Exercise Sheet 5

Model Checks, Model Selection, Multivariate Distributions

Solutions. You do not need to submit your exercise solutions to us. Of course you are free to ask questions during tutorial hours! Exercise solutions will be published on Moodle at the same time with the exercise sheet. Use solutions responsibly, otherwise you risk not being able to solve the exam to a satisfactory level.

Exercise 1 Negative Binomial Distribution Derivation

In this exercise you will reproduce the derivation of the negative binomial distribution.

a) The negative binomial distribution gives the probability distribution

$$p(k|r,p) = \binom{k+r-1}{k} (1-p)^k \, p^r$$

over how many k failures are needed before r successes are reached in a binomial experiment, with expectation and variance

$$E[k|r,p] = \frac{r\,(1-p)}{p}, \quad \mathrm{Var}[k|r,p] = \frac{r\,(1-p)}{p^2}.$$

Show that by setting $E[k|r,p] \equiv \lambda$, the probability distribution becomes

$$p(k|r,\lambda) = \frac{(r+k-1)!}{(r-1)!\,(r+\lambda)^k} \, \frac{\lambda^k}{k!} \, \frac{1}{(1+\frac{\lambda}{r})^r},$$

with $\mathrm{Var}[k|r,\lambda] = \lambda\,\left(1+\frac{\lambda}{r}\right)$.

Hint: Use the definition of the binomial coefficient via factorials.

b) Show that $\lim_{r\to\infty} p(k|r,\lambda) = \mathrm{Pois}(k|\lambda)$.

Hint: See lecture slides.

c) The Poisson likelihood forms a conjugated family with a Gamma prior distribution. Does the negative binomial distribution also form a conjugated family with a Gamma-distributed prior?

Exercise 2 More Toilet Paper Data

After your preliminary A/B study (previous exercise sheet), you collected a whole year of weekly data (toilet paper weight in kg) in order to give more precise predictions to the facility management so that they can plan their (reduced) budget better. The data are stored in `toilet_paper_data.npy` (you may load them with `np.load()`).

a) Like last time, you use a Normal likelihood $y \sim N(\mu, \sigma^2)$ and assume weak empirical Bayes priors for μ and σ: $\mu \sim N(\bar{y}, \hat{\sigma})$ and $\sigma \sim \text{Exp}(1/\hat{\sigma})$, where $\bar{y}$ is the empirical mean and $\hat{\sigma}$ the empirical standard deviation of your data $\mathbf{y}$.

   Simulate the posterior distribution $P(\mu, \sigma|\mathbf{y})$ with PyMC and visualize the marginal distributions $P(\mu|\mathbf{y})$ and $P(\sigma|\mathbf{y})$. What are the 90% highest density intervals for μ and σ?

b) Use `pm.plot_pair()` to also visualize the full joint distribution in a contour plot. Does the posterior look more or less (multivariate) normally distributed? Use `np.cov()` to compute the covariance matrix of the multivariate normal distribution. Are there strong correlations between μ and σ in the posterior?

   Hint: `np.cov()` required a numpy array with each variable (samples for μ and samples for σ) as a column.

c) Compute samples from the predictive distribution using PyMC and use them to compute the root mean squared error (RMSE) and mean absolute error (MAE) averaged over your data and your predictive distribution as done in the lecture.

d) You are not sure, whether using the normal distribution as likelihood (or sampling probability) was appropriate. To this end, perform a posterior predictive check as done in the lecture. Are the observed data within the range of typical predictions by the model? If not, what tendencies do you observe? What might be the reason for the discrepancy?

Exercise 3 Robust Estimation with Student's T Distribution

(This exercise is a continuation of Exercise 1)

After asking back, the facility management tells you that there was a large conference with many external people in the building on one day (where you have the outlier in your data). You consider two different alternatives:

1. Remove the outlier and re-use the same normal model.
2. Keep the outlier and use a model that is capable of taking outliers into account.

Bayesians often prefer to be subjective in their choice of model but rather like to stay objective in their data. Modifications of data are always dangerous, since they are more difficult to document or reproduce than changes to the model

a) You choose a Student's t distribution as likelihood, since it is known to be more robust against outliers than the normal distribution. In addition to mean μ and standard deviation σ, the Student's t distribution also has a parameter ν. In statistical t-tests, ν has typically a non-trivial interpretation as number of degrees of freedom. Here we take a simpler path and use ν to set the Student's t distribution in relation with the normal distribution.

   Use PreliZ' visualisation functionality to compare different Student's t distributions with the normal distribution. In particular, set $\mu = 0$ and $\sigma = 1$ and use $\nu \in \{2, 3, 5, 10, 20\}$. For what limit of ν is the Student's t distribution equal to the normal distribution? In what differs the Student's t distribution from the normal distribution? Observe in particular also the tails of the different distributions.

   Hint: Use the functions `pz.StudentT(mu, sigma, nu).plot_pdf()` and `pz.Normal(mu, sigma).plot_pdf()` and make sure the figure has an appropriate size (e.g. `figsize=(8,6)`) and region of interest (e.g. `plt.xlim(-4,4)`) so that the differences are clearly visible.

b) Repeat your modelling efforts from Exercise 1 using a Student's t distribution as likelihood instead of a normal distribution. To this end, use a Gamma($s = 10, r = 1.1$) prior distribution for ν (that you should visualize beforehand). What are the 90% highest density intervals for μ, σ and ν? From looking at ν, do you expect the resulting Student's t likelihood to have much fatter tails than the previously used normal likelihood?

c) Repeat the posterior predictive test that you did in Exercise 1 for your new model and compare it to the old one with normal likelihood. How do they differ?

d) Compute again RMSE and MAE. What are the differences to Exercise 1 where a normally distributed likelihood was assumed? Were you expecting smaller or larger values compared to Exercise 1? What could be the reason that RMSE and MAE are not as much smaller as you wish them to be?

e) Compare both models (normal vs Student's t likelihood) in terms of expected log predictive density (ELPD). Can you clearly favour one model of the other?

