sim \pi(x_{0:t}|y_{1:t})$$

and

$$\tilde{\omega}(x_{0:t}^{(i)}) = \frac{\omega(x_{0:t}^{(i)})}{\sum_{i=1}^{B}\omega(x_{0:t}^{(i)})}$$

Then,

$$\begin{aligned}\mathbb{P}(x_{0:t}|y_{1:t}) &\approx \hat{P}_B(x_{0:t}|y_{1:t}) \\ &= \sum_{i=1}^{B}\tilde{\omega}(x_{0:t}^{(i)})\delta_{x_{0:t}^{(i)}}(x_{0:t})\end{aligned}$$

Now, we can write

$$\hat{I}_B(f_t) = \int f_t(x_{0:t})\hat{P}_B(x_{0:t}|y_{1:t})\,\mathrm{d}x_{0:t}$$

(b) Sample Impotance Resample (S.I.R.) (Particle Filter)
We now want to find $\hat{P}_B(x_{0:t+1}|y_{1:t+1})$ by changing $\{x_{0:t}^{(i)} : i = 1,\dots,B\}$. In order for this to happen, we require some condition:

- $\pi(x_{0:t+1}|y_{1:t+1})$ must admit a marginal distribution at t:

$$\begin{aligned}&\pi(x_{0:t+1}|y_{1:t+1}) + \pi(x_{0:t}|y_{1:t})\pi(x_{t+1}|x_{0:t}, y_{1:t+1}) \\ &\quad = \pi(x_0)\prod_{k=1}^{t+1}\pi(x_k|x_{0:k-1}, y_{1:k}) \qquad \text{doing this recursively}\end{aligned}$$

Let $\tilde{\omega}(x_{0:t}^{(i)}) = \tilde{\omega}_t^{(i)}$. Then

$$\tilde{\omega}_t^{(i)} \propto \tilde{\omega}_{t-1}^{(i)}\frac{\mathbb{P}\left(x_t^{(i)}\middle|x_{t-1}^{(i)}\right)\mathbb{P}\left(y_t\middle|x_t^{(i)}\right)}{\pi(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})}$$

Example 20.

$$\begin{aligned}\pi(x_{0:t}|y_{1:t}) &= \mathbb{P}(x_{0:t}) \\ &= \mathbb{P}(x_0)\prod_{k=1}^{t}\mathbb{P}(x_k|x_{k-1})\end{aligned}$$

This then gives an approximation $\hat{P}_B(x_0)$ using $\tilde{\omega}_0^{(i)}$. Doing this, we have

$$\tilde{\omega}_t^{(i)} \propto \tilde{\omega}_{t-1}^{(i)}\mathbb{P}\left(y_t\middle|x_t^{(i)}\right)$$

1. *Initialize: For $i = 1,\dots,B$, $x_0^{(i)} \sim \mathbb{P}(x_0)$.*
2. *IS step: At iteration t, $\tilde{x}_t^{(i)} \sim \mathbb{P}\left(x_t\middle|x_{t-1}^{(i)}\right)$. Then, set $\tilde{x}_{0:t}^{(i)} = (x_{0:t-1}^{(i)}, x_t^{(1)})$ for $i = 1,\dots,B$. Lastly, we can calculate the weights $\tilde{\omega}_t^{(i)} = \mathbb{P}\left(y_t\middle|\tilde{x}_t^{(i)}\right)\tilde{\omega}_{t-1}^{(i)}$.*
3. *Seletion step: Sample with replacement $\{x_{0:t}^{(i)} : i = 1,\dots,B\}$ from $\{\tilde{x}_{0:t}^{(i)} : i = 1,\dots,B\}$ with probability $\{\tilde{\omega}_t^{(i)} : i = 1,\dots,B\}$.*

20. November 19th, 2014

21. December 1st, 2014

Definition 7 (Hamiltonian Monte Carlo)**.** *Uses gradients so we can take larger step1s without sacrificing acceptance probability.*

We wish that the acceptance probability is close to 1, that is the value of the density at the starting point and the ending point is similar. We introduce another parameter γ, which is independent to our desired θ. We can then sample from a γ. Given this γ, we can sample a θ by going along the contours of the posterior at γ.

- This is a Metropolis within Gibbs scheme with overlapping parameter blocks p and (θ, p).
- HMC **does not** solve the problem of sampling from multiple modes very well.