Empirical Analysis II 2020 Problem Set 1

Chase Abram, Jeanne Sorin*

July 5, 2020

1. *Suppose there is a sample of data on de-meaned log prices $p_t \in \mathbb{R}$ and de-meaned log quantities $q_t \in \mathbb{R}$, $t = 1, \ldots, T$. Assume they are generated by unobserved iid supply and demand shocks $\epsilon_{S,t} \in \mathbb{R}$ and $\epsilon_{D,t} \in \mathbb{R}$, which are uncorrelated and normalized to have unit variance. Assume more specifically, that supply and demand shocks follow a normal distribution*

$$\begin{bmatrix} \epsilon_{D,t} \\ \epsilon_{S,t} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{I})$$

where $\mathbf{I}$ is the 2×2 identity matrix. The data is then assumed to be generated according to

$$\begin{bmatrix} p_t \\ q_t \end{bmatrix} = \mathbf{A} \begin{bmatrix} \epsilon_{D,t} \\ \epsilon_{S,t} \end{bmatrix}$$

where the unknown 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} p_D & p_S \\ q_D & q_S \end{bmatrix}$$

encodes the movements of prices and quantities due to a positive one-standard deviation demand shock in the first column and due to a positive one-standard deviation supply shock in the second column.

(i) *Given $\mathbf{A}$, calculate the distribution for $[p_t, q_t]'$, i.e., calculate their variance-covariance matrix Σ.*

We are given that the shock vector is distributed $\mathcal{N}(\mathbf{0}, \mathbf{I})$, so since $[p_t, q_t]'$ is simply the shock vector premultiplied by $\mathbf{A}$, the Σ will simply be $\mathbf{AIA}' = \mathbf{AA}'$.

$$\begin{aligned} \begin{bmatrix} \varepsilon_{D,t} \\ \varepsilon_{D,t} \end{bmatrix} &\sim \mathcal{N}(0, \mathbf{I}) \\ \begin{bmatrix} p_d & p_s \\ q_d & q_s \end{bmatrix} \begin{bmatrix} \varepsilon_{D,t} \\ \varepsilon_{D,t} \end{bmatrix} &\sim \mathcal{N}(0, \begin{bmatrix} p_d & p_s \\ q_d & q_s \end{bmatrix} \mathbf{I} \begin{bmatrix} p_d & q_d \\ p_s & q_s \end{bmatrix}) \\ &\sim \mathcal{N}(0, \begin{bmatrix} p_d^2 + p_s^2 & q_d p_d + p_s q_s \\ q_d p_d + q_s p_s & q_s^2 + q_d^2 \end{bmatrix}) \end{aligned}$$

*This solution set benefited greatly from the comments and points raised by Manav Chaudhary, George Votja, and Dan Ehrlich.

(ii) *Given* **A**, *calculate the demand elasticity and supply elasticity (and define them first).*

The demand elasticity is $\frac{\%\Delta Q_D}{\%\Delta P_D}$, and since we have log data, these percentage changes will simply be the "average" differentials of the log data, so we can consider the price and quantity changes due to supply/demand shocks. For demand, we need supply shocks to identify movement along the demand curve, so the demand elasticity is $\varepsilon_D = \frac{q_S}{p_S}$. Analogously, we use demand shocks to identify movement along the supply curve, so the supply elasticity ($\frac{\%\Delta Q_S}{\%\Delta P_S}$) is $\varepsilon_S = \frac{q_D}{p_D}$.

More precisely, for the price elasticity of demand, we don't observe q_d, p_d but instead

$$\begin{aligned} q_t &= q_s\varepsilon_{s,t} + q_d\varepsilon_{D,t} \\ p_t &= p_s\varepsilon_{s,t} + p_d\varepsilon_{D,t} \end{aligned}$$

As we study movements along the demand curve we set $\varepsilon_{D,t} = 0$ so that movements of q_t and p_t actually capture movements of q_d and p_d:

$$\begin{aligned} q_t &= q_s\varepsilon_{s,t} = q_d \\ p_t &= p_s\varepsilon_{s,t} = p_d \end{aligned}$$

Rewriting the price elasticity of demand and using the chain rule

$$\begin{aligned} \frac{\partial q_d}{\partial p_d} &= \frac{\partial q_t}{\partial \varepsilon_{s,t}} \frac{\partial \varepsilon_{s,t}}{\partial p_t} \\ &= q_s \cdot \frac{1}{p_s} \\ &= \frac{q_s}{p_s} \end{aligned}$$

Similarly for the price elasticity of supply.

$$\begin{aligned} \frac{\partial q_s}{\partial p_s} &= \frac{\partial q_t}{\partial \varepsilon_{d,t}} \frac{\partial \varepsilon_{d,t}}{\partial p_t} \\ &= q_d \cdot \frac{1}{p_d} \\ &= \frac{q_d}{p_d} \end{aligned}$$

(iii) *Let*

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12} & \hat{\Sigma}_{22} \end{bmatrix}$$

be the MLE ("maximum likelihood estimator") for the variance-covariance matrix of $[p_t, q_t]'$. *Suppose that* $\hat{\Sigma}$ *is positive definite. Impose, that* $p_S = 0$. *Given* $\hat{\Sigma}$, *state the set of all MLE's for* $\mathbf{A}$, *and calculate them explicitly. Is* $\mathbf{A}$ *identified? If not, how many solutions are there?*

We are looking for $\mathbf{A}$s such that $\hat{\Sigma} = \mathbf{A}\mathbf{A}'$, and the upper right entry of $\mathbf{A}$ is zero. Therefore we can explicitly solve for the upper left entry of $\hat{A}$, since $\hat{\Sigma}_{11} = A_{11}^2$, so $A_{11} = \pm\sqrt{\hat{\Sigma}_{11}}$.[1] This knowledge allows us to pin down the bottom left entry, since $A_{11}A_{21} = \hat{\Sigma}_{12}$. Then we are left with a two possible solutions for A_{22} since $A_{21}^2 + A_{22}^2 = \hat{\Sigma}_{22}$.