Exercise 4 Railway Switch Maintenance

For resource planning (personnel and pre-stored material), SBB wants to assess how often their railway switches turn up defective and have to be maintained. To this end, they collected for 10 years the weekly counts of defective railway switches (fictitious data, stored in `railway_switch_data.npy`).

In this exercise you want to assess whether it is better to use a Poisson or a negative binomial likelihood for your Bayesian model. Since you know that models with more parameters typically pay a price in terms of additional uncertainty, you proceed very carefully.

a) Run posterior simulations for two models:

   1. A model with Poisson likelihood and a Normal$(\bar{y}, \hat{\sigma})$ prior on λ, where $\bar{y}$ is the empirical data mean and $\hat{\sigma}$ the empirical standard deviation.
   2. A model with negative binomial likelihood with the same prior on λ and a Gamma$(2, 0.1)$ prior on r.

   Quickly visualize the priors used before the simulation - do they make sense for you? Visualize the marginal posterior distributions of the involved parameters and assess their uncertainty.

   Hint: PyMC does not use exactly the same parameter names for the prior distributions as used here and in the lecture. Make sure you find the correct mappings.

b) Perform a posterior predictive check of both models and plot the side-by-side. Which model do you prefer after the posterior predictive check?

c) Compute the Bayes factor between the marginal likelihoods for the two models. Which model explains the data better in your belief? Also try to compute the marginal likelihoods $p(d|\mathcal{M}_{\text{Pois}})$ and $p(d|\mathcal{M}_{\text{NegBin}})$. Would you expect that they are so extremely small?

   Hint: Use SMC as introduced in the lecture and the companion notebook.

d) Compare the models in terms of ELPD. Is there a clear winner when involving the one-standard-error rule?

e) Compute predictive RMSE and MAE for both models and compare them.

Exercise 5 Estimating Churn Proportions

You work at a national health insurance company. Recently, significantly more customers than usual have terminated their insurance policies with your company and changed to somebody else. In the insurance jargon, the number of customers terminating their contract is called *churn*.

The management has ordered an investigation that was carried out by the sales department: they have tried to contact 200 customers that have left and asked them about the reason why they had terminated their contract. They could actually reach 150 customers and out of these 112 gave a suitable answer. Finally, the sales departement has put the different churn reasons into three main categories:

1. premium ('prämie') too high ($k_1 = 82$ cases)
2. unsufficient coverage ($k_2 = 24$ cases)
3. unsuitable insurance models ($k_3 = 6$ cases)

As a data scientist, you are aware that this is a restricted sample and decide to use a Bayesian model with multinomial likelihood and flat Dirichlet prior to model the true underlying proportions π_1, π_2 and π_3 of your entire insurant population including uncertainty.

a) Simulate the posterior with PyMC and visualize the marginal posterior distributions of π_1, π_2 and π_3. What 90% HDIs can you communicate back to the management? Formulate a sentence.

b) Use Dirichlet-multinomial conjugacy and PreliZ to visualize your posterior on a probability simplex.

c) Your superior is unsure about your results. She says: "It's very typical that people switch contracts because of the premium. I would have used a Dirichlet([10,1,1]) prior that reflects this."

   Use PreliZ to visualize this prior on a simplex and marginalized (to look at π_1 specifically) and compute the prior mean and standard deviation for π_1. Finally, re-run your simulation and show that your results are still pretty similar and that your posterior is quite insensitive to this choice of prior.

Exercise 6 The Multivariate Normal Distribution

Even though the multivariate normal distribution plays a big role in Bayesian inference, it will not play a major role in the first part of this lecture, since the `bambi` library does everything for us in hiding complexity. However, the multivariate normal distribution will become more important, when you will work with Gaussian processes in the second half.

The following three exercises are meant to get to know the multivariate normal distribution a little bit better.

a) Load the credit rating data (from `credit_data.csv` in the exercise materials) and select `'Income'` as x and `'Rating'` as y.

   Compute (classical) estimators for $\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and the covariance matrix Σ.

   Visualize the resulting normal distribution $N(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ using a contour plot and overplot with the data. Ponder (qualitatively) whether the normal distribution is a good model for this data.

   Hints: You may use `np.cov()` to compute the covariance and code in the notebook `'week5_multivariate_distributions.ipynb'` to produce the contour plot.

b) Show that if $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$, then

$$x_1 + x_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

   Hints: Compute $E[x_1 + x_2]$ and $\mathrm{Var}[x_1 + x_2]$ and use linearity in the first and that $\mathrm{Var}[x_1 + x_2] = \mathrm{Var}[x_1] + \mathrm{Var}[x_2]$ for independent x_1 and x_2.

c) Let $X \sim N(\boldsymbol{\mu}, \Sigma)$ be a random vector. Show that $A\,X + \mathbf{b} \sim N(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

   Hints: Compute $E[A\boldsymbol{\mu} + \mathbf{b}]$ and use linearity. Compute $\mathrm{Cov}(A\boldsymbol{\mu} + \mathbf{b})$ and use that $\mathrm{Cov}(A\boldsymbol{\mu}+\mathbf{b}) = \mathrm{Cov}(A\boldsymbol{\mu})$ and $\Sigma = \mathrm{Cov}(X) = E\left[(X - E[X])\,(X - E[X])^T\right] = E[X\,X^T] - E[X]\,E[X]^T$.