Final Project

1 Guidelines

1.1 Goal

To apply a combination of the computational and inferential techniques addressed in this class to a real statistical problem which is of interest to you. Some possible lines of inquiry are: (1) propose an algorithm or model and evaluate its performance with simulated data; (2) implement an efficient piece of software for a model of scientific interest; (3) analyze an actual dataset and present your scientific findings.

1.2 Format

Groups of 1-3. STAT 840/CM 760 students must come up with their own projects. STAT 440 students may come up with their own or choose from one of the projects below. The project itself consists of a 6-10 page report in the style of a scientific article. You must include an abstract and references (details below), but aside from that you are free to modify the following template:

1. **Abstract:** 100-200 words. A good abstract is like a good advertisement: catchy and to the point. Address the following without getting into technical details: (1) Why is the problem important? (2) What is the challenge/what do you plan to do? (3) What is your solution/what did you discover?

2. **Introduction:** The introduction is like the abstract but with more detail. This is the place to cite the work of others on the problem. Your project must contain a minimum of 6 references. The introduction typically ends with an outline of the remainder of the article.

3. **Methodology:** This is where things get more technical in terms of notation, models, equations, etc. Whatever is too technical however (e.g., proofs, long simplifications of equations) should go to the Appendix.

4. **Results:** You must apply your methods to a real dataset, and/or to a simulated dataset with carefully chosen parameters (i.e., relevant to a specific application, taken from an

existing article, etc). Depending on your project, this would be the place to answer some of the following questions: Does your method beat the competitor? Which model best fits the data? Does the best model actually fit the data? What is your conclusion about the dataset that you are studying?

5. **Discussion:** Summarize your results and conclusions. Most importantly, you must point out some shortcomings of your methodology/analysis but in a positive way, i.e., talk about directions for future research.

6. **Appendix:** Include all computer code, math, extra analyses you did, and whatever else you think is important but don't have space to include in the 10 pages.

1.3 Grading

This is an open-ended project which you should think of as an opportunity for statistical thinking rather than a set of questions for which you must provide the correct, unique answer. Aside from basic presentation (legibility, labeling of figure axes, etc.), here are some elements that I will be looking for in terms of evaluation:

- **Presentation:** Clarity and organization of ideas. Make sure to emphasize your main results. Please display results graphically whenever possible. If you must include tables of numbers, round off the decimals that nobody cares about. As always, figure axes should be labeled and code must be commented. All reported parameter estimates must also include standard errors or interval estimates.

- **Computing:** Magnitude of the computing challenge (e.g., fitting a non-standard model is harder than estimating the mean and variance of a univariate Normal). Efficiency of coding. If you are less interested in a data analysis, a possible project would produce a fast and efficient implementation of an inference algorithm with potential applications. If you are really up to the task, it is possible to hardwire "for-loops" in `C++` instead of `R` using the package `Rcpp`...

- **Statistics:** In a nutshell, I am looking for how competently you are relating the tools to the actual scientific problem at hand. Models and/or methods must be properly explained/justified. If you are interested in making claims about a particular dataset, then you should be selecting between competing models and assessing goodness-of-fit. If you are interested in evaluating a computational methodology, or perhaps the value of a misspecified model (assuming they are all wrong but some are useful), then you should be choosing relevant and meaningful parameters to simulate data in order to understand the operational characteristics under a "ground truth". Remember: I view statisticians as lawyers that don't

get paid enough to lie. It's an argument: there's no universally optimal approach, each has its own merits and drawbacks. The picture is complete when you understand them both.

1.4 Deadlines

The project is due on Friday, April 4. A brief description of your project (more than just "I'm doing project 2") along with team members is due at the beginning of class on Thursday, March 13.

2 Sample Projects

Feel free to modify these projects, inspire yourself, answer only some of the questions, etc. Naturally, if you'd like to discuss ideas or clarifications I am happy to assist.