Explicitly, we find 4 elements in the set of MLEs for $\mathbf{A}$, so $\mathbf{A}$ is not identified:

$$MLE(\mathbf{A}) = \left\{ \begin{bmatrix} a_1\sqrt{\hat{\Sigma}_{11}} & 0 \\ \frac{\hat{\Sigma}_{12}}{a_1\sqrt{\hat{\Sigma}_{11}}} & a_2\sqrt{\hat{\Sigma}_{22} - \frac{\hat{\Sigma}_{12}^2}{\hat{\Sigma}_{11}}} \end{bmatrix} : a_1 \in \{-1, 1\}, a_2 \in \{-1, 1\} \right\}$$

To double-check that our square roots don't cause problems, note that the positive definiteness of $\hat{\Sigma}$ implies $\hat{\Sigma}_{ii} = e_i'\hat{\Sigma}e_i > 0$, and the bottom right element of $\hat{\mathbf{A}}$ is real because $|\hat{\Sigma}| = \hat{\Sigma}_{11}\hat{\Sigma}_{22} - \hat{\Sigma}_{12}^2 > 0$ implies the term under the $\sqrt{\cdot}$ in the bottom right is positive (along with $\hat{\Sigma}_{11} > 0$).

(iv) *Suppose that* $\hat{\Sigma} = \mathbf{I}$. *Without imposing* $p_S = 0$, *state the set of all MLE's for* $\mathbf{A}$ *and parameterize them explicitly. Is* $\mathbf{A}$ *identified? If not, what is the dimensionality of the set?*

Since we removed the assumption $p_S = 0$, the solution set gets a little funkier. We need

$$\begin{aligned} \mathbf{I} &= \hat{\mathbf{A}}\hat{\mathbf{A}}' \\ &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} \\ &= \begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & c^2 + d^2 \end{bmatrix} \end{aligned}$$

This is a non-linear system of 3 equations in 4 unknowns. The non-linearity means that there is no standard dimensionality in the sense that the set of MLEs is not a vector space. However, we may allow a to be a free variable over $[-1, 1]$, and then the magnitude (but not sign) of b will be pinned down by $a^2 + b^2 = 1$. Given a choice for the sign of b, the magnitudes of c and d will also be pinned down, but not necessarily their signs (though choosing one sign will imply a choice for the other). So the dimensionality of the set is at least one-dimensional (though constrained), and

[1] I will use $\sqrt{\cdot}$ and/or fractional exponents to denote the *positive* square root, and will be explicit about including the negative square root (either through $\pm$ or a sign parameter).

arguably something more than that, because signs may also be free, for some given values of a. To be precise, I believe the Hausdorff dimension of the set is 1. So $\mathbf{A}$ is not identified.

Explicitly the set of MLEs for $\mathbf{A}$ is

$$MLE(\mathbf{A}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} : \begin{cases} a \in [-1,1] \\ b^2 = 1 - a^2 \\ ac + bd = 0 \\ d^2 = 1 - c^2 \end{cases} \right\}$$

We can parameterize this set by allowing $t_1 \in [-1,1]$, $t_2 \in \{-1,1\}$, and $t_3 \in \{-1,1\}$ and considering the parameterization γ:

$$\gamma(t_1,t_2,t_3) = \begin{bmatrix} t_1 & t_2(1-t_1^2)^{\frac{1}{2}} \\ t_3(1-t_1^2)^{\frac{1}{2}} & -t_1t_2t_3 \end{bmatrix}$$

Intuitively, this parameterizes four curves in the 2×2 matrix space $\mathbb{R}^{2\times 2}$

(v) *Impose, additionally, that the supply elasticity is strictly positive and the demand elasticity is strictly negative (i.e., that the supply curve is positively sloped and the demand curve is negatively sloped). How do these sign restrictions narrow the set of all MLE's for A? Parameterize the set explicitly.*

Supply elasticity being positive means $\frac{q_D}{p_D} > 0$, and demand elasticity being negative means $\frac{q_S}{p_S} < 0$. The strictness of these criteria rule out $t_1 \in \{-1,0,1\}$. Furthermore the sign restrictions mean that $\frac{t_1}{t_3} > 0$, and $-t_1t_3 < 0$ (the second relation is redundant).

Then our new parameterization[2], ξ, is defined by letting $t_1 \in (-1,1)\setminus\{0\}$, $t_2 \in \{-1,1\}$:

$$\xi(t_1,t_2) = \begin{bmatrix} t_1 & t_2(1-t_1^2)^{\frac{1}{2}} \\ \frac{t_1}{|t_1|}(1-t_1^2)^{\frac{1}{2}} & -\frac{t_1^2}{|t_1|}t_2 \end{bmatrix}$$

So we still cannot identify $\mathbf{A}$, but we have reduced the set of MLEs so that we now only have two curves in $\mathbb{R}^{2\times 2}$ (each also now missing a point in the "middle"), as opposed to the four we had above.

(vi) *For the cracks and for extra points: In the last part, suppose that $\hat{\Sigma}$ was some positive semi-definite matrix, though not necessarily the identity matrix. When would be the set of identified A have only finitely many entries, and how many?*

[2] I recognize that using periodic functions is another way to do this, and that way might even seem more "natural", but this was how my brain worked it out. It is equivalent.

First, note that if there exists a solution $\mathbf{A}$, then

$$\hat{\Sigma}' = (\mathbf{A}\mathbf{A}')' = \mathbf{A}\mathbf{A}' = \hat{\Sigma}$$

So $\hat{\Sigma}$ must be symmetric. Now consider there are three cases for a positive semi-definite matrix: both eigenvalues are positive, one eigenvalue is positive and the other is zero, or both eigenvalues are zero.

The key point is then that symmetric matrices are diagonalizable, which means we only have to care about matrices similar to the form solved above!

$$\begin{aligned}
\hat{\Sigma} &= P^{-1}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}P && \text{(For some invertible } P\text{)}\\
&= P^{-1}\mathbf{B}\mathbf{B}'P\\
&= (P^{-1}\mathbf{B}P)(P^{-1}\mathbf{B}'P)\\
&= \mathbf{A}\mathbf{A}'
\end{aligned}$$

Where λ_i are the eigenvalues of $\hat{\Sigma}$, and $P' = P^{-1}$ by the symmetry of $\hat{\Sigma}$ (this is a theorem somewhere). Now, by comparing the first and second lines, we only have to worry about finding $\mathbf{B}$ such that

$$\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix} = \mathbf{B}\mathbf{B}'$$

If both eigenvalues are positive, then the solution set (for $\mathbf{B}$) will be similar to above, where $t_1 \in [-\sqrt{\lambda_1}, \sqrt{\lambda_1}]$, $t_2 \in \{-1, 1\}$ and $t_3 \in \{-1, 1\}$.

$$\gamma(t_1, t_2, t_3) = \begin{bmatrix} t_1 & t_2(\lambda_1 - t_1^2)^{\frac{1}{2}} \\ t_3\left(\frac{t_1^2\lambda_2 - \lambda_1\lambda_2}{2t_t^2 - \lambda_1}\right)^{\frac{1}{2}} & -t_1t_2t_3\left(\frac{\lambda_2}{2t_1^2 - \lambda_1}\right)^{\frac{1}{2}}\end{bmatrix}$$

When we impose the elasticity restriction, we kill one of the free signs.

$$\gamma(t_1, t_2, t_3) = \begin{bmatrix} t_1 & t_2(\lambda_1 - t_1^2)^{\frac{1}{2}} \\ \frac{t_1}{|t_1|}\left(\frac{t_1^2\lambda_2 - \lambda_1\lambda_2}{2t_t^2 - \lambda_1}\right)^{\frac{1}{2}} & -\frac{t_1^2}{|t_1|}t_2\left(\frac{\lambda_2}{2t_1^2 - \lambda_1}\right)^{\frac{1}{2}}\end{bmatrix}$$

If exactly one eigenvalue is positive (without loss of generality say λ_1), then $a = b = 0$, but we have freedom within $c^2 + d^2 = 1$, so for $t_1 \in [-\sqrt{\lambda}_2, \sqrt{\lambda}_2]$, $t_2 \in \{-1, 1\}$.

$$\gamma(t_1, t_2) = \begin{bmatrix} 0 & 0 \\ t_1 & t_2(\lambda_2 - t_1^2)^{\frac{1}{2}} \end{bmatrix}$$

Imposing the elasticity sign restriction kills the sign freedom

$$\gamma(t_1) = \begin{bmatrix} 0 & 0 \\ \frac{t_1^2}{|t_1|} & -(\lambda_2 - t_1^2)^{\frac{1}{2}} \end{bmatrix}$$

If both eigenvalues are zero, then the only solution is $\mathbf{B} = \mathbf{0}$. But then we cannot make the elasticity restrictions strict...

To sum up, after we impose the elasticity restrictions

- $\hat{\Sigma}$ asymmetric implies no solutions
- $\hat{\Sigma}$ symmetric and positive definite implies a continuum of solutions
- $\hat{\Sigma}$ symmetric with exactly one positive eigenvalue implies a continuum of solutions
- $\hat{\Sigma}$ symmetric and both eigenvalues are zero implies sort of 1 solution in the $\mathbf{0}$ matrix, but really 0 solutions, because then we cannot get strict elasticity signs

These results are consistent with those found in Prof. Uhlig's 2017 paper (though this paper was only brought to my attention after I finished the problem). The final point, which is I suppose an important one, thought I would argue it follows quite obviously from my parameterizations, is that another way we can get identification (1 solution), or at least finitely many values, is to, in any of the above scenarios, on any element including a t_1, impose an exact choice for its value. For example, if I impose in the example immediately above that $A_{21} = 0$, then I immediately get exactly one solution to the problem, because the elasticity restrictions rule out the others. In the more general cases above, this same trick still kills the continuum of solutions, but if there are still free signs, then there will be a multiplicity of solutions (maybe 2, maybe 4).

2. *Consider the problem of rolling two fair dice, X_1 at date $t = 1$ and X_2 at date $t = 2$. Let $\omega \in \Omega = \{(1,1),(1,2),\ldots,(6,6)\}$ be the outcome of these two dice roles, i.e. $X_1(\omega) = \omega_1$ and $X_2(\omega) = \omega_2$. Let $S = X_1 + X_2$ be the sum of the two dice roles.*

(i) *What is the smallest σ-algebra $\mathcal{F}_0$ on Ω?*

One definition of a σ-algebra $\mathcal{F}$ is that it is a set of subsets of Ω such that

(a) $\varnothing \in \mathcal{F}$

(b) $A \in \mathcal{F}$ implies $\Omega \setminus A \in \mathcal{F}$

(c) $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$ implies $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$

From (a) and (b), we deduce that $\Omega \setminus \varnothing \equiv \Omega \in \mathcal{F}$, so any σ-algebra must contain the sets $\varnothing$ and Ω. If these two sets alone comprise a σ-algebra, then this must be the smallest σ-algebra. They do!

Immediately, since the empty set and whole set are complements, we can see that (a) and (b) hold. Now consider a countable union of sets $\{A_i\}$ such that each set is either $\varnothing$ or Ω. If all of these sets are $\varnothing$, then the union is $\varnothing \in \mathcal{F}$, and if one or more of these sets is Ω, then the union is $\Omega \in \mathcal{F}$. So $\mathcal{F}_0 \equiv \{\varnothing, \Omega\} \equiv \{\varnothing, \{(1,1),(1,2),\ldots,(6,6)\}\}$ is the smallest σ-algebra on Ω.