2.1 A Heteroscedastic Regression Model for Survival Analysis

In biostatistics, survival analysis is concerned with modeling the lifetimes of individuals after the onset of a disease, conditional on various predictors. Let $T > 0$ be the lifetime of an individual after the onset of a disease, and $x = (x_1, \ldots, x_p)$ be a set of predictors. Two standard survival models both model the individual's hazard function as opposed to T itself. That is, if $f(t \mid x)$ and $F(t \mid x)$ are the conditional PDF and CDF of T, the hazard function is defined as

$$\lambda(t \mid x) = \frac{f(t \mid x)}{1 - F(t \mid x)}.$$

The first common model for survival data is Cox' proportional hazard model (CPH),

$$\lambda(t \mid x) = \lambda_0(t) \exp(x'\beta),$$

where $\lambda_0(t)$ is a baseline hazard model that is often estimated non-parametrically from the data. The second common model for survival data is the accelerated failure time model (AFT):

$$\lambda(y \mid x) = e^{-x'\beta} \lambda_0 \left(y e^{-x'\beta} \right).$$

This is implied if $f(t \mid x) = e^{-x'\beta} f_0 \left(t e^{-x'\beta} \right)$ and $f_0(t)$ is some baseline PDF which must be specified beforehand. It turns out that the AFT assumption is equivalent to a linear model

$$z = x'\beta + \epsilon,$$

where $z = \log(T)$ is the log of the survival time and $\exp(\epsilon) \sim f_0(t)$.

Here, we are going to break from both the CPH and AFT assumptions and investigate the following heteroskedastic linear model (HLM):

$$z \mid x, w \sim \mathcal{N}\Big(x'\beta, \exp(w'\gamma)\Big). \tag{2.1}$$

For $\theta = (\beta, \gamma)$, the log-likelihood for this model is

$$\ell(\theta \mid Z, X, W) = -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(z_i - x_i'\beta)^2}{\exp(w_i'\gamma)} + w_i'\gamma\right],$$

which is not convex. However, for fixed γ, you can easily find the MLE of β using the following pseudo-`R` commmand:

$$\texttt{lm(}Z\ \texttt{~}\ X\texttt{, weights =}\ \exp(-W\gamma)\texttt{)}$$

What is less well-known (or perhaps unknown?) is that you can then find the MLE of γ for fixed β using `glm` with a Gamma distribution:

$$\texttt{glm((}Z - X\beta\texttt{)}^2\ \texttt{~}\ W\texttt{, family = Gamma("log"))}$$

Iterating between these two steps appears to converge to a unique maximum $\hat{\theta}$.

2.1.1 Questions

- See useful the HLM model (2.1) is under model misspecifications. That is, simulate data from a different model and try to fit it with HLM. To do this, you must find a survival analysis model in the literature – with meaningful parameters – as a benchmark.
- One of the features of survival analysis data is the censoring of observations. That is, patients commonly drop out of the study before their survival time is recorded, such that we only know that $z > t$ for such patients, not the value of z itself.

  Develop a methodology to fit the HLM model with censored observations. One suggestion is to adapt the EM algorithm from HW3 for probit regression. You can use the fact that if $X \sim \mathcal{N}(0,1)$ with CDF $\Phi(x)$, then

$$E[X \mid X > a] = \frac{1}{\sqrt{2\pi}}\frac{e^{a^2/2}}{\Phi(-a)}, \qquad E[X^2 \mid X < a] = 1 + \frac{a}{\sqrt{2\pi}}\frac{e^{a^2/2}}{\Phi(-a)}.$$

- The `survival` package in `R` contains the dataset `colon` on the effect of chemotherapy on survival with colon cancer. Censoring in `colon` is denoted by the variable `status`, with

0 corresponding to a censored observation. You can fit both CPH and AFT models with `survival` using the functions `coxph` and `survreg`, respectively. See which of these models best explains the data.

- Expand the model to include heavy tails, i.e.,

$$\frac{z_i - x_i'\beta}{\exp(\frac{1}{2}w_i'\gamma)} \,|\, \beta, \gamma, \nu \overset{\text{iid}}{\sim} t_{(\nu)}.$$

2.2 Stochatic Volatility Modeling for Financial Markets

The file `stat440-finance.csv` consists of daily observations of three financial assets between January 2, 1990 and February 28, 2014. The first few rows of the dataset are:

```
        date   gspc msft   vix
1 1990-01-02 359.69 0.45 17.24
2 1990-01-03 358.76 0.45 18.19
3 1990-01-04 355.67 0.47 19.22
4 1990-01-05 352.20 0.45 20.11
5 1990-01-08 353.79 0.46 20.26
6 1990-01-09 349.62 0.46 22.20
```

Here `date` is the date, `gspc` is the S&P 500 stock market index, `msft` is the stock price of Microsoft, and `vix` is a measure of the implied volatility associated with `gspc`. Historically, the Black-Scholes model for the log of a financial asset $X_t = \log(S_t)$ on day t is given by

$$X_{t+1} = X_t + \alpha + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

A *stochastic* volatility model instead proposes that

$$X_{t+1} = X_t + (\alpha + V_t) + V_t^{1/2}\epsilon_t,$$

where $V_t > 0$ is another stochastic process – the volatility – which determines the standard deviation of X_t (and also partly its mean). Typically, V_t is assumed to be unobserved which makes computations extremely challenging. Here, we are going to investigate what happens if V_t is related to some observable measure of volatility, such as `vix`.

Let $X_t = \log(\texttt{gspc})$ and $V_t = \texttt{vix}$ on day t. A discrete-time stochastic volatility model for these data is

$$\begin{aligned} X_{t+1} &= X_t + (\alpha + \kappa V_t) + V_t^{1/2}\epsilon_{1t} \\ A_{t+1} &= A_t - \gamma(A_t - \mu) + \epsilon_{2t}, \end{aligned} \quad , \quad \epsilon_t = (\epsilon_{1t}, \epsilon_{2t}) \overset{\text{iid}}{\sim} \mathcal{N}_2(0, \Sigma) \tag{2.2}$$

where $A_t = \log(V_t)$ and $\gamma > 0$. The model parameters are $\theta = (\alpha, \kappa, \gamma, \mu, \Sigma)$. For data $X = (X_0, \ldots, X_T)$ and $A = (A_0, \ldots, A_T)$, we can estimate θ by the following method. Let

$$Y_1 = \begin{bmatrix} \dfrac{X_1 - X_0}{V_0^{1/2}} \\ \vdots \\ \dfrac{X_T - X_{T-1}}{V_{T-1}^{1/2}} \end{bmatrix}, \quad Y_2 = \begin{bmatrix} A_1 - A_0 \\ \vdots \\ A_T - A_{T-1} \end{bmatrix}, \quad Z_1 = \begin{bmatrix} V_0^{-1/2} & V_0^{1/2} \\ \vdots & \vdots \\ V_{T-1}^{-1/2} & V_{T-1}^{1/2} \end{bmatrix}, \quad Z_2 = \begin{bmatrix} -A_0 & 1 \\ \vdots & \vdots \\ -A_{T-1} & 1 \end{bmatrix},$$

$$\beta_{4\times 1} = \begin{bmatrix} \alpha & \kappa & \gamma & \gamma\mu \end{bmatrix}'.$$

A convenient choice of prior for β and Σ is

$$\begin{aligned} \Sigma &\sim \text{Inv-}\mathcal{W}_2(\Psi, \nu) \\ \beta \,|\, \Sigma &\sim \mathcal{N}_4(\lambda, \Omega). \end{aligned}$$

In this case, the conditional posteriors are

$$\begin{aligned} \Sigma \,|\, \beta, A, X &\sim \text{Inv-}\mathcal{W}_2\left(\Psi + \hat{D}, T + \nu\right) \\ \beta \,|\, \Sigma, A, X &\sim \mathcal{N}_4\left(B\lambda + (1-B)\hat{\beta}, (1-B)\hat{V}\right), \end{aligned}$$

where

$$\hat{D} = \begin{bmatrix} (Y_1 - Z_1\beta_1)'(Y_1 - Z_1\beta_1) & (Y_1 - Z_1\beta_1)'(Y_2 - Z_2\beta_2) \\ (Y_2 - Z_2\beta_2)'(Y_1 - Z_1\beta_1) & (Y_2 - Z_2\beta_2)'(Y_2 - Z_2\beta_2) \end{bmatrix}, \quad \hat{V}^{-1} = \left(\Sigma^{-1} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) \circ \begin{bmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{bmatrix}$$

$$\hat{\beta} = \hat{V}\left(\left(\Sigma^{-1} \otimes \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \circ \begin{bmatrix} Z_1'Y_1 & Z_1'Y_2 \\ Z_2'Y_1 & Z_2'Y_2 \end{bmatrix}\right) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad B = \hat{V}(\hat{V} + \Omega)^{-1},$$

and $\otimes$ and $\circ$ respectively denote the Kronecker product and (elementwise) Hadamard product between two matrices. For the improper prior

$$\pi(\beta, \Sigma) \propto |\Sigma|^{(\nu+2+1)/2},$$

the conditional posteriors have the same form but with $B = 0$ and $\Psi = 0$.