(ii) *What is the smallest σ-algebra $\mathcal{F}_1$ on Ω, so that $X_1 : \Omega \to \mathbb{R}$ is measurable?*

For X_1 to be measurable ($\mathcal{F}_1$-measurable), we need that for any Borel set of $\mathbb{R}$, the pre-image of X_1 is in $\mathcal{F}_1$: for all $B \in \mathcal{B}(\mathbb{R})$, $X_1^{-1}(B) \in \mathcal{F}_1$. Since X_1 can only take on finitely many values, $\{1,\ldots,6\}$, we can construct Borel sets to capture each value individually, i.e. $B_i = (i - \frac{1}{2}, i + \frac{1}{2})$, so that $X_1^{-1}(B_i) = X_1^{-1}(\{i\}) = \{(i,1),(i,2),\ldots,(i,6)\}$. This leads us to conjecture that the $\mathcal{F}_1 \equiv \sigma(\{\{(i,1),(i,2),\ldots,(i,6)\} \mid i \in \{1,\ldots,6\}\})$, which is the smallest σ-algebra generated by the sets that have the same value for the first entry, and all possible entries for the second entry (for example $\{(i,1),(i,2),\ldots,(i,6)\}$ is $X_1^{-1}(\{i\})$). By the definition of σ-algebra, this set will also contain all the possible unions, such as the set which has all ω with a 1 or a 3 in the first entry.

All that remains to check is that we may take an arbitrary $B \in \mathcal{B}(\mathbb{R})$ and have $X_1^{-1}(B) \in \mathcal{F}_1$. Let $B \in \mathcal{B}(\mathbb{R})$. Then B contains some subset A of $\{1,2,\ldots,6\}$ (or none them, in which case $X_1^{-1}(B) = \varnothing$). Then $X_1^{-1}(B) = X_1^{-1}(A) = X_1^{-1}(\bigcup_{i\in A}\{i\}) = \bigcup_{i\in A} X_1^{-1}(\{i\}) \in \mathcal{F}_1$ (by definition of generated σ-algebra). So X_1 is $\mathcal{F}_1$-measurable, and since any σ-algebra such that X_1 is measurable must contain $\{\{(i,1),\ldots,(i,6)\} \mid i \in \{1,\ldots 6\}\}$ (by above reasoning), $\mathcal{F}_1 \equiv \sigma(\{\{(i,1),\ldots,(i,6)\} \mid i \in \{1,\ldots 6\}\})$ **must be the smallest such σ-algebra.**

(iii) *What is the largest σ-algebra $\mathcal{F}_2$ on Ω?*

The largest possible σ-algebra would be the power set 2^Ω (the set containing all subsets of Ω), and we can show that $\mathcal{F}_2 \equiv 2^\Omega$ is a σ-algebra. First, $\varnothing \in 2^\Omega$. Second, if $A \in 2^\Omega$, then trivially $\Omega \setminus A \in 2^\Omega$, because 2^Ω contains all possible subsets of Ω. Third, again trivially we have that for $A_i \in 2^\Omega$, $i \in \mathbb{N}$, then $\bigcup_{i\in\mathbb{N}} A_i \in 2^\Omega$, because 2^Ω contains all possible subsets of Ω, and any union of subsets is a subset of Ω. So 2^Ω is a σ-algebra, and it is the largest possible because no subsets of Ω exists outside of 2^Ω, so $\mathcal{F}_2 \equiv 2^\Omega$.

(iv) *Comment on the statement: "$\mathcal{F}_t$ represents the information known at date t.". Remark: $(\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2)$, a sequence of σ-algebras with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$, is called a filtration.*

This statement makes sense. At time zero, no information is known, other than that the outcome ω will be in Ω. So $\mathcal{F}_0 \equiv \{\varnothing, \Omega\}$ summarizes this information well, because it essentially tells us nothing about what we can expect the final outcome to be. If the outcome is $(5, 3)$, then we have no way of "narrowing down" our options for what ω is at time 0 (nothing has been thrown).

At time 1, we receive information about the first die roll, so obtain information about the outcome. Given a roll i, we are able to say within which subset of Ω, within $\mathcal{F}_1$, ω lives. For example, if the outcome is $(5, 3)$, then the information available at time 1 is that $\omega \in \{(5,1), (5,2), \ldots, (5,6)\}$. Even if we were given only partial information, such as that the roll was odd, we would have the information that $\omega \in \{(1,1), \ldots, (1,6), (3,1), \ldots, (3,6), (5,1), \ldots, (5,6)\}$.

At time 2, we receive the second die roll, and have full information, so we may narrow down where ω lives to a subset which is a singleton, which is ω. For example, if the outcome is $(5, 3)$, then at time 2 we know $\omega \in \{(5,3)\}$.

In summary $\mathcal{F}_t$ is information in the sense that it tells us how granular/specific we can be in saying what ω is, after time t.

(v) *Let μ be the counting measure on Ω, i.e.*

$$\mu(A) = \#A$$

counts the elements in A for all $A \in \mathcal{F}_2$. Consider the probability measure

$$\mathcal{P}(A) = P((X_1, X_2) \in A)$$

Find the density $f : \Omega \to \mathbb{R}$ with respect to the counting measure μ on Ω, so that

$$\mathcal{P}(A) = \int_A f(\omega)\mu(\mathrm{d}\omega)$$

The key detail is that the dice are fair, so that all outcomes are equally likely. There are 36 possible outcomes (6 for each die), so $f(\omega) = \frac{1}{36}$ for all $\omega \in \Omega$. We can easily check that this density works:

$$\begin{aligned}\int_A f(\omega)\mu(\mathrm{d}\omega) &= \sum_{\omega\in A} f(\omega)\mu(\{\omega\}) \\ &= \sum_{\omega\in A} \frac{1}{36}\cdot 1 \\ &= \frac{\#A}{36} \\ &= P((X_1, X_2)\in A) \\ &= \mathcal{P}(A)\end{aligned}$$

(vi) *Find a $\mathcal{F}_1$-measurable function $g:\Omega\to\mathbb{R}$, so that*

$$\int g(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega) = \int S(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega)$$

for all $A\in\mathcal{F}_1$, where $\mathbf{1}_A(\omega)=1$, if $\omega\in A$ and 0 else.