2.2.1 Questions

- Plot the `gspc` and `vix` indices and pick two time periods that look interesting to you. Fit the stochastic volatility model (2.2) to each of the time periods and compare the estimates of θ. What features of the data in each period cause the estimates of θ to differ?

- An important quantity of interest for models such as (2.2) is the ability to make accurate 1-day and 30 day predictions for X_t. Pick a time period of 252 days (1 year) to "train" the model. Let's call the data for this time period $X_{1:252} = (X_1, \ldots, X_{252})$ and $A_{1:252} = (A_1, \ldots, A_{252})$. Based on this data, the d-day predictive distribution of X_t is

$$p(X_{252+d} \,|\, X_{1:252}, A_{1:252}) = \int p(X_{252+d} \,|\, X_{1:252}, A_{1:252}, \theta) p(\theta \,|\, X_{1:252}, A_{1:252}) \, \mathrm{d}\theta \tag{2.3}$$

To obtain a histogram of (2.3), use the following method:

for $i = 1$ to `nreps`
  $\theta \sim p(\theta \,|\, X_{1:252}, A_{1:252})$
  for $t = 0$ to $d - 1$
    $(\epsilon_{1t}, \epsilon_{2t}) \sim \mathcal{N}_2(0, \Sigma)$
    $X_{252+t+1} = X_{252+t} + (\alpha + \kappa V_{252+t}) + V_{252+t}^{1/2} \epsilon_{1t}$
    $A_{252+t+1} = A_{252+t} - \gamma(A_{252+t} - \mu) + \epsilon_{2t}$
  end
  $X_i^{(d)} \leftarrow X_{252+d}$
end

Find the quantile of where the real value of log(`gspc`) lies on day $252 + d$. Now, refit the model using on extra day, i.e., repeat the procedure above to obtain the histogram of $p(X_{253+d} \,|\, X_{1:253}, A_{1:253})$. Find the quantile for the real value of log(`gspc`) on day $253 + d$ and move on to predict $p(X_{254+d} \,|\, X_{1:254}, A_{1:254})$. Do this until you have d-day quantiles for several hundred days. How often is the real value of log(`gspc`) way out in the tails of the predicted distribution $p(X_{T+d} \,|\, X_{1:T}, A_{1:T})$?

- Use any model you like to jointly describe the time series for `gspc` and `msft` – perhaps adapting (2.2), perhaps also including `vix` if that seems to help. Assess whether including `msft` in the model for `gspc` can make better 1-day predictions about `gspc`. You may wish to consider the notion of Granger causality.

2.3 Efficient Simulation of Normals and Inverse-Wisharts

Many important models based on Normal data have posterior distributions which can be efficiently sampled using Normals and Inverse-Wisharts. For example:

1. **Multivariable regression.** Let $X_{n\times p} = [x_{ij}]$ and $Y_{n\times q} = [y_{ij}]$ be matrices of p predictors and q responses for each of n observations. For a matrix of errors $E_{n\times q} = [\epsilon_{ij}]$, the multivariate regression model is

$$Y = X\beta + E, \quad (\epsilon_{i1}, \ldots, \epsilon_{iq}) \overset{\text{iid}}{\sim} \mathcal{N}_q(0, V), \tag{2.4}$$

where $\beta_{p\times q} = [\beta_{ij}]$ and $V_{q\times q}$ are unknown parameters to be estimated from the data. The conjugate prior for this model is

$$V \sim \text{Inv-}\mathcal{W}_q(\Psi, \nu)$$
$$\text{vec}(\beta) \,|\, V \sim \mathcal{N}_{pq}\left\{\text{vec}(\Lambda), V \otimes \Omega^{-1}\right\},$$

where $\text{vec}(\cdot)$ is the vectorization operator, $\otimes$ is the Kronecker product, and Inv-$\mathcal{W}_q(\Psi, \nu)$ is an Inverse-Wishart distribution. The parameters of the prior are $\Psi_{q\times q}$, $\nu_{1\times 1}$, $\Lambda_{p\times q}$, and $\Omega_{p\times p}$. For this choice of prior, the posterior distribution is

$$V \,|\, Y, X \sim \text{Inv-}\mathcal{W}_q(\Psi + S + C, \nu + n)$$
$$\text{vec}(\beta) \,|\, V, Y, X \sim \mathcal{N}_{pq}\left\{\text{vec}\left[A\Lambda + (I - A)\hat{\beta}\right], V \otimes (X'X + \Omega)^{-1}\right\},$$

where

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad S = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$
$$A = (X'X + \Omega)^{-1}\Omega \quad C = \hat{\beta}'(X'X)\hat{\beta} + \Lambda'\Omega\Lambda - (X'X\hat{\beta} + \Omega\Lambda)'(X'X + \Omega)^{-1}(X'X\hat{\beta} + \Omega\Lambda).$$