At time 1, we only have part of the total sum, but the second die is independent of the first, so we can "guess" that it will be equal to the average die roll of a fair die, which is 3.5. So our natural choice is $g(\omega)=\omega_1+\frac{7}{2}$. This function is clearly $\mathcal{F}_1$-measurable, because for any Borel set B of $\mathbb{R}$, some subset A of $\{4.5, 5.5, \ldots, 9.5\}$ will be contained in B (or none will), and by the same reasoning made in part (ii), the pre-image A will be contained in $\mathcal{F}_1$.

We may check that this satisfies the given equation:

$$\begin{aligned}\int g(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega) &= \sum_{\omega\in A} g(\omega)\mu(\{\omega\}) \\ &= \sum_{\omega\in A}(\omega_1+\frac{7}{2})\cdot 1 \\ &= \sum_{\omega\in A}\omega_1\cdot 1 + \sum_{\omega\in A}\frac{7}{2}\cdot 1 \\ &= \sum_{\omega\in A}\omega_1\cdot 1 + \sum_{\omega\in A}\omega_2\cdot 1 \\ &= \sum_{\omega\in A}(\omega_1+\omega_2)\cdot 1 \\ &= \sum_{\omega\in A} S(\omega)\cdot\mu(\{\omega\}) \\ &= \int S(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega)\end{aligned}$$

We are above able to say $\sum_{\omega\in A}\frac{7}{2} = \sum_{\omega\in A}\omega_2$ because $A \in \mathcal{F}_1$, and every set in $\mathcal{F}_1$ has this property, precisely because $\mathcal{F}_1$ gives no information about ω_2. In other words, regardless of how many different ω_1 values (say v) are in a given set in $\mathcal{F}_1$, we will always have the same distribution of ω_2 values (we will have v of each value in $\{1,\ldots,6\}$).

(vii) *Comment on the statement: "g is the conditional expectation $g = E_1[S]$ of S at date $t = 1$".*

This statement makes sense intuitively because g captures the expected sum of the two dice rolls given the outcome of the first dice ω_1. The conditional expectation $E_1[S]$ at time $t = 1$ has two components: w_1 which is known, and w_2 which is totally unknown, as the two rolls are independent, so our guess is a mixture of a known componenet and best guess for the unknown component.

More rigorously, this statement is mathematically true in that g is a $\mathcal{F}_1$-measurable function, which is integrable, and which agrees with the true sum $S(\omega)$ over sets $A \in \mathcal{F}_1$ in the sense

$$\int_A g(\omega)\mu(\mathrm{d}\omega) = \int_A S(\omega)\mu(\mathrm{d}\omega)$$

This is exactly the definition for the conditional expectation of S given $\mathcal{F}_1$[3].

[3]I referenced Billingsley's *Probability and Measure*.

3. (Sufficiency Principle) *Suppose that x_1 and x_2 are drawn independently from $N(\theta, 1)$, where θ is an unknown parameter. Let $\bar{x} = (x_1 + x_2)/2$ and let $x = (x_1, x_2)$. For notation and as convention, allow "densities" like $\mathbf{1}_z(x_1)$, say, which put a point mass of unity at $z = x_1$ and are zero for all other values of x_1. Also allow them as a piece of joint densities.*

(i) *Explicitly state the joint density $f(x \mid \theta)$.*

If we are given $\theta = \theta_0$, then x is comprised of two independent draws from $N(\theta_0, 1)$. Therefore x is a 2-dimensional normal random variable with no off-diagonal elements (actually an identity variance-covariance), and mean θ for both components: .

$$x \mid \theta \sim \mathcal{N}\Big(\begin{bmatrix} \theta \\ \theta \end{bmatrix}, \mathbf{I} \Big)$$

Therefore the density (with a small abuse of notation because little x is used to index the possible values for x) is (letting $\Theta = [\theta, \theta]'$)

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{(2\pi)^{\frac{2}{2}} |\mathbf{I}|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x - \Theta)' \mathbf{I}^{-1} (x - \Theta) \Big) \\ &= \frac{1}{2\pi} \exp\Big(-\frac{1}{2}(x - \Theta)'(x - \Theta) \Big) \end{aligned}$$

(ii) *Explicitly state the density $f(x_1 \mid \theta)$ and the density $f(\bar{x} \mid \theta)$.*

By itself, $x_1 \mid \theta \sim N(\theta, 1)$, so its density is just

$$f(x_1 \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(x_1 - \theta)^2 \Big)$$

For $\bar{x}$, we may use the fact that the sum of independent normal random variables will be normal with means and variances equal to the sums of their means and variances. We also use that scaling a normal random variable by α scales its mean by α and its variance by α^2.

$$\begin{aligned} \bar{x} \mid \theta &\sim \frac{1}{2}(x_1 + x_2) \mid \theta \\ &\sim \frac{1}{2}\mathcal{N}(2\theta, 2) \\ &\sim \mathcal{N}(\theta, \frac{1}{2}) \end{aligned}$$

Therefore

$$f(\bar{x} \mid \theta) = \frac{1}{\sqrt{\pi}} \exp\Big(-(\bar{x} - \theta)^2 \Big)$$

(iii) *Calculate and explicitly state the conditional density $f(x \mid x_1 = z, \theta)$, stating it as a density $f(x \mid x_1)$, if it can be done (can it?). Can you describe it as a normal distribution $\mathcal{N}(m, \Sigma)$ for some mean m and some variance-covariance matrix Σ?*

The trick is to notice that we essentially are just dealing with $f(x_2 \mid \theta)$, except now we are technically in $\mathbb{R}^2$, and we want to put zero weight where $x_1 \neq z$, but normally distribute the weight over the set where $x_1 = z$, using the facts that the mean for x_2 is θ and its variance is 1.