2. **Hierarchical Normal Models.** The model is

$$y_i \,|\, \mu_i, V_i \overset{\text{ind}}{\sim} \mathcal{N}_q(\mu_i, V_i), \qquad \mu_i' \overset{\text{iid}}{\sim} \mathcal{N}_q(x_i'\beta, A),$$

where the V_i and x_i are known matrices and vectors of size $q \times q$ and $p \times 1$, and μ_i are unknown random effects, and $\beta_{p\times q}$ and $A_{q\times q}$ are the model parameters. Letting $Y = (y_1, \ldots, y_n)$, $X = (x_1, \ldots, x_n)$ and $\mu = (\mu_1, \ldots, \mu_n)$, we have seen in HW1 that

$$\mu_i \,|\, Y, \beta, A \overset{\text{ind}}{\sim} \mathcal{N}_q\Big(B_i x_i'\beta, (I - B_i)y_i, (I - B_i)V_i\Big),$$

where $B_i = V_i(V_i + A)^{-1}$. On the other hand, note that

$$\ell(\beta, A \,|\, Y, \mu) = \ell(\beta, A \,|\, \mu),$$

Such that conditioned on everything else, we can consider the likelihood for β and A as being equivalent to the likelihood for

$$\mu = X\beta + [\epsilon_{ij}]_{n\times q}, \quad (\epsilon_{i1}, \ldots, \epsilon_{iq}) \overset{\text{iid}}{\sim} \mathcal{N}_q(0, A),$$

which is precisely the multivariable regression model (2.4).

3. Another example of (2.4) is for autoregressive time series models. Let $x_0, \ldots, x_N$ denote q-dimensional time series observations from an autoregressive model of order d:

$$x_t = \sum_{i=1}^{d} x_{t-i}A_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, V),$$

where the A_i and V are $q \times q$ parameter matrices to be estimated from the data. Ignoring the effect of the first d observations (which is reasonable if $N \gg d$) leads to a likelihood of the form (2.4), where

$$Y_{N-d,q} = \begin{bmatrix} x_d \\ x_{d+1} \\ \vdots \\ x_N \end{bmatrix}, \quad X_{N-d,qd} = \begin{bmatrix} x_{d-1} & \cdots & x_0 \\ x_d & \cdots & x_1 \\ \vdots & & \vdots \\ x_{N-1} & \cdots & x_{N-d} \end{bmatrix}, \quad B_{qd\times q} = \begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix}$$

A central component of each of these models is an effective implementation of the so-called Normal-Inverse-Wishart distribution on random variables X and V:

$$\begin{aligned} V &\sim \text{Inv-}\mathcal{W}_d(\Psi, \nu) \\ X \,|\, V &\sim \mathcal{N}_d(\mu, V/\kappa). \end{aligned}$$

2.3.1 Questions

1. Implement simulation and density evaluation algorithms for the Normal-Inverse-Wishart distribution. For the former, use the Bartlett decomposion to generate L, the lower Choleski factor of V^{-1}: $LL' = V^{-1}$. Next, simulate $X = \mu + L^{-1}Z$ where $Z \sim \mathcal{N}_d(0, I)$ using the method of forward substitution for linear systems with triangular matrices. Hard-code this in `C++` with an interface to `R` using the `Rcpp` package.

2. Use this to perform either either a multivariable regression or fit an autoregressive model to a dataset of your choice, and comment on the analysis.

3. If you don't want to analyze a dataset, implement a Gibbs sampler from the multivariate hierarchical model. Once again, please use `Rcpp`. Simulate data with $n = 27$, $x_i = 1$, $\beta = (0, 0)$,

$$V_i = \begin{bmatrix} 148.84 & 142.74 \\ 142.74 & 490.63 \end{bmatrix}, \quad A = \begin{bmatrix} 3.38 & -.77 \\ -.77 & 2.55 \end{bmatrix}.$$

These values are taken from the paper by Everson & Morris (2000): "Inference for multivariate normal hierarchical models". If you do want to analyze data, you can use the data on 27 hospitals provided in the paper.