Therefore the density is

$$f(x \mid x_1 = z, \theta) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(x_2 - \theta)^2 \Big) \mathbf{1}_z(x_1)$$

This distribution can be described as a normal distribution, we just have to say that the variance for x_1 is 0. So

$$x \mid x_1 = z, \theta \sim \mathcal{N}\left(\begin{bmatrix} z \\ \theta \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

Note that $|\Sigma| = 0$, which is why we cannot write the density in the "standard" way, but must include the Dirac delta density $\mathbf{1}_z$. More intuitively, we can recall that constant random variables are normal random variables with zero variance; I think of them as "collapsed" distributions. Here, we are considering the intermediate case where there is "collapse" in the x_1 dimension, but the x_2 dimension looks "standard", so the support is $\{z\} \times \mathbb{R}$.

Note that we cannot write the density as a function of x_1 alone (we need θ). This is because x_1 alone gives no information about x_2, so θ is still informative.

(iv) *Calculate and explicitly state the conditional density $f(x \mid \bar{x}, \theta)$, stating it as a density $f(x \mid \bar{x})$, if it can be done (can it?). Can you describe it as a normal distribution $\mathcal{N}(m, \Sigma)$ for some mean m and some variance-covariance matrix Σ? (Hint: it may be easiest to write $x_j = \theta + \epsilon_j$ for $j = 1, 2$. Then, write $x_1 = \bar{x} + \nu_1$, and figure out the relationship between ν_1 and ϵ_1, ϵ_2.)*

We use a sort of variant on the hint that was more clear to us. We can use the above answers to condition, as long as we are careful to check our support. The idea is that in the numerator of the fraction below, we are considering almost the same density $f(x \mid \theta)$ except we also want the event $\bar{x} = \frac{x_1 + x_2}{2}$ to occur, so we need to tack on the Dirac delta "density" $\mathbf{1}_{\bar{x}}$. The denominator is just $f(\bar{x} \mid \theta)$. Normally in Bayesian analysis the normalizing constant might be nasty, and we would just normalize after finding the shape, but in this case the normalizing constant happens to be quite easy to find.

$$
\begin{aligned}
f(x|\bar{x},\theta) = f(x,\bar{x}|\frac{x_1+x_2}{2},\theta) &= \frac{f(x,\bar{x}|\theta)}{f(\bar{x}|\theta)} \\
&= \left[(\frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{2}[(x_1-\theta)^2+(x_2-\theta)^2]))/(\frac{\sqrt{2}}{\sqrt{2\pi}}\exp(-(\bar{x}-\theta)^2))\right]\mathbf{1}_{\bar{x}}(\frac{x_1+x_2}{2}) \\
&= \left[\frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{2}[(x_1-\theta)^2+(x_2-\theta)^2]))/(\frac{1}{\sqrt{\pi}}exp(-(\frac{x_1+x_2}{2}-\theta)^2)\right]\mathbf{1}_{\bar{x}}(\frac{x_1+x_2}{2}) \\
&= \left[\frac{1}{2\sqrt{\pi}}\exp\left(-\frac{1}{2}[(x_1-\theta)^2+(x_2-\theta)^2]+(\frac{x_1+x_2}{2}-\theta)^2\right)\right]\mathbf{1}_{\bar{x}}(\frac{x_1+x_2}{2}) \\
&= \frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{4}(x_1-x_2)^2)\mathbf{1}_{\bar{x}}(\frac{x_1+x_2}{2})
\end{aligned}
$$

We were able to write the density as $f(x \mid \bar{x})$, in this case, because if the $\bar{x}$ is known then we can cancel out the θs so they do not appear in the final form.

We can write this as a normal distribution by noting if either component increases, the other component must exactly decrease by that amount to preserve the mean. Additionally, the mean for both components will be $\bar{x}$, since our best guess would be for both to equal the given mean. Of course, their own variances don't change.

$$
x \mid \bar{x} \sim \mathcal{N}\left(\begin{bmatrix}\bar{x}\\ \bar{x}\end{bmatrix}, \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\right)
$$

Notice that $|\Sigma| = 0$, which explains why we were unable to write our "density" without using a Dirac delta fudge thing.

(v) *Based on your results, is $T(x) = \bar{x}$ a sufficient statistic? Is $T(x) = x_1$ a sufficient statistic?*

$T(x) = \bar{x}$ is a sufficient statistic while $T(x) = x_1$ is not. Why is that? $T(x) = \bar{x}$ summarizes all information available in the data to help you recover θ. Loosely speaking $\bar{x}$ is the most precise statement you can make about θ given your data. On the contrary, by considering $T(x) = x_1$, you lose information about x_2 that is not included in x_1 as both are drawn independently.

If we consider the technical definition given in class, we can note that $x \mid x_1$ is a distribution which still very much depends on θ (θ is in the pdf and the $\mathcal{N}(\cdot,\cdot)$ form), since x_1 is uninformative about x_2. On the other hand $x \mid \bar{x}$ is a distribution which only depends on $\bar{x}$ (neither the pdf nor the $\mathcal{N}(\cdot,\cdot)$ form contain a θ).

4. *Consider the testing procedure in example 3, where we stop at some $T = T^{max}$. Write a program to numerically answer the following questions (one decimal precision is fine) for $T^{max} \in \{50, 100, 200, 1000\}$.*

(i) *What is the size of the test per the announced stopping rule?*

Code is provided at the end of this question. I found these sizes by running a large number of trials with the given stopping rule (meaning either stop when rejected or when $T^{\max}$ is reached), then calculating the proportion of the time that the hypothesis $\theta = 0$ was rejected, for critical value 1.96.

The sizes are as follows

- 50: size $\approx .30$
- 100: size $\approx .35$
- 200: size $\approx .40$
- 1000: size $\approx .50$

(ii) *What is the appropriate critical value, for a test of size 5%?*

My algorithm here was similar to above, except that I started with critical value 1.96 for each $T^{\max}$, and incremented the critical value until the size was at or below 5%. The critical values are as follows:

- 50: critical value ≈ 2.77
- 100: critical value ≈ 2.86
- 200: critical value ≈ 2.94
- 1000: critical value ≈ 3.05

(iii) *For the cracks and extra points: What is the size of the test per the announced stopping rule and what is the appropriate 5% critical value for $T^{max} = \infty$?*

When $T^{\max} = \infty$, the size of the test will be 1, and for any given critical value rejection will occur with probability 1, so the critical value is ∞. Mathematically, consider that for each draw, the "worst" possible scenario would be if the average of the draws thus far were 0, because then there would be the lowest probability of rejection for that draw (the previous draws are biasing as heavily as they can for no rejection). Therefore, the probability of rejection on any given draw is greater than or equal to

$$\begin{aligned} P(\text{rej. at } t) &= P(|\bar{x}| > c) \\ &\geq P(A_t) \\ &= P(\frac{|x_t|}{t} > c) \\ &= P(|x_t| > tc) \\ &= 2(1 - \Phi(tc)) \end{aligned}$$

In this formulation, each rejection is an event independent of the previous draws (because we are always considering the "worst" possible scenario, so our event only depends on the current draw). Therefore, the second Borel-Cantelli Lemma tells us that if

$$\sum_{t=1}^{\infty} 2(1 - \Phi(tc))$$

diverges, then the $P(\limsup A_t) = 1$, so rejection occurs with probability 1. This unfortunately appears to be a dead-end, due to lack of closed-form analysis available, but I am mildly confident that this sum does diverge.

Consider instead the tail σ-algebra of the sequence $\{A_t\}$: $\bigcap_{t=1}^{\infty} \sigma(A_t, A_{t+1}, \ldots)$ (here we must allow for the idea that we will keep drawing, and just mark down each rejection that should have occurred). By Kolmogorov's Zero-One law, all events in this σ-algebra have probability 0 or 1. The event that, from some time onward, a rejection never occurs, lies in this σ-algebra. The same is true for, from some time onward, a rejection always occurs. If either of these events had probability 1, then the other would as well, since the zero mean but decreasing variance will just mean that as time goes on, $\bar{x}$ should vary less, centered on whatever value it is currently at, and there is certainly positive probability of rejecting at any given time. So then both events must have probability zero, because they cannot both have probability 1. But then there is probability 1 that, at some point in time, at least 1 rejection will occur. So we have the results stated at the beginning, i.e. size= 1, and for any c rejection will occur with probability 1 (so at least 5% of the time), so $c = \infty$.

Here is the code in `Julia`:

```
using Pkg
Pkg.add("Distributions")

# H0 mean
theta = 0
var = 1
# number of iterations
iter = 10000

# Each Tmax
for Tmax in [50, 100, 200, 1000]
    # Count number of rejections
    rej = 0
    for j in 1:iter
        # first draw
        x = randn()
        # scale
```

```
            x = var*x + theta
            # initialize avg
            avg = x
            # intially don't reject
            test = false
            # Count number of draws
            t = 1
            # accept until reject or Tmax
            while (!test) && (t < Tmax)
                y = randn()
                avg = (t*avg + y)/(t+1)
                t += 1
                if abs(avg) > 1.96/sqrt(t)
                    test = true
                end
            end

            if test
                rej += 1
            end
        end

        println("For ", Tmax, ", size is approx. ", rej/iter)
    end

    # Each Tmax
    for Tmax in [50, 100, 200, 1000]
        # Start with standard critical value
        g = 1.96
        # c value not found yet
        found = false
        # search until c value found
        while !found
            # Count number of rejections
            rej = 0
            for j in 1:iter
                # first draw
                x = randn()
                # scale
                x = var*x + theta
                # initialize avg
                avg = x
                # don't reject
                test = false
                # Count number of draws
```

```
            t = 1
            # accept until reject or Tmax
            while (!test) && (t < Tmax)
                y = randn()
                avg = (t*avg + y)/(t+1)
                t += 1
                if abs(avg) > g/sqrt(t)
                    test = true
                end
            end

            if test
                rej += 1
            end
        end

        # see if size has reached 5%
        if rej/iter <= 0.05
            found = true
            # indexing correction
            g -= 0.01
        end

        # increment c value of consideration
        g += 0.01
    end
    println("For ", Tmax, ", critical value is approx. ", g)
end
```

(Hint: I specified a grid of potential critical values $c \in [2, 5]$ as well as using $c = 1.96$. I then took $n = 10000$ draws of $(x_1, \ldots, T^{max})$ and counted all rejections according to the announced stopping rule. I found c to get the appropriate size.)

5. *Suppose that* $x = [x_1, x_2]'$ *is distributed according to*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathcal{N}(\mathbf{0}, \Sigma)$$

where the unknown variance-covariance matrix Σ *can be written as*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

i.e. the unknown parameter is

$$\theta = \begin{bmatrix} \sigma_1^2 \\ \sigma_{12} \\ \sigma_2^2 \end{bmatrix}$$

(i) *Is this an exponential family? If so, state* k *and state the "ingredients"* $c_i(\theta), T_i(x), i = 1, \ldots, k, d(\theta), S(x), A$.

Let's start by writing the pdf (assuming $|\Sigma| \neq 0$).

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x - \mathbf{0})'\Sigma^{-1}(x - \mathbf{0})\Big) \\ &= \frac{1}{2\pi(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}[x_1, x_2]' \begin{bmatrix} \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} & -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} & \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{bmatrix} [x_1, x_2]\Big) \\ &= \exp\Big(-\ln(2\pi) - \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2) - \frac{1}{2}[\frac{\sigma_2^2 x_1 - \sigma_{12}x_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \frac{\sigma_1^2 x_2 - \sigma_{12}x_1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}]'[x_1, x_2]\Big) \\ &= \exp\Big(-\ln(2\pi) - \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2) - \frac{1}{2}\frac{\sigma_2^2 x_1^2 + \sigma_1^2 x_2^2 - 2\sigma_{12}x_1x_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\Big) \end{aligned}$$

We can see that the above density can be written in the exponential family with $k = 3$ by "ingredients"

- $c_1(\theta) = -\frac{\sigma_2^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)}$, $c_2(\theta) = -\frac{\sigma_1^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)}$, $c_3(\theta) = \frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$
- $T_1(x) = x_1^2$, $T_2(x) = x_2^2$, $T_3(x) = x_1x_2$
- $d(\theta) = -\ln(2\pi) - \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)$
- $S(x) = 0$
- $A = \mathbb{R}$

(ii) *Find the $(k+1)$-st parameter conjugate prior, using the exponential family approach, where you are allowed to simply say that "$\omega(t_1, \ldots, t_{k+1})$ is such that the prior integrates to unity," (i.e., do not calculate it explicitly).*

Directly, from the Prop. given in class, the 4-parameter conjugate prior is

$$\begin{aligned}\pi(\theta; (t_1, \ldots, t_{k+1})) &= \exp\Big(\sum_{j=1}^{3} c_j(\theta) t_j + t_4 d(\theta) - \ln(\omega(t_1, \ldots, t_4)) \Big) \\ &= \exp\Big(- \frac{\sigma_2^2 t_1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} - \frac{\sigma_1^2 t_2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} + \frac{\sigma_{12} t_3}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ &\qquad - t_4\big(\ln(2\pi) + \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\big) - \ln(\omega(t_1, \ldots, t_4)) \Big)\end{aligned}$$

(iii) *Suppose $t_1 = 0, \ldots, t_{k+1} = 0$, and suppose that you have two observations,*

$$x^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

State the resulting posterior.

The idea here is that we use the given conjugate prior (with all $t_i = 0$) and update using the data. Our prior is the 4-parameter conjugate distribution above, and when we update with the first datum, our posterior will be a new 4-parameter conjugate distribution, which we can make our new prior, then update with the second datum.

We update with $x^{(1)}$, then $x^{(2)}$, then write out the explicit form, using the Prop. from class to avoid the intermediate algebra. I slightly abuse notation in that I use $\omega(\cdot)$ throughout as a fudge factor to make the density proper.

$$
\begin{aligned}
\pi(\theta \mid x^{(1)}, x^{(2)}) &= \pi(\theta \mid x^{(1)}) f(\theta \mid x^{(2)}) \\
&= \pi(\theta; (0_1, \dots, 0_4)) f(x^{(1)} \mid \theta) f(x^{(2)} \mid \theta) \\
&= \pi(\theta; (T_1(x^{(1)}, \dots, T_3(x^{(1)}), 1) f(x^{(2)} \mid \theta) && \text{(Slide 28)} \\
&= \pi(\theta; (T_1(x^{(1)}) + T_1(x^{(2)}), \dots, T_3(x^{(1)}) + T_3(x^{(2)}), 2) && \text{(Slide 28)} \\
&= \exp\Big(- \frac{\sigma_2^2(T_1(x^{(1)}) + T_1(x^{(2)}))}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} - \frac{\sigma_1^2(T_2(x^{(1)}) + T_2(x^{(2)}))}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} + \frac{\sigma_{12}(T_3(x^{(1)}) + T_3(x^{(2)}))}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\
&\qquad - 2\big(\ln(2\pi) + \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\big) - \ln(\omega(\cdot))\Big) \\
&= \exp\Big(- \frac{\sigma_2^2(1^2 + 1^2)}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} - \frac{\sigma_1^2(1^2 + (-1)^2)}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} + \frac{\sigma_{12}((1 \times 1) + (1 \times (-1)))}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\
&\qquad - 2\big(\ln(2\pi) + \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\big) - \ln(\omega(\cdot))\Big) \\
&= \exp\Big(- \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} - 2\big(\ln(2\pi) + \frac{1}{2}\ln(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)\big) - \ln(\omega(\cdot))\Big)
\end{aligned}
$$

(iv) *(For the cracks only) How would you generalize this to $x \in \mathbb{R}^m$? Does the prior have a closed form? (Hint: look up "(inverse) Wishart distribution". Perhaps examine $m = 1$ first.).*

So... we were somewhat given the answer in the hint. The inverse Wishart distribution is the general name for the family of priors we found in part (ii), generalized to n-dimensional normal distributions. Instead of just having real number parameters in the density, however, we use a matrix representation.

Formally, we can always demeam our distribution so we have a p-variable normal $\mathcal{N}(\mathbf{0}, \Sigma)$. Then our conjugate prior has an inverse Wishart distribution $\mathcal{W}^{-1}(\mathbf{\Psi}, \nu)$, where $\mathbf{\Psi}$ is $p \times p$ and $\nu \in \mathbb{R}$.[4] Then after n observations, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, the poster distribution is $\mathcal{W}^{-1}(\mathbf{\Psi} + \mathbf{X}\mathbf{X}', \nu + n)$. Of course, then we can use this as the prior for further updating, etc.

[4] I'm using Wikipedia's notation